The Application of Work Tasks in connection with the Evaluation of Interactive Information Retrieval Systems: Empirical Results

Pia Borlund

Department of Information Studies, Royal School of Library and Information Science,
Birketinget 6, DK-2300 Copenhagen S, Denmark

Peter Ingwersen

Department of Information Studies, Royal School of Library and Information Science,
Birketinget 6, DK-2300 Copenhagen S, Denmark

Abstract

This paper reports on the empirical findings of the meta-evaluation of the application of work tasks in connection with the evaluation of interactive information retrieval (IIR) systems.

The purpose of the meta-evaluation is to uncover if it is recommendable to apply work tasks to future evaluation of IIR systems. It is investigated if any search behavioural differences exist between test persons' treatment of their own real information needs *versus* simulated information needs. The hypothesis is that if no difference exists one can correctly substitute real information needs with simulated information needs through the application of simulated work task situations. We are also interested in learning what defines a 'good' work task situation.

The empirical results of the meta-evaluation provide positive evidence of the application of simulated work task situations in connection with evaluation of IIR systems. The results also point to that tuning of work task situations towards the group of test persons is of importance due to motivation of the test persons. Furthermore, the results of the evaluation show that different versions of semantic openness of the simulated situations make no difference to the test persons' search treatment. Finally, it is verified that there exists a general pattern of assessment behaviour for the test persons. This verification makes yet another experimental reason for permutation of work tasks between test persons, in order to avoid bias of the retrieval results.

1 Introduction

The concept of work task is widely applied within the field of Information Science, for instance in relation to systems design and development [e.g. 1-4], studies of user behaviour and task complexities [e.g. 5-7], as well as at a theoretical level when modelling information retrieval (IR) use and interaction [e.g. 8]. Recently, the concept of work task has been introduced as a core component in relation to the evaluation of interactive information retrieval (IIR) systems [e.g. 9-12]. The popularity of the work task based evaluation approaches is due to an increasing demand for hybrid evaluation approaches which combine two main approaches to the evaluation of IIR systems, the system-driven and the cognitive user-centred approaches, respectively. Basically, the demand can be explained by the fact that IR systems have become more interactive, and consequently cannot be evaluated without including the interactive seeking and retrieval processes [13]. This is so because IIR systems are systems where the user dynamically conducts searching tasks and correspondingly reacts to systems responses over session time. Consequently, the foci of IIR system evaluation include all the user's activities of interaction with the retrieval and feedback mechanisms as well as the retrieval outcome itself.

The essential reason for the employment of the work task based approaches to the evaluation of IIR systems, is that the actual application of work tasks and the involvement of test persons makes it possible to facilitate the evaluation in a way which is close to actual information seeking and IR processes, though still in a relatively controlled evaluation environment.

A work task, which is a short 'cover story' that describes an IR requiring situation, ensures the experiment realism by simulating a real work task situation. Like in real life, the work task situation is to be seen as the cause of the 'breakdown situation', a cognitive state which creates an information need which has to be satisfied in order for the user to be able to deal with the situation and to be able to move on. The issue of realism is also ensured the experiment by the involvement of test persons (potential users) who, based on the provided work task, develop individual and subjective information needs. Individually, the test persons interactively search,

modify, and assess relevance of the retrieved information objects in relation to their information needs and the underlying work task situation. In addition, the work task provides experimental control by being the same for all of the test persons. Due to the control one can compare the search results across the systems and/or system components as well as across the group of test persons.

This paper reports on the empirical results from a meta-evaluation of the application of work tasks intended for the evaluation of IIR systems. The purpose of the meta-evaluation is to uncover whether or not the application of work tasks is recommendable for future evaluation of IIR systems. It is investigated if the use of simulated work task situations result in realistic seeking and retrieval processes compared to the seeking and retrieval processes of users' own information needs. Furthermore, there is an interest in learning what makes a 'good' work task with reference to semantic openness, as well as how tuned work tasks ought to be towards the group of test persons. The meta-evaluation is carried out by the involvement of 24 test persons (university students), who each carried out searches according to one personal information need and four simulated information needs (triggered by work tasks). The test setting is a full-text online system applying parts of the TREC data collection (the newspapers: the Herald, and the Financial Times). The data is collected through transaction logs of the test persons' search activities as well as through post-search interviews with the test persons.

The structure of the paper is as follows: section two briefly presents the definition of the applied work task concept, including the number of functions which the work task serves. Section three outlines the experimental test setting; presents the research questions, and reports on the empirical results of the meta-evaluation. The final section summarises the major points and results presented in the paper.

2 Simulated work task situations

In our work we refer to work tasks as 'simulated work task situations' [10]. By doing so we simply point to the fact that the application of work tasks in this context, i.e. evaluation of IIR systems, is done by the simulation of situations where IR work tasks arise. A simulated work task situation is a semantically rather open description of a scenario of a given IR requiring situation (for an example, see appendix 1). A simulated work task situation serves two main functions, as it:

1) triggers and develops a simulated information need by allowing for user interpretations of the situation, leading to cognitively individual information need interpretations as in real-life;
2) is the platform against which situational relevance is measured [10, pp. 227-228].

More specifically the simulated work task situation helps to describe to the test person:

- the source of the information need;
- the environment of the situation;
- the problem which has to be solved; and
- serves to make the test person understand the objective of the search [10, p. 229].

In the present meta-evaluation we apply two different versions of the work task component. The reason is that we are interested in learning about the consequences (if any) of the level of semantic openness of simulated work task situations presented to the test persons. This leads to a sim-1 version and a sim-2 version (see appendix 1). The sim-1 version is constituted by a *simulated work task situation* as well as of an *indicative request*. The indicative request is a suggestion to the test person about what to search for. The indicative request is not to be seen as an example of the underlying need of the particular simulated work task situation in use. According to the type of test collection and knowledge of the test persons, a definition of the topic on search can be included. Empirical results reported by Spink, Greisdorf, and Bateman [14] support the application of a definition of the topic on search. For instance, in a test situation where a domain specific collection is applied by test persons with little knowledge of the topic. Because the more the test persons know about the information requiring problem, the better they can identify the need and formulate the requests/queries which results in focused retrievals [14, p. 118]. A feasibility study [10] of the applicability of work tasks, which included definitions, revealed that the test persons' search statements mainly were generated based on the simulated work task situations, and less by the definitions or the more specific indicative requests. In other words, the search statements and the subsequent search results were not biased by the application of definitions and indicative requests.

A sim-2 version is constituted by the *simulated work task situation*, only. The exclusion of the indicative request (and definition) makes the sim-2 version more semantically open. The employment of the two sim-

versions as well as topically different 'sims' are applied with the intention of learning about how specific or semantically open one may express simulated information need situations, as well as how tuned simulated work task situations ought to be in order to be perceived and accepted by the test persons.

3 The Meta-Evaluation

3.1 Experimental Conditions

This section reports on the meta-evaluation of the use of simulated work task situations. By meta-evaluation we mean evaluation of an evaluation sub-component, in this case the work task component. The meta-evaluation was carried out in Spring 1998 at the University of Glasgow, Scotland.

The test setting is a full-text online system applying TREC data (the Herald, and the Financial Times collections[1]) and a probabilistic based retrieval engine [15]. The search activities were logged, including the test persons' relevance scores for the retrieved documents. The test persons indicated the relevance scores by use of a slider[2], which was incorporated into the interface, and shown at the screen next to the title field and the field viewing the full-text documents. The test persons assessed *usefulness* of the retrieved documents. They based their relevance assessments on either the title or the full-text of the document. In addition, the test persons indicated the degree of relevance of the assessed documents according to the relevance categories of: low, medium, and high relevance.

24 university students volunteered as test persons to the experiment. The test persons (20 male students, 4 female students) are from various academic fields and educational levels, e.g. computing, mathematics, geography, biochemistry, language, English history, psychology etc., and are undergraduate, and graduate students, respectively.

3.2 Experimental Procedure

When the test persons signed up for the experiment they were asked to prepare a real information need, which they should bring with them to the actual experiment. They were also asked to describe why and for what use the information need was relevant. The real information need was required to be of personal interest and of importance to the test persons. At the same time the information need should be able to be met by a newspaper full-text collection. Further, the test persons were informed that the time span of the applied data collection runs from January 1995 to May 1997. The test persons participation in the experiment took place over two sessions. The experiment took 3 hours on average per test person. They were paid each £ 5 per hour, which they were paid at the end of the second session. The test persons participated in the experiment one by one. The experiment consisted of the following six steps:

1) brief questionnaire as to the test persons searcher skills and experiences;
2) introduction to the experiment;
3) explanation of the test person's part (job) in the experiment;
4) demo of the retrieval system;
5) execution of six search tasks (1 training, 1 real need, and 4 simulated situations);
6) brief post-search interview after each completed search task.

The test persons were instructed that their job was to retrieve as many useful documents as it would take to satisfy their information need, and that their job was not about to retrieve as many relevant documents as possible. The retrieval should stop when the need was satisfied, or when the test persons felt it was not possible to satisfy the information need with the actual newspaper collection. The evaluation was carried out with no time limitation during the evaluation process.

Each test person applied one personal information need and five simulated situations (simulated needs), of which one was for a training purpose. All of the test persons applied both sim-1 and sim-2 versions. During the experiment, the search tasks (except the training task) were permuted between the test persons in such a way that none of the test persons carried out the search tasks in the same order. This was done in order to neutralise any

[1] The two collections, the Herald and the Financial Times, were merged for the purpose of the evaluation, so that the test person were not forced to choose between the two collections. The merging of the two collections also served to avoid yet another variable when analysing the data, namely what collection had been searched.

[2] Hidden to the test persons, the slider assigned a relevance value on the scale from 0-10 according to the final indicated position of the slider. Zero refers to non-relevant, the value ten indicates most relevant. The relevance category 'low' covers the range from zero to three, 'medium' from four to seven, and 'high' from eight to ten.

effect on the results derived from the test persons experiencing system and domain knowledge through the various search tasks. The simulated situations were mainly constructed based on TREC-6 Interactive Topics, with one exception, sim A. The idea of sim A was borrowed from one of the four test persons who participated in the pilot study of the experimental setting of the experiment, prior to the actual meta-evaluation experiment.

The collected data consist of 24 sets of responses to the questionnaires, 120 protocols of shorter verbal statements (i.e., the post-search interviews), and the 120 corresponding search transaction logs. Inference statistical tests, prior to the actual statistical tests, indicates no bias of the results due to the unbalance of the sample of the test persons.

3.3 Research Questions

It is of importance to future use of work tasks in connection with the evaluation of IIR systems, that we gain knowledge about the functionality as well as about the consequences of the application of work tasks. The overall research question is: how does the application of the work task effect the measurable search behaviour and the assessment output when different types of information needs[3] are involved, and the structure[4] and topical types[5] of the underlying simulated situations are varied? More specifically we are interested in learning:

1) Can simulated information needs substitute real information needs?
2) What makes a 'good' simulated situation with reference to semantic openness and types of topics of the simulated situations?
3) Does there exist a pattern of assessment behaviour between the test persons, as to the order of search jobs?

Basically, we try to answer the research questions by uncover if there are any differences in the test persons' behavioural treatment of the various types of information needs during the searching. Research question no. 1 is concerned with the test persons' treatment of their own real information needs *versus* the simulated information needs. Research question no. 2 deals with the test persons' treatment of (1) the two versions, sim-1 *versus* sim-2; and (2) the topical different simulated situations, sim A, B, C, and D *versus* the test persons' treatment of their own real information needs. The 3rd research question is a consequence of the feasibility study reported on in [10]. In the study it was found, though based on a rather small sample of test persons and search tasks, that the test persons had similar assessment behaviour. Regardless of the search job in question they all assessed their first search job to a middle degree of precision, the second search job to a high degree of precision, and the third search job to a low degree of precision. If it is the case that the existence of a certain pattern can be verified, it is yet another reason why the order of the search jobs ought to be permuted between the test persons. By looking into the order of the search tasks and the corresponding precision values, we want to investigate this phenomena further.

3.4 Data Analysis and Results

In this section we compare the test persons' treatment of the different types of information needs in relation to various attributes of search interaction. The test persons' treatment of their own real information needs is considered the baseline. By comparing the various search attributes it is our hope to verify characteristics of simulated situations and especially simulated work task situations, which may improve our understanding of the application of work tasks in relation to future evaluation of IIR systems.

3.4.1 Research Question No 1: Can Simulated Information Needs Substitute Real Information Needs?

This research question is investigated by chi-square testing for differences between the test persons' behavioural treatment of the two types of information needs in relation to various attributes. The search attributes, per search task, are:

- number of search runs;
- average number of search terms;

[3] The different types of information needs refer to real information needs *versus* simulated information needs, respectively.

[4] By structure we refer to the fact that two different versions of simulated situations are applied to trigger the simulated information needs. One version (sim-1)which consists of a simulated work task situation and a indicative request; and the second version (sim-2) which consists of the simulated work task situation, only.

[5] Topical type refers to the various types of different 'cover stories' of which the different simulated work task situations consist of (sim A, sim B, sim C, and sim D).

- number of unique search terms per test person;
- search time spent;
- distribution of applied relevance categories: low, medium, and high;
- distribution of applied relevance categories: non-relevant, low, medium, high, and relevant;
- number of relevance assessments, distributed on applied source (full-text/title).

The inference statistical tests are carried out based on the following hypotheses:

H_0: No difference exists between real information needs *versus* simulated information needs when measure in relation to ... each of the listed search attributes, one by one.
H_a: Difference exists between real information needs *versus* simulated information needs when measured in relation to ... each of the listed search attributes, one by one.

All, except one, of the chi-square tests lead to no rejection of the H_0 hypothesis (see appendix 2). These results are interpreted as a strong indication of no difference between real *versus* simulated information needs. Based on the overall results, we conclude:

1) one can substitute real information needs with simulated information needs through the application of simulated work task situations; and
2) one can mix simulated and real information needs.

The only analysis that reveals a different result is the test of applied source of the assessment (full-text/title). This test results in the support of the H_a hypothesis, which states that there *is* a difference between the two basic types of information needs. This result is supported by the descriptive results of mean values of the numbers of relevance assessments based on either full-text, or titles. On average the test persons made approximately five more relevance assessments based on full-text, when treating their own information needs, than compared to their treatment of the simulated information needs. These descriptive results are interesting compared to the former chi-square test concerning the search time spent per task which led to no rejection of the H_0 hypothesis. Interestingly, because one intuitively assumes that it takes longer time to read/assess relevance of full-text documents than of the titles of documents. A condition which, however, is supported by the descriptive mean value of search time spent per task: The test persons spent 7 minutes more when dealing with their own personal information needs than the simulated information needs. However, this is *not* statistically significant.

The result of the test persons preferring to asses their own information needs based on full-text as to titles is probably due to the motivation of the test persons. The test persons might have been more interested in treating their own real information needs than the simulated information needs. This is in line with findings by Morita and Shinoda [16] who found that users spend more time reading articles of interest to them than articles which are of no interest to them. This informs us about the importance of applying simulated work task situations which can attract the attention of the test persons at a level equal to real information needs, in order to achieve realistic search behaviour.

Finally, answers and comments by the test persons to the post-search interviews support the preliminary conclusions and results of the application of simulated information needs. The test persons were asked whether or not they found the simulated situations realistic. 87.5% answered YES opposed to the 12.5% of the test persons who answered NO. The high number of test persons who approved of the simulated situations support why there, in general, is no difference in the treatment of the two basic types of information needs. In addition, we wish to point out that there is no correlation between the test persons who did not find the simulated situations realistic and test persons who, at the same time, were not able to retrieve documents that satisfied their information needs.

3.4.2 Research Question No. 2: What Makes a 'Good' Simulated Situation?

This research question is approached differently. First, we carry out the chi-square tests of differences between real information needs *versus* sim-1 (simulated work task situation + indicative request) *and* the semantically more open sim-2 versions (simulated work task situation, only). Followed by the chi-square tests of the two sim-versions, sim-1 *versus* sim-2. Secondly, chi-square tests are carried out in regard to sim A, B, C, and D *versus* real information needs. In addition, the inference statistical results are viewed in relation to the descriptive results of the qualitative part of the collected data.

The chi-square tests concerning real needs *versus* sim-1, and real needs *versus* the sim-2 version are carried out with reference to the previous list of search attributes (see section 3.4.1). The underlying idea for

these tests is to investigate if any of the sim-versions result in a treatment similar to the test persons' treatment of their own information needs.

The tests reveal the same pattern of results as in the tests regarding research question no. 1. The tests reveal no indication of the existence of differences between the test persons' treatment of the involved sim-versions compared to the treatment of the real information needs (e.g., see appendix 2). Again, the test of the applied source (full-text/title) of the relevance assessments leads to the support of the alternative hypothesis.

Further, chi-square analyses are carried out concerning the sim-1 *versu*s sim-2 versions (e.g., see appendix 2). These analyses confirm that no differences exist between the two versions. Based on this we conclude that: None of the sim-versions are better than the other. Or put differently, when applying work tasks to the evaluation of IIR systems it makes no difference if one makes use of the work task version which consists of both the 'simulated work task situation' and the 'indicative request' (sim-1) or just the 'simulated work task situation' (sim-2) which is semantically more open.

In the post-search interview the test persons were asked if it made any difference to them, if they had had either a sim-1 version or a sim-2 version. 29% replied YES – it made a difference, and 71% of the test persons said NO – it made no difference. This support the inference statistical results of no differences between the real information needs and the two simulated versions. All of the YES-answers are in favour of the indicative requests. Interestingly, the test persons explain their 'YES' differently. A few of the test persons say that the indicative requests made it easier to generate the queries as they picked the search terms from the indicative requests. One test person says it was helpful because the indicative request helped him understand what was expected from him. Others simply state they preferred having both. Finally, one of the test persons says he did not use the indicative request in relation to the query formulation, but had found it useful when scanning for relevant information. This tell us, though it is not statistically significant, that different types of applications have been employed in relation to the sim-1 version. This indicates that the employment of the indicative requests can be constructively applied in combination with the simulated work task situations.

Further chi-square tests of topically different sim A, B, C, and D *versus* real information needs are carried out with the purpose of identifying which of the sim A, B, C, and D that have been treated most similar to the real information needs. For the first time, the results of the chi-square analyses (see appendix 2) as regard the source (full-text/title) of the relevance assessments lead to no rejection of the H_0 hypotheses. Thus no difference exists between real information needs and sim A and B, respectively. This means that sim A and sim B are based on similar proportion of full-text/title relevance assessments as the real information needs. We interpret the application of 'source' (full-text/title) as a sign of *motivation.* The more full-text based assessments, the higher motivation. So far, the analyses of real needs have resulted in more full-text based assessments as compared to the treatment of the simulated needs. This change with sim A and sim B. These two simulated work task situations have motivated the test persons to treat the work tasks similar to their own information needs. This make sim A and sim B to examples of 'good' work tasks as opposed to sim C and sim D.

The chi-square analyses of real needs *versus* sim C *or* D with reference to each of the search attributes, result in the usual set of 'no rejections' of H_0 hypotheses with the same exceptions as earlier. Thus, on the overall level sim C and sim D are just as 'good' as the sim-1 and sim-2 versions. However, they are at least better than sim A and sim B in one respect as the chi-square tests of sim C *or* D *versus* the real needs lead to no support of the alternative hypothesis in regard to the search attribute: Average number of search terms per task. The mean values of the average number of search terms per search task show that the test persons applied more terms in their treatment of the information needs triggered by sim A (mean: 4.03) and sim B (mean: 4.62) than their personal needs (mean: 3.28).

We interpret the results as follows: of sim A, B, C, and D, sim A and B are the best to motivate the test persons as these work tasks are treated with a similar proportion of full-text based relevance assessments as the real information needs. However, the revealed differences in the treatment of real information needs *versus* sim A *or* sim B with reference to the average number of search terms per search task show that sim A and sim B did not result in a search behaviour similar to the one of the real information needs. Sim A and sim B are closest in nature to real needs.

When the test persons were asked about which of the simulated work task situations they preferred, 79.2% voted sim A: ('graduate') as either their first or second priority. 41.7% voted the training task ('Hubble Telescope') as either their first or second priority. The simulated work task situations which the test persons liked the less were sim D ('burglary') and sim C ('Arran'). 45.8% of the test persons did not like sim D, and 29.2% did not approve of sim C. One test person though, voted sim D as her first priority, but she had also recently been paid a visit by a burglar – she explained. Sim B ('student job') was perceived rather neutral by the test persons.

Nevertheless, based on the test persons comments in this respect, we do have an idea of what makes a good simulated situation. The test persons comments reveal three characteristics of what makes a good simulated work task situation: (1) the situation has to be of a nature to which the test persons can *relate* and in which they

can *identify* themselves; (2) the *topic* of the situation has to be of *interest* to the group of test persons; and (3) the situation has to provide enough imaginative *context* in order for the test persons to be able to apply the situation. Characteristics which are of importance for the tuning of work tasks towards the group of test persons.

3.4.3 Research Question No. 3: The Pattern of Assessment Behaviour Between the Test Persons

This question is of interest due to indicative findings in the earlier mentioned feasibility study [10]. In the study it was found that a consistent pattern of assessment behaviour existed between the test persons as to the calculated level of precision of their relevance assessments and the corresponding order of search jobs. If there exists a consistent pattern of assessment behaviour, this is yet another reason that search jobs should be permuted between the test persons, in order to avoid biasing the retrieval results. It is another reason, in addition to the possibilities of the test persons gaining system and domain knowledge during the experiment.

Precision is computed as the sum of the number of relevant assigned documents (according to the relevance categories: low, medium, and high relevance) divided by the total number of relevance assessed documents. By precision priority we refer to the ranked position of the calculated precision values which the various search jobs have resulted in. Table 1 shows in a matrix format the accumulated precision priorities for each search job, listed according to the order of the search jobs (sim + real).

Order of search job:	1st priority	2nd priority	3rd priority	4th priority	5th priority
1	5	8	3	3	5
2	6	7	3	6	2
3	2	5	7	6	4
4	6	3	4	3	8
5	5	2	8	7	2

Table 1: The distribution of the no. of search jobs with reference to the degree of the calculated precision values and the corresponding order of the search jobs (sim + real).

In table 1, we see that search job no. 1 is primarily assigned second priorities. Just as second priorities also dominates the 2nd search job, though accompanied by first and fourth priorities. Third and fourth priorities are mainly assigned the 3rd search job. The 4th search job is dominated by fifth priorities, and finally the 5th search job is dominated by third and fourth priorities. The assessment behaviour of the test persons results in decreasing priorities which correspond to the decreasing order of search jobs, until the final search job where the priorities increase. Table 2, summarises the results of table 1. The dominating priorities are emphasised in bold. It is interesting to note that regardless the order of execution of the personal needs the pattern is the same.

Order of search job:	Priority of relevance assessments:
1	**2nd**
2	1st, **2nd**, and 4th
3	**3rd** and 4th
4	**5th**
5	**3rd** and 4th

Table 2: Assessment pattern of search jobs (sim + real).

The revealed pattern can be seen as an illustration of the test persons' motivation and concentration during the experiment – despite that the experiment was carried out over two sessions. A result of how motivation and concentration of the test persons influence their relevance assessment behaviour. The hypothesis could be that the test persons start the experiment highly motivated and concentrated. (They are positive and they want to do well. The positive attitudes and their own interest in doing well may make them easier to be pleased by the retrieved documents, and consequently they make more approvals). Later on in the experiment they may loose motivation and concentration. (They get tired and bored, and do no longer feel nervous and insecure about the experiment and their own performance. This results in a lower approval behaviour). Finally, in the end the test

persons regain concentration and motivation. (They do this because they know the experiment will soon be over, and because of that they want to finalise the experiment with an honest contribution).

Based on these findings we conclude that there does exist a general pattern of assessment behaviour for the test persons. The general pattern has implications for the application of work tasks in connection with the evaluation of IIR systems, which means that the order of work tasks must be permuted between the involved test persons. Permutation of work tasks is mandatory in order to neutralise the effect the assessment behaviour might have on the retrieval results.

4 Summary and Conclusions

Recently the work task concept has been introduced and applied as a core component in relation to evaluation of IIR systems. This paper reports on the empirical results of the evaluation of the application of work tasks in connection with the evaluation of IIR systems. The purpose of the meta-evaluation is to uncover if the application of work tasks is recommendable for future evaluation of IIR systems. We try to answer the following three research questions:

1) Can one substitute real information needs with simulated information needs?
2) What makes a 'good' work task with reference to semantic openness and types of topics?
3) Does there exist a pattern of assessment behaviour between the group of test persons, as to the order of search jobs?

With reference to research question no. 1 it is concluded that one can substitute real information needs with simulated information needs through the application of simulated work task situations; and that one can mix simulated and real information needs.

As for research question no. two, it is concluded that it makes no difference to the test persons and their search behaviour if the applied work tasks are constituted of both the 'simulated work task situation' and the 'indicative request' (sim-1) or just of the 'simulated work task situation' (sim-2) which is semantically more open. Further it is found that sim A and sim B are 'good' examples of work tasks as opposed to sim C and sim D. These findings were supported by qualitative data on the test persons' preferences of work tasks. 79.2% of the test persons liked sim A ('graduate') the best, and 45.8% liked sim D ('burglary') the least. In addition, the test persons' comments uncover three main characteristics of good simulated work task situations, namely: (1) the situation has to be of a nature to which the test persons can *relate* and in which they can *identify* themselves; (2) the *topic* of the situation has to be of *interest* to the group of test persons; and (3) the situation has to provide enough imaginative *context* in order for the test persons to be able to apply the situation.

The analysis of the third and final research question lead to the verification of a general pattern of assessment behaviour for the group of test persons. It is furthermore pointed out that the existence of this pattern has implications for the application of work tasks. It implies that the order of work tasks must be permuted between the involved test persons. Permutation of work tasks is mandatory in order to neutralise the effect the assessment behaviour might have on the retrieval results.

The overall conclusion is that the work task as an experimental component is recommendable for future evaluation of IIR systems.

5 Acknowledgement

The authors wish to thank Professor Keith van Rijsbergen, and Doctoral students Ian Ruthven and Jane Reid, all from University of Glasgow, for their support and help with the execution of the experiment. The authors also wish to thank the three anonymous MIRA reviewers for their helpful and constructive comments.

6 References

1 Allen, B.L. Information tasks: Towards a user-centered approach to information systems. San Diego, Academic Press. 1996.
2 Diaper, D. Task observation for human computer interaction. In: Diaper, D. ed. Task analysis for human-computer interaction. Chichester, Ellis Horwood. 1989, 210-237.
3 Rasmussen, J., Pejtersen, A.M. and Goodstein, L.P. Cognitive systems engineering. N.Y., John Wiley & Sons. 1994.

4 Mulhem, P. and Nigay, L. Interactive information retrieval systems: From user centered interfaces design to software design. In: Frei, H-P., Harman, D., Schäuble, P. and Wilkinson, R. eds. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Zurich, Switzerland, 1996. ACM Press. 1996, 326-334.

5 Borgman, C. All users of information retrieval systems are not created equal: An exploration into individual differences. Information Processing & Management, (25)3, 1989, 237-252.

6 Kuhlthau, C.C. Seeking meaning: a process approach to library and information science. Norwood, NJ: Ablex Publishing, 1993.

7 Byström, K. and Järvelin, K. Tasks complexity affects information seeking and use. Information Processing & Management, (31)2, 1995, 191-213.

8 Ingwersen, P. Information retrieval interaction. London: Taylor Graham, 1992.

9 Brajnik, G., Mizzaro, S. and Tasso, C. Evaluating user interfaces to information retrieval systems: A case study on user support. In: Frei, H-P., Harman, D., Schäuble, P. and Wilkinson, R. eds. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Zurich, Switzerland, 1996. ACM Press. 1996, 128-136.

10 Borlund, P. and Ingwersen, P. The development of a method for the evaluation of interactive information retrieval systems. Journal of Documentation, (53)3, 1997, 225-250.

11 Jose, J.M., Furner, J. and Harper, D.J. Spatial querying for image retrieval. In: Croft, B.C, Moffat, A., van Rijsbergen, C.J., Wilkinson, R. and Zobel, J., eds. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Melbourne, Australia, 1998. ACM Press. 1998, 232-240.

12 Reid, J. A new, task-oriented paradigm for information retrieval: implications for evaluation of information retrieval systems. 1998, 11 pages. (To be presented at the CoLIS 3 conference (Concepts of Library and Information Science) in Dubrovnik, Croatia, May 1999).

13 Robertson, S.E. and Hancock-Beaulieu, M.M. On the evaluation of IR systems. Information Processing & Management, (28)4, 1992, 457-466.

14 Spink, A., Greisdorf, H., and Bateman, J. From highly relevant to not relevant: examining different regions of relevance. Information Processing & Management, (43), 1998, 599-621.

15 Campbell, I. RdSystem technical notes. University of Glasgow, Department of Computing Science, Glasgow. 1990. [Unpublished].

16 Morita, M. & Shinoda, Y. Information filtering based on user behaviour analysis and best match text retrieval. In: Croft, W.B. and van Rijsbergen, C.J., eds. Proceedings of the 17th ACM Sigir Conference on Research and Development of Information Retrieval. Dublin, 1994. London: Springer, 1994, 272-281.

Appendix 1 (1/2)

The two versions, sim-1 and sim-2, with reference to the five topically different simulated situations: sim A, B, C, D, and the Training task situation, as applied in the meta-evaluation.

Simulated situation A	
Sim-1	**Sim-2**
Simulated work task situation: After your graduation you will be looking for a job in industry. You want information to help you focus your future job seeking. You know it pays to know the market. You would like to find some information about employment patterns in industry and what kind of qualifications employers will be looking for from future employees. **Indicative request:** Find for instance something about future employment trends in industry, i.e. areas of growth and decline.	**Simulated work task situation:** After your graduation you will be looking for a job in industry. You want information to help you focus your future job seeking. You know it pays to know the market. You would like to find some information about employment patterns in industry and what kind of qualifications employers will be looking for from future employees.

Simulated situation B	
Sim-1	**Sim-2**
Simulated work task situation: You have got a new student job with a local branch of one of the wildlife protection organisations. Your responsibility is to maintain and update the web pages of the organisation. You have been informed that the organisation's next big campaign will be on how to prevent the decline of wildlife species, focusing on the situation in Europe. As a new member of staff you feel you need some basic background information so you have decided to investigate the European situation with particular reference to problems caused by environmental and climate changes. **Indicative request:** Find for instance something about the steps taken to save species of wildlife.	**Simulated work task situation:** You have got a new student job with a local branch of one of the wildlife protection organisations. Your responsibility is to maintain and update the web pages of the organisation. You have been informed that the organisation's next big campaign will be on how to prevent the decline of wildlife species, focusing on the situation in Europe. As a new member of staff you feel you need some basic background information so you have decided to investigate the European situation with particular reference to problems caused by environmental and climate changes.

Appendix 1 (2/2)

Simulated situation C

Sim-1	Sim-2
Simulated work task situation: Some friends of yours are about to visit you and as a surprise you are planning a trip for all of you to the Isle of Arran. You have heard rumours that some of the ferries to Arran are less safe than others. In addition to this you have recently seen the movie Titanic. You would therefore like to retrieve information about the causes of safety problems on ferries as well as some information about how to prevent accidents. **Indicative request:** Find for instance something about ferry sinking.	**Simulated work task situation:** Some friends of yours are about to visit you and as a surprise you are planning a trip for all of you to the Isle of Arran. You have heard rumours that some of the ferries to Arran are less safe than others. In addition to this you have recently seen the movie Titanic. You would therefore like to retrieve information about the causes of safety problems on ferries as well as some information about how to prevent accidents.

Simulated situation D

Sim-1	Sim-2
Simulated work task situation: There has been a burglary in your flat. Among the things stolen was an old and unique piece of jewellery with a high value of affect. You called the police, who were not very hopeful of getting the jewellery back. They said that there had been several such burglaries in the areas within the previous few months. You're interested in finding out about similar cases and more specifically the details and the consequences of the crimes. **Indicative request:** Find for instance something about the buying or selling of stolen objects.	**Simulated work task situation:** There has been a burglary in your flat. Among the things stolen was an old and unique piece of jewellery with a high value of affect. You called the police, who were not very hopeful of getting the jewellery back. They said that there had been several such burglaries in the areas within the previous few months. You're interested in finding out about similar cases and more specifically the details and the consequences of the crimes.

Simulated training situation (tr)

Sim-1	Sim-2
Simulated work task situation: The other night you were at a party where the Hubble Telescope was discussed as one of the other guests knew quite a lot about this subject. Now you want to improve your own knowledge of this topic and more specifically you want to know about the Telescope's technical drawbacks and scientific achievements. **Indicative request:** Find for instance something about the launch of the Hubble Telescope in 1991.	**Simulated work task situation:** The other night you were at a party where the Hubble Telescope was discussed as one of the other guests knew quite a lot about this subject. Now you want to improve your own knowledge of this topic and more specifically you want to know about the Telescope's technical drawbacks and scientific achievements.

Chi-square tests

Hypothesis: H_0: No difference exists between **real information needs** *versus* **simulated information needs** when measure in relation to...

H_a: Difference exists between **real information needs** *versus* **simulated information needs** when measure in relation to...

Search attributes:	$\alpha^2 c$	X^2
number of search runs per task	6.64	0.09
average number of search terms per task	6.64	1.93
number of unique search terms per task per test person	6.64	0.41
search time spent per task	6.64	2.84
distribution of applied relevance categories: low, medium, and high	9.21	0.15
distribution of applied relevance categories: non-relevant, low, medium, high, and relevant.	13.28	0.37
number of relevance assessments, distributed on applied source (full-text/title)	6.64	15.02

Hypothesis: H_0: No difference exists between **sim-1** *versus* **real information needs** when measure in relation to...

H_a: Difference exists between **sim-1** *versus* **real information needs** when measured in relation to...

Search attributes:	$\alpha^2 c$	X^2
number of search runs per task	6.64	0.19
average number of search terms per task	6.64	0.79
number of unique search terms per task per test person	6.64	0.82
search time spent per task	6.64	1.37
distribution of applied relevance categories: low, medium, and high	9.21	0.27
distribution of applied relevance categories: non-relevant, low, medium, high, and relevant.	13.28	0.63
number of relevance assessments, distributed on applied source (full-text/title)	6.64	11.1

Appendix 2 (1/4)

Hypothesis: H_0: No difference exists between **sim-2** *versus* **real information needs** when measure in relation to...
H_a: Difference exists between **sim-2** *versus* **real information needs** when measured in relation to...

Search attributes:	α^2c	X^2
number of search runs per task	6.64	0.06
average number of search terms per task	6.64	0.19
number of unique search terms per task per test person	6.64	0.11
search time spent per task	6.64	3.81
distribution of applied relevance categories: low, medium, and high	9.21	0.22
distribution of applied relevance categories: non-relevant, low, medium, high, and relevant.	13.28	0.94
number of relevance assessments, distributed on applied source (full-text/title)	6.64	13.83

Hypothesis: H_0: No difference exists between **sim-1** *versus* **sim-2** when measure in relation to...
H_a: Difference exists between **sim-1** *versus* **sim-2** when measured in relation to...

Search attributes:	α^2c	X^2
number of search runs per task	6.64	0.21
average number of search terms per task	6.64	0.42
number of unique search terms per task per test person	6.64	0.74
search time spent per task	6.64	1.14
distribution of applied relevance categories: low, medium, and high	9.21	0.74
distribution of applied relevance categories: non-relevant, low, medium, high, and relevant.	13.28	2.86
number of relevance assessments, distributed on applied source (full-text/title)	6.64	0.29

Appendix 2 (2/4)

Hypothesis: H_0: No difference exists between **sim A** *versus* **real information needs** when measure in relation to...
H_a: Difference exists between **sim A** *versus* **real information needs** when measured in relation to...

Search attributes:	$\alpha^2 c$	X^2
number of search runs per task	6.64	0.83
average number of search terms per task	6.64	9.19
number of unique search terms per task per test person	6.64	0.09
search time spent per task	6.64	0.56
distribution of applied relevance categories: low, medium, and high	9.21	0.11
distribution of applied relevance categories: non-relevant, low, medium, high, and relevant.	13.28	1.35
number of relevance assessments, distributed on applied source (full-text/title)	6.64	1.85

Hypothesis: H_0: No difference exists between **sim B** *versus* **real information needs** when measure in relation to...
H_a: Difference exists between **sim B** *versus* **real information needs** when measured in relation to...

Search attributes:	$\alpha^2 c$	X^2
number of search runs per task	6.64	0.17
average number of search terms per task	6.64	6.94
number of unique search terms per task per test person	6.64	0.18
search time spent per task	6.64	0.13
distribution of applied relevance categories: low, medium, and high	9.21	0.13
distribution of applied relevance categories: non-relevant, low, medium, high, and relevant.	13.28	0.31
number of relevance assessments, distributed on applied source (full-text/title)	6.64	1.85

Appendix 2 (3/4)

Hypothesis: H_0: No difference exists between **sim C** *versus* **real information needs** when measure in relation to...
H_a: Difference exists between **sim C** *versus* **real information needs** when measured in relation to...

Search attributes:	α^2c	X^2
number of search runs per task	6.64	0.44
average number of search terms per task	6.64	0.18
number of unique search terms per task per test person	6.64	0.19
search time spent per task	6.64	5.78
distribution of applied relevance categories: low, medium, and high	9.21	.023
distribution of applied relevance categories: non-relevant, low, medium, high, and relevant.	13.28	0.22
number of relevance assessments, distributed on applied source (full-text/title)	6.64	29.67

Hypothesis: H_0: No difference exists between **sim D** *versus* **real information needs** when measure in relation to...
H_a: Difference exists between **sim D** *versus* **real information needs** when measured in relation to...

Search attributes:	α^2c	X^2
number of search runs per task	6.64	0.09
average number of search terms per task	6.64	0.17
number of unique search terms per task per test person	6.64	0.09
search time spent per task	6.64	4.55
distribution of applied relevance categories: low, medium, and high	9.21	0.02
distribution of applied relevance categories: non-relevant, low, medium, high, and relevant.	13.28	0.91
number of relevance assessments, distributed on applied source (full-text/title)	6.64	41.28

Appendix 2 (4/4)