Reconsideration of the simulated work task situation
A context instrument for evaluation of information retrieval interaction
Borlund, Pia; Schneider, Jesper Wiborg

Published in:
IIiX’10 Proceeding of the Third Symposium on Information Interaction in Context, New Brunswick, NJ, USA, August 18-21, 2010

DOI:
10.1145/1840784.1840808

Publication date:
2010

Document version
Publisher’s PDF, also known as Version of record

Citation for published version (APA):
Reconsideration of the Simulated Work Task Situation:
A Context Instrument for Evaluation of Information Retrieval Interaction

Pia Borlund
Royal School of Library and Information Science
Fredrik Bajers Vej 7K
DK-9220 Aalborg East, Denmark
Phone: (+45) 98 15 79 22
pb@iva.dk

Jesper W. Schneider
Royal School of Library and Information Science
Fredrik Bajers Vej 7K
DK-9220 Aalborg East, Denmark
Phone: (+45) 98 15 79 22
jws@iva.dk

ABSTRACT
The present paper reports on the initial study and the preliminary findings of how the concept of simulated work task situation is reported used in the research literature. The overall objective of the study is in a systematic manner to learn how and for what types of evaluations the concept is applied. In particular we are interested to learn whether the recommendations for how to apply simulated work task situations are followed.

The preliminary findings indicate a need for clarifications of the recommendations of how to use simulated work task situations. Particularly with respect to ‘realism’ of the simulated work task situations, which is emphasised through the need for tailoring of the simulated work task situations towards the group of study participant to ensure the depicted situations are realistic and interesting from the participants’ point of view. Likewise it seems that the recommendation to involve the study participants’ own information needs (to function as baseline of search interaction) is generally neglected in the reported studies.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – search process.

General Terms
Experimentation, Human Factors, Performance, and Reliability.

Keywords

1. INTRODUCTION
This paper looks at how the concept of a simulated work task situation is reported used in the research literature. The present study serves as a starting point for further exploration of the use of this concept for empirical user evaluations. The overall objective is to identify how the concept of simulated work task situation is applied, and for what types of evaluations it is used. In particular we want to learn about the use, and unintentional use of the concept in order to clarify and improve the recommendations for the application of simulated work task situations. The reported use of this concept and the insight gained will help to set directions for future research on the refinement of the overall ‘IIR evaluation model’ [e.g., 14], which the concept of simulated work task situation forms part of.

The concept of simulated work task situation was introduced in 1997 in a paper by Borlund and Ingwersen [8] as an instrument for user-authentic evaluation of IR system effectiveness and user satisfaction with retrieved information. The concept has further found use in behavioural studies of information searching and user-system interaction [e.g., 2; 74]. In 2000 Borlund developed the framework for interactive IR (IIR) systems evaluation known as the ‘IIR evaluation model’. An evaluation model that includes the application of simulated work task situations on the basis of a set of empirically based recommendations of how to use this concept [11; 12; 14]. The IIR evaluation model is composed of three parts:

Part 1: A set of components which aims at ensuring a functional, valid, and realistic setting for the evaluation of IIR systems (i.e., the involvement of potential users as study participants; the application of individual and potentially dynamic information need interpretations; and the assignment of multidimensional and dynamic relevance assessments).

Part 2: Empirically based recommendations for the application of the concept of a simulated work task situation; and

Part 3: A call for alternative performance measures capable of managing non-binary based relevance assessments.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

IIiX 2010, August 18–21, 2010, New Brunswick, New Jersey, USA.
Copyright 2010 ACM 978-1-4503-0247-0/10/08...$10.00.

1 Alternative with respect to the performance measures of recall and precision traditionally employed.
Section 1 presents a general introduction to the evaluation of Information Retrieval (IR) systems. Section 2 focuses on the multidimensional relevance of the model, highlighting the importance of understanding user-oriented approaches that can inform future evaluations.

Section 3 outlines the empirical data analysis conducted in the study. It discusses how simulated work task situations are used to generate relevance assessments that are more reflective of real-world scenarios.

Section 4 provides detailed insights into the types of evaluations that are possible with the proposed model and demonstrates the validity and improvement of user-oriented IR approaches through application.

The paper concludes with concluding statements that emphasize the potential for further research and development in the field of IR evaluation.
describe an imaginary situation of, e.g., a middle-aged woman looking for information on climacteric related conditions. The described situation ought to be authentic, relevant, and realistic to the university students – males and females – so that it leads to realistic interpretation of, and interaction with the simulated information needs. The advice to tailor the simulated work task situations entails homogeneity of the group of study participants. They need to have something in common, which can form the foundation for development and application of the simulated work task situations. It is furthermore recommended to employ a combination of simulated work task situations (simulated information needs) and the study participants’ genuine information needs – both when pilot testing and when carrying out the actual evaluation. This means that the study participants should bring with them real personal information needs which they search as part of the evaluation. Hence, genuine information needs function as a baseline against the simulated information needs. In addition, the genuine information needs provide information about the systems’ effect on real information needs. The inclusion of genuine information needs is also useful in the pilot test of the test setting, test procedure, and the study participants’ perceptions of the simulated work task situations as the study participants’ personal, real information needs can inspire to ‘realistic’ and user-adaptable simulated work task situations. Let this also be a reminder of the ever good test practice of pilot testing prior to actual evaluation. The final piece of advice is to permute the order of search jobs between the study participants so that no study participant is given the same simulated work task and their own personal information need in the same order. This is to neutralise any effect on the results in terms of bias of search interaction and relevance assessment behaviour of the study participants as well as the study participants’ increasing system knowledge and knowledge of domain topicality of the simulated work tasks situations.

With this brief introduction to the concept of simulated work task situation and the summary of the empirically based recommendations of how to use simulated work task situations we close section 2 and move on to section 3 below. Section 3 presents how the research literature that reports on the use of simulated work task situations is identified and grouped.

3. METHODS

In order to investigate the use of the concept of simulated work task situation as reported in the research literature, the research literature in question needs to be 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 is carried out on the basis of the six publications authored by Borlund [8-12; 14] in which the concept of simulated work task situation is presented (e.g., see Table 1). The citation count as of May 2009 is 193 for the six publications. The distribution of the received citations with respect to the six publications is depicted in Table 1.

On the basis of the citation analysis 157 individual publications are identified. Another 41 individual publications are identified as a result of searching online repositories. The repositories were searched during August of 2009. In total 198 individual publications cite one or more of the six publications by Borlund. Paper copies are made of every single publication. One publication, however, is represented only by an abstract in English as the main text is in Japanese [72].

In order to get an overview of the 198 publications they are (roughly) organised into six categories according to the focus and content of the publications. The categories are the following six:

1. Empirical evaluation by use of simulated work task situations;
2. Empirical evaluation, but not by use of simulated work task situations;
3. Theoretical evaluation;
4. Relevance issues;
5. Performance measures; and
6. ‘Other’.

Category 1 contains papers that report on empirical evaluation by use of simulated work task situations [e.g., 76; 88]. Category 2 concerns papers that report on empirical evaluation, but that do not evaluate by use of simulated work task situations [e.g., 18]. The papers in category 2 might, however, refer to simulated work task situations as a potential way of evaluation. Category 3 holds papers on evaluation of IR systems, but that do not report on any actual empirical evaluation, i.e., papers that discuss approaches to evaluation or propose new ways for evaluation [e.g., 6; 35]. The 4th category contains papers that deal with the concept of relevance and relevance issues [e.g., 53]. The before mentioned Japanese paper by Sagara [72] is categorised as a ‘relevance paper’. Category 5 deals with papers concerned with performance measures [e.g., 42]. The 6th category of “Other” includes a mixture of papers for example the ARIST review by Cool [19, p. 15] on the concept of situation in Information Science in which she refers to simulated work task situation as an approach to represent the salient aspects of a person’s IR problematic situation. Or the paper by Byström and Hansen [17] that presents a theoretical discussion of the concept of work task—including simulation of work tasks as by Borlund. The distribution of the 198 papers with respect to the six categories is shown in Table 2.

### Table 1. Distribution of received citations to publications by Borlund.

<table>
<thead>
<tr>
<th>Citation analysis (May 2009)</th>
<th>Web of Science®</th>
</tr>
</thead>
<tbody>
<tr>
<td>Borlund &amp; Ingwersen, 1997 (JDOC)</td>
<td>[8]</td>
</tr>
<tr>
<td>Borlund &amp; Ingwersen, 1998 (SIGIR)</td>
<td>[9]</td>
</tr>
<tr>
<td>Borlund &amp; Ingwersen, 1999 (MIRA)</td>
<td>[10]</td>
</tr>
<tr>
<td>Borlund, 2000 (JDOC)</td>
<td>[12]</td>
</tr>
<tr>
<td>Borlund, 2003 (INFO RESEARCH)</td>
<td>[14]</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
</tr>
</tbody>
</table>
The further analysis of the present paper is made with respect to the 85 papers of category 1, that is, papers which report of actual empirical evaluations by use of simulated work task situations [1-4; 7; 16; 20-23; 25-31; 34; 36-41; 44-47; 48-52; 54; 56-68; 70; 71; 73; 74; 76-104; 106; 107]. These papers are studied with particular focus on the methodological aspects of the reported evaluations. The initial findings of the preliminary reading and analysis of the 85 papers are presented in section 4. Before moving on to the results section the limitations of the present study should be addressed.

There are limitations to the present study as it does not include monographs [e.g., 33; 105], or doctoral theses [e.g., 5; 24; 55; 69; 75] or recent work like the paper by Kelly [43] as these are not included in Web of Science® or indexed at the time of searching. It is the plan that these publications will be included in a future analysis. For the purpose of the future analysis an exhaustive citation analysis is planned by use of SCOPUS. The objective is to verify any further publications that are to be included in the overall study. The planned citation analysis will also include the JASIST publication by Borlund from 2003 on the concept of relevance in IR [13]. The citation analysis from May 2009 shows that this particular paper yields 63 citations. While the paper mainly focuses on the concept of relevance it does touch upon the concept of simulated work task situation, and ought for that reason to be included. It might be considered a limitation that the paper is not included in the present analysis. Nevertheless, the Borlund relevance paper was omitted due to its primary focus on relevance. Another limitation to the present study is that only publications that make explicit reference pointers to any of the six publications by Borlund are included. In other words, studies that employ simulated work task situations, but do not make any direct referral to the publications by Borlund are not included. Though these studies might be potentially relevant and of interest to our study they are hard to identify when no direct reference pointers are given, hence they are not included.

4. RESULTS and FUTURE WORK
The preliminary reading of the 85 papers reveals a variety of types of evaluations for which simulated work task situations are used, for example for evaluation of: relevance feedback [70; 71]; search behaviour of online museum accessibility [74]; image retrieval [40]; retrieval of broadcasts [26]; XML retrieval [50-52; 64; 65; 81-85]; cross-language retrieval [1; 41]; information access and search performance of visually impaired [21]; journalists’ satisfaction with search results [7]; mobile information systems [4; 25; 56]; and how to offer users strategic help via the interface [16] – just to mention some.

The first indication of a future work is given by the relatively high numbers of studies carried out by use of simulated work task situation. This indicates a genuine need for this evaluation instrument, and acceptance of it by the IR and information seeking communities. Though often used it is not the same as being validated. Only one study has validated the use of simulated work task situations and that is the thesis work by Borlund [11; 12]. This calls for a validation study of construct validity of simulated work task situations. So it does because the incitement to use the concept of simulated work task situations in one’s study of IR system interaction is to obtain reliable results of humans’ interaction with the IR system(s). This makes it crucial to know whether reliable results that reflect genuine information interaction and assessment behaviour of humans are obtained. In other words it is vital to know what it requires of a simulated work task situation to prompt the desired realistic behaviour. The concept of simulated work task situation was put forward in 1997, and since then the Internet has become popular and widespread and probably is the main information source to most users. It is a fact that the Internet makes us independent information searchers (true end-users) as well heavy users of information as access to information is just a click away. Hence a validation study that can either corroborate or falsify the human search and assessment behaviour of the first study is in place. The impact of the Internet is also visible in several of the 85 studies as quite many focuses on web retrieval and web information access [e.g., 1; 3; 20; 21; 26; 31; 38; 44; 46-48; 56-58; 64; 65; 68; 74; 86; 87; 89; 95; 96; 99-101]. The numerous studies indirectly support the need for a validation study that can verify whether the more experienced information searchers of today follow the same interaction pattern as earlier, or whether changes should be made to the design and use of simulated work task situations.

A final note in this respect: Within Information Science we do not have a strong tradition for repeating previous studies, as we seem to be more focused on the novelty of research. But in order to strengthen our research and the approaches we apply and build upon in Information Science validation studies (no matter how boring and indifference they might seem) ought to be acknowledged and carried out, since it is the way to further develop and mature our field of science, our research practices, and the methods we use for research.

The 85 cases of reported use of simulated work task situations clearly demonstrates that the use of the concept is applied with the purpose of achieving reliable user-system interactions as they would take place in real life. This, however, requires a certain loyalty towards the recommendations put forward with reference to the use and construction of the simulated work task situations. Here the preliminary reading brings attention to at least three issues of evaluation practice, which are with respect to: 1) tailoring of simulated work task situations towards the group of study participants; 2) ensuring the simulated work task situations are of interest to the study participants; and 3) the use of the participants’ genuine information needs in combination with simulated work task situations.

The tailoring of the simulated work task situations towards the group of study participants is not always that successful. The
purpose with tailoring is to make sure that the participants can identify themselves with the situations presented, and hence ensure realism of the evaluation. There are cases [e.g., 39; 67] where university students are to imagine themselves being either members of an information broker company, or imagine how they are to prepare for an one-hour talk on the subjects ‘tea’ or ‘everyday drinks for old people’ – a talk the participants further are to imagine to give in two weeks time at the neighbourhood library to an audience of ordinary people [67, p. 540-541]. These examples are hardly realistic given the study participants’ personal situations as, e.g., computer science students, engineer students, or information science students. In our opinion, these studies do not provide the optimal condition for ‘realistic’ IR interaction and hence are questionable. In order to put power behind the emphasis of tailoring of simulated work task situations, and to simply verify this recommendation, it would be interesting to further investigate the effect of non-tailored simulated work task situations compared to tailored situations. Even the result of no effect would be useful to know for future evaluations.

With reference to the recommendation to ensure the study participants find the simulated work task situations topical interesting in order to achieve dedicated IR interaction during testing, the research literature [e.g., 1; 2; 39] presents examples of how this is solved alternatively by inviting the study participants to select the simulated work task situations/topics they prefer to search. In the study by Airio [1] the study participants get to choose between 10 simulated work task situations of which they select four that they would like to search. In the case of Arapakis and colleagues [2] the study participants are to search three simulated work task situations at each their level of difficulty. For each level of difficulty the study participants can choose between two scenarios of simulated work task situations. The study by Joho and his team [39] is another example of how the study participants get to choose three topics out of 15 TREC topics to be inserted in the context of a simulated work situation. Joho and his co-workers [39] further normalise for the possible effect of ‘interestness’. “The 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” [39, p. 92]. Though the solution might be to invite the study participants to choose the task(s) they prefer to search, the danger does exit in terms of providing too many different simulated work task situations to choose between that generalisation of search interaction cannot be made across the group of study participants. To create topically interesting simulated work task situations require partly an insight and understanding of the study participants’ typical information needs, and partly that the group of participants is homogeneous by sharing interests that can be used for searching. We do not really know enough about how the issue of interest or relevancy of simulated work task situations affects the study participants’ and their system interaction and relevance assessments. This lack of knowledge obviously calls for further studies.

With respect to the recommendation to use a combination of simulated work task situations and the study participants’ personal, genuine information needs within the same evaluation; this recommendation is very rarely followed. The purpose for doing so is to allow for the study participants’ personal genuine information needs to act as baseline for the search behaviour and relevance assessments prompted by the simulated work task situations. Hereby one has an instrument to compare, interpret, and validate the participants’ interaction patterns achieved by use of the simulated work task situations. So far only Blomgren, Vallo, and Byström [7] have included the study participants’ own information needs. They 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” [7, 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. The familiarity of the task requirements and a higher motivation lead to better values for Precision, RHL index and Satisfaction than they did in relation to simulated work tasks” [7, p. 66].

The results and experiences by Blomgren, Vallo, and Byström [7] confirm two things: 1) it is important to involve also the study participants’ real information needs because the search behavior on the basis of the real information needs provides an indication of how (realistically) the search behavior of the simulated work task situations can be interpreted; and 2) tailoring of the simulated work task situations is essential for engaged and reliable search interaction by the study participants.

As previously mentioned Arapakis and colleagues [2] evaluate by use of different levels of difficulty of simulated work task situations. This might make perfect sense, but is, however, not validated. In other words, we do not know whether the various levels of difficulty results in the expected behaviour, or what the expected behaviour is. This needs further investigation and validation so that appropriate recommendations and guidelines for construction of this type of simulated work task situations can be made. The study by Bell and Ruthven [3] 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 [32, pp. 116-117] known as 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., 95; 96; 100; 102] with IR carried out with respect to ‘fact search’, ‘decision search’, and ‘background/exploratory search’ as well as by Toms and her group of colleagues [88]. The study by Toms and her colleagues is very interesting in that it compliments the work by Bell and Ruthven [3] and validates issues of relevance for task constructions, and hereby adds to our understanding of how different types of search tasks and task structures lead to different search efforts of the study participants. Clearly further research is needed to get a deeper understanding of the sub-components of work tasks and their effects on IR system use and interaction. As such we are in line with Toms and co-workers components of work tasks and their effects on IR system use and interaction. As such we are in line with Toms and her group of colleagues [88]. The study by Toms and her colleagues is very interesting in that it compliments the work by Bell and Ruthven [3] and validates issues of relevance for task constructions, and hereby adds to our understanding of how different types of search tasks and task structures lead to different search efforts of the study participants. Clearly further research is needed to get a deeper understanding of the sub-components of work tasks and their effects on IR system use and interaction. As such we are in line with Toms and co-workers [88; 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…”. In other words, more research is needed, and with those words we close the reporting of preliminary results.

5. CONCLUSIONS
The preliminary study of the 85 evaluations that use the concept of simulated work task situation demonstrate the wide use and acceptance of this concept for evaluation of IR system interaction. At the same time the reported evaluations also illustrate the need for clarification of the guidelines and recommendations of how to use simulated work task situations for evaluation, which have lead to identification of numerous relevant future studies. The present study confirms the need for a careful and thorough analysis of the existing research literature as planned of which this is the preliminary analysis and reporting. The objective of the future analysis is identical to this study, namely, to learn how, and for what types of evaluations the concept is applied. In particular, to learn about the use of the concept of simulated work task situation in order to improve and refine the recommendations for the application of the concept and hereby to set directions for future research on simulated work task situations.

6. ACKNOWLEDGMENTS
Sincere thanks go to the librarians of the Library of the Royal School of Library and Information Science for invaluable help in tracing down less assessable publications — special thanks to Head of Library Lisbeth Rasmussen and ‘cybrarian’ Karen Margrethe Örnstrup. The four anonymous referees are also due thanks for their patience and their generosity with many constructive comments and considerations that have improved the present paper and positively will impact the future work. Thank you!

7. REFERENCES


