

Ordinary Search Engine Users assessing Difficulty, Effort, and Outcome for Simple and Complex Search Tasks

Georg Singer, Ulrich
Norbisrath
Institute of Computer Science,
University of Tartu
Tartu, Estonia
{First.Last}@ut.ee

Dirk Lewandowski
Hamburg University of Applied
Sciences
Hamburg, Germany
Dirk.Lewandowski@haw-
hamburg.de

ABSTRACT

Search engines are the preferred tools for finding information on the Web. They are advancing to be the common helpers to answer any of our search needs. We use them to carry out simple look-up tasks and also to work on rather time consuming and more complex search tasks. Yet, we do not know very much about the user performance while carrying out those tasks – especially not for ordinary users. The aim of this study was to get more insight into whether Web users manage to assess difficulty, time effort, query effort, and task outcome of search tasks, and if their judging performance relates to task complexity. Our study was conducted with a systematically selected sample of 56 people with a wide demographic background. They carried out a set of 12 search tasks with commercial Web search engines in a laboratory environment. The results confirm that it is hard for normal Web users to judge the difficulty and effort to carry out complex search tasks. The judgments are more reliable for simple tasks than for complex ones. Task complexity is an indicator for judging performance.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval - *search process*

H.3.7 [Information Storage and Retrieval]: Digital Libraries - *user issues*

1. INTRODUCTION

People use search engines for all kinds of tasks, from simply looking up trivia to planning their holiday trips. While looking up dates and facts usually is an endeavor limited in time and also effort, more complex tasks usually can take much longer than expected. Carrying out those tasks with current search engines might also cause much more effort than expected. This disparity between expected and real effort for such tasks is partly due to search engines not sup-

porting those types of tasks explicitly [24], but also due to users having little understanding about different task types.

Singer et al. [15] have taken Marchionini's definition of exploratory search [14] and focus especially on the search concepts discovery, aggregation and synthesis. Their idea is that these are the most time consuming activities and they also cause the most search effort when fulfilling an information need and make a search task complex. They define complex search tasks as at least requiring one of the elements aggregation (finding several documents to a known aspect), discovery (detecting new aspect), and synthesis (synthesizing the found information into a single document). Complex tasks typically require going through those steps multiple times. [23].

According to a key note speech with the title "Search isn't Search" by Stefan Weitz (Microsoft) given at the SMX Conference 2009 [21], only 1 in 4 queries are successful and many queries yield terrible satisfaction. Many search queries are actually not isolated efforts towards finding a single fact but instead are part of sessions, close to 50% of sessions are longer than 1 week and people are increasingly using search to make decisions (66% of search users).

The above mentioned dissatisfaction of users in terms of unsuccessful queries might partly be caused by users not being able to judge the task effort properly and therefore their experience not being in line with their expectations. The problem is that little research exists about a reasonably big sample of ordinary Web search engine users carrying out complex search tasks and examining their ability to judge task effort. In this paper, we present a study in which a larger number of ordinary Web search engine users carried out selected simple and complex search tasks in a laboratory environment. Before each task we asked the users to rate their expectations regarding task difficulty, task effort and task outcome. Then we had them carry out the tasks and after each task asked them to do the rating again, this time them being aware of the real effort. We were especially interested, if the judging performance varied between simple and complex tasks and whether good searchers were also better judges. Finally we investigated whether the judging performance depended on task complexity or rather simply on the individual person.

The rest of this paper is structured as follows: First, we review the literature on users estimating the complexity of search tasks, followed by studies on simple and complex search (and therefore also covering exploratory search). Then, we give some necessary definitions and state our re-

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, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

IiiX'12 Nijmegen, The Netherlands.

Copyright 2012 ACM 978-1-4503-1282-0/2012/08 ...\$15.00.

search questions. After that, we describe our methods, followed by the results. These are discussed, and in the conclusions section we sum up the outcomes and limitations of our research and give some directions for future research.

2. RELATED WORK

According to Li, “Tasks are activities people attempt to accomplish in order to keep their work or life moving on” ([11], p. 1823). Usually tasks have an ultimate goal and information searching is an activity to find relevant information to achieve that goal [19]. Information searching can be the result of an interest or of a work task. A work task is a task that appears in the work context. Its goal is work related [11]. Work tasks can be the origin for information-seeking tasks and information search tasks [3, 4, 8, 2, 1]. Li [11] defines information-seeking tasks as being related to people’s general information needs. Such needs can be satisfied by searching through multiple sources, including books in libraries, papers and also digital information systems. Once people start searching with information systems, the information-seeking task usually becomes a search task. A family might for example be faced with the task to plan a holiday trip. Resulting out of this work task, the search task, to use the Internet and Web search engines to find children-friendly hotels at a certain destination might arise.

Task complexity can be either objective or subjective [13]. As far as information science is concerned, objective task complexity is poorly researched. According to Li and Belkin [12], task complexity relates to the number of sub-tasks that need to be carried out. Subjective task complexity reflects how complex the person, who carries out the task, sees it [13]. Byström and Järvelin [5] have developed a task categorization accounting for task complexity from an automatic information processing task to a genuine decision task. According to Byström and Järvelin task complexity is mainly defined by users having to deal with “a priori determinability of, or uncertainty about, task outcomes, process, and information requirement” ([5], p. 194). They state three types of information needs in tasks: problem information (specific requirements of the problem dealt with), domain information (facts, concepts, rules and laws about the domain the problem is located in) and problem solving information (known methods to tackle this problem). Their findings show that for automatic information processing tasks both the level of motivation to carry out the task is high (as people are quite sure they will be able to solve it) and also only problem information is needed. In known, genuine decision tasks (which was the highest level of complexity they investigated in this experiment), the level of ambition is also high (and higher than expected). They credit this to the level of education of their study participants. What clearly distinguished this task from the simple one, was the level of problem solving information, which was required in terms of systems to use and experts to ask. Bell and Ruthven [2] carried out a user study with 30 people who were asked to work on three groups of search tasks (tasks organized in three complexity levels) and afterwards rate the complexity of each task on a 5-point scale. The goal was to test the above described model by Byström and Järvelin. They observed that assessment of completion and task complexity were inversely correlated. The more complex people perceived a task, the less confident they felt, when they completed that task. In addition they found that

a task is perceived as more complex if the task contains little information about what information is needed and what amount of information should be retrieved. Also subjective factors like previous knowledge about topics related to that task had to be taken into account as influencing factors for the perception of complexity. Gwizdka and Spence [7] conducted a study with 27 undergraduate psychology students in which they were required to fulfill a look-up task. They examined the relationship between searcher’s activities and subjective post-task difficulty and finding predictors for subjective task complexity. They found that task time, time per click, pages visited, unique pages visited, revisit ratio and back-button use were good predictors for subjective task complexity. White and Livonen [22] conducted a study with 54 experienced Web searchers and had them rate 16 search questions regarding complexity. Their results show that users perceive closed/predictable source questions easy, open/unpredictable source questions difficult. In addition the study participants agreed that “searchability, clarity, familiarity/currency, public knowledge, simplicity, and specificity” were important aspects that made a task either simple or complex. Li et al. [13] conducted a survey containing 100 university students in China. They observed that objective task complexity measures were more indicative for task complexity than subjective ones. The main objective predictors for task complexity were: number of words in the task description, number of languages needed to interpret search results and the number of domain areas, that the task involved. In addition the objective complexity criteria were more helpful to predict complexity. In the information science community the two concepts complexity and difficulty are sometimes used as being identical and sometimes they are used as being different. Gwizdka [6] has done a question-driven, web-based information search study with 48 participants (students, mean age 27 years) aimed at understanding the cognitive load when carrying out web search tasks (recording them and analyzing their respective actions). The study participants were required to carry out a primary task and in parallel a secondary task to measure their cognitive load on the primary task. The results confirm that subjective task difficulty and objective difficulty are in line and that study participants tended to underestimate task difficulty. The author credited this to the high degree of Internet search experience among the participants and their relatively young age. The study also shows that subjective difficulty was more strongly related with user effort than objective difficulty. The author interprets this as the subjective difficulty more truly showing the searcher’s cognitive effort. Gwizdka’s definition of task difficulty overlaps with the common definition of subjective task complexity. Vakkari and Huuskonen [20] conducted a study with 41 medical students to investigate how the search effort impacted search output and task outcome. They found that in case of bad retrieval results, the study participants worked harder to achieve desired task outcomes. They conclude that measures for search process and task outcome need to be added to classic IR measures.

3. DEFINITIONS

Objective vs. subjective

As we have seen, researchers use different concepts to distinguish simple tasks from more complex ones, either describing

them as “complex” or “difficult”, sometimes interchangeably. To help guide the reader of this paper, we give some definitions, which we will use throughout this paper.

A *task* is an abstract description of activities to achieve a certain goal [9, 11].

Search is the process of finding information.

A *search task* is a piece of work concerning the retrieval of information related to an information need. The search is usually carried out with IR systems [11].

A search task is *complex* if it requires at least one of the elements aggregation, discovery and synthesis [15]. It typically requires reviewing many documents and synthesizing them into a desired format.

A search task is *difficult* if a lot of cognitive input is needed to carry out the task.

A search task requires increased *effort* if the user needs either more cognitive effort to understand the task and formulate queries (time effort), or more mechanical effort (number of queries, number of pages visited, browser tabs opened and closed).

4. RESEARCH QUESTIONS

In the study task complexity is the independent variable. Search performance, searchers’ assessments of task difficulty, query effort, time effort and outcome are dependent variables.

To guide our research, we formulated the following research questions:

RQ1: Can users assess difficulty, effort and task outcome for simple search tasks?

RQ2: Can users assess difficulty, effort and task outcome for complex search tasks?

RQ3: Are there significant performance differences between assessing simple and complex search tasks?

RQ4: Does the users’ ability to judge if the information they have found is correct or not depend on task complexity?

RQ5: Is there a correlation between the overall search performance (ranking in the experiment) and the ability to assess difficulty, time effort, query effort, and task outcome for complex tasks?

Here we investigate the correlation between search performance and judging performance.

RQ6: Does the judging performance depend on task complexity or simply the individual user?

This question investigates the association between judging performance, task complexity and the individual user.

5. RESEARCH METHOD

The results presented in this paper are based on a body of data gathered in the course of a larger experiment in August 2011. One additional article using distinct parts of the data and describing different aspects has been published [18], and one article has been submitted for review [16]. The following description of the research design is based on Singer et al. [16]. The experiment was conducted in August 2011 in

Basic Data	Gender		
Age Span	Female	Male	Total
18-24	5	4	9
25-34	9	7	16
35-44	7	8	15
45-54	8	8	16
55-59	3	1	4
Total	32	28	60

Table 1: Demography of user sample

Hamburg, Germany. Participants were invited to the university, where they were given a set of search tasks (see below) to fulfill. The study was carried out in one of the university’s computer labs, where each participant had her own computer and was instructed to work on the search tasks independently. Participants were not observed directly, but their browser interactions were recorded using the Search-Logger plug-in [17]. While the tasks were presented in a certain order to the participants, they were allowed to choose the order of the tasks according to their wishes, and it was also possible to interrupt a task, work on another one, and later return.

We recruited a sample of 60 volunteers, using a demographic structure model. The aim was to go beyond the usual user samples consisting mainly of students, often experienced searchers from information science or computer science, and also, to increase the sample size. As a user sample of the intended size could not be representative, we wanted at least to make sure that adults from various age ranges, and also men and women alike, were considered. For details on the sample, see Table 1. The effective number of study participants providing data to our study was reduced to 56, as the data of 4 (2 females, 2 males) out of the 60 users was corrupt and could therefore not be used.

The search experiment consisted of 12 search tasks. As our experiment was conducted in Germany, the language of the tasks was German. A prerequisite for all tasks was that a correct answer had to be available somewhere in public websites in German as of August 2011. The study participants had 3 hours to complete the experiment.

Simple tasks are characterized by asking the users to find simple facts. The needed information is contained in one document (web site) and can be retrieved with one single query. Complex tasks on the other hand are formulated in a way that the users have enough context to comprehend the task situation but the tasks are still characterized by uncertainty and ambiguity [10]. There is no single correct answer retrievable and the required information is spread over various documents (web sites). Fulfilling the task typically requires issuing multiple queries, aggregating information from various sources and synthesizing the information into a single solution document [15]. The tasks were as follows ((S) marks simple, and (C) complex tasks):

1. **(S)** When was the composer of the piece “The Magic Flute” born?
2. **(S)** How hot can it be on average in July in Aachen/Germany?
3. **(S)** How many opera pieces did Verdi compose?
4. **(S)** When and by whom was penicillin discovered?

5. (S) How many Euros do you get if you exchange 10.000 units of the currency of Lithuania?
6. (S) Joseph Pulitzer (1847-1911) was a well-known journalist and publisher from the U.S. The Pulitzer Prize carries on his name. In which European country was Pulitzer born?
7. (C) How high is the state debt of Italy in comparison to their gross domestic product (GDP) in June 2011 in %?
8. (C) What are the most important five points to consider if you want to plan a budget wedding?
9. (C) You were offered the job to run a local Goethe Institute (responsible for German language and cultural education) abroad. The chance is high that you will be sent to Astana (Kazakhstan). Please collect facts and information (about half a page) about the political situation in Kazakhstan and the living quality.
10. (C) What is the name of the creature on the following picture and who is the author? Hint: this Austrian writer is also well known in Germany. (Illustration omitted for copyright reasons)
11. (C) Are there differences regarding the distribution of religious affiliations between Austria, Germany, and Switzerland? Which ones?
12. (C) There are five countries whose names are also carried by chemical elements. France has two (31. Ga – Gallium and 87. Fr – Francium), Germany has one (32. Ge – Germanium), Russia has one (44. Ru – Ruthenium) and Poland has one (84. Po – Polonium). Please name the fifth country.

We arranged the task sequence so that users could alternatively solve simple and complex ones and they could always switch between tasks if they wanted. The aim was to keep the participants interested, and to not discourage participants through a series of complex search tasks, which they might be unable to solve.

We implemented users' judgments as binary responses to questionnaire items. We added questionnaires before starting and after finishing each search task. Prior to each task we used the following statements that users could rate with yes or no: 1) This task is easy 2) It will take me less than 5 minutes to complete the task 3) I will need fewer than 5 queries to complete the task 4) I will find the correct information. After the participants had completed that task we asked them to rate the following statements with yes or no: 1) The task was easy 2) I took me less than 5 minutes to complete the task 3) It needed fewer than 5 queries to complete the task 4) I have found the correct information.

We understood effort as comprising the cognitive effort to understand the task and formulate queries (time) and also the mechanical effort to carry out the task (number of queries). This is different than in Jacek Gwizdka [6]. In his study, he measures the cognitive load or mental effort. For our study, measuring cognitive effort directly was not feasible, therefore we measured the indicators time effort and query effort.

We compared users' subjective values for the question whether they thought in advance they would find the correct information (yes or no) with the objectively graded outcome that they submitted. The objective result is a manual review of all the answers given by the participants of our study. The solutions the study participants provided were bench-marked against this optimal solution developed by the researchers on a scale, correct, partly correct, wrong, no solution submitted. In case of simple tasks, each task had exactly one solution. If the solution was correct and complete, the task was graded correct. If it was correct and not complete (like only mentioning who invented Penicillin but not when in above mentioned Task 4), it was graded "partly correct". If the solution was wrong, it was graded "incorrect". In the case of complex tasks, it was less trivial. As the tasks were quite open, there was no single right or wrong solution possible. If the solution provided by the users covered all aspects that the optimal solution also contained it was marked "correct". If the solution covered fewer aspects it was graded "partly correct". If the solution did not cover any aspects the solution was "incorrect". For both, simple and complex tasks, if no solution was submitted, the task was graded "unanswered".

To understand the relation between search performance (ranking of the user in the experiment) and the ability to estimate task difficulty, task effort, and task outcome, we ordered the users according to their ranking in the experiment. We ranked the users first by the number of correct answers and then, in cases of users with the same number of correct answers, by "partly correct" answers.

We ran paired-sample t-tests (assuming unequal variances) to analyze the statistical significance of our results for RQ3- RQ5.

6. RESULTS

In this section we present the results of our study, which are also used to answer the research questions stated in Section 4. We first show, how ordinary Web search engine users manage to judge difficulty, effort and task outcome for simple tasks. Next we present the results of the same analysis but this time applied to complex tasks. Then we highlight the differences in judging performance between simple and complex tasks. We show, how task complexity impacts the ability of the users to judge whether the information they have found is correct or not. Next we examine, if better searchers are better at judging than worse searchers. Finally we investigate whether the judging performance depends on the task complexity or the individual user only.

It is important to point out that the numbers that we are presenting here are subjective difficulty, subjective effort, and subjective ability to find the correct information. Subjective here means the individual judgment of the searcher in answering a question of our pre-task and post-task questionnaires. If a task was subjectively difficult, this does not necessarily mean that this would also be the case on an objective level as outlined in the related work section [5].

RQ1: Can users assess difficulty, effort and task outcome for simple search tasks?

Table 2 outlines our findings regarding users and their ability to estimate difficulty, effort (in terms of time used and queries entered) and being able to find the correct result for

	# of tasks	%
difficulty		
incorrect	29	9.8
correct	266	90.2
time effort		
incorrect	27	9.1
correct	268	90.8
query effort		
incorrect	38	12.9
correct	257	87.1
ability to find right result		
incorrect	16	5.4
correct	279	94.6

Table 2: Users judging simple search tasks

simple search tasks. “# of tasks” represents the number of simple tasks that have been processed by the study participants. The total number of tasks (correct plus incorrect ones) should have been $56 \times 6 = 336$ (56 valid users x 6 tasks). It is slightly lower due to invalid or not given answers or not fulfilled tasks. % shows the percentage of the number of judged tasks to the total valid answers for tasks. We graded an answer as correct when the users’ self-judged values were the same in the pre-task questionnaire and the post-task questionnaire. For example if they judged a task to be difficult in the pre-task questionnaire and after carrying out the task stated again that it was a difficult task, the judgment was graded as correct.

For all parameters (difficulty, time effort, query effort, and result finding ability) approximately 90% of the users managed to match estimated and experienced values for simple tasks. However, in our study the users had slightly more trouble estimating the time effort needed than the query effort (in terms of estimating going over a threshold of numbers of queries).

RQ2: Can users assess difficulty, effort and task outcome for complex search tasks?

Table 3 outlines our findings on users and their ability to estimate difficulty, effort, and being able to find the correct result for complex search tasks. For all parameters about 70% of all tasks were judged correctly with slightly worse estimations for query effort and result-finding skills.

RQ3: Are there significant differences between assessing simple and complex search tasks?

To analyze if there exist significant differences between how users assess certain parameters (difficulty, time effort, query effort, search success), we have compared the differences between users’ pre-task estimate and post-task experience based values for 295 simple tasks and 286 complex tasks as outlined in Table 4 (100% means a 100% probability to judge the right difficulty; pre-task estimate and post task evaluation are totally in line). We used paired sample t-tests to compare the results from simple and complex tasks. We paired average difficulty for simple tasks and average diffi-

	# of tasks	%
difficulty		
incorrect	95	33.2
correct	191	66.8
time effort		
incorrect	99	34.6
correct	187	65.3
query effort		
incorrect	91	31.8
correct	195	68.2
ability to find right result		
incorrect	78	27.2
correct	208	72.8

Table 3: Users judging complex search tasks

	difficulty (%)	time effort (%)	query effort (%)	task outcome (%)
Simple tasks (n=295)	90±2	91±2	87±2	95±1
Complex tasks (n=286)	67±3	65±3	68±3	73±3
p-value	<0.001	<0.001	<0.001	<0.001

Table 4: Correctly judged tasks per dependent variable (mean values over tasks)

culty for complex tasks. We followed the same procedure for time effort, query effort and task outcome.

In the case of simple tasks, users are significantly better at estimating all four parameters: difficulty, time effort, query effort and search success, i.e. the difference between their pre-task estimate and their post-task experience based value for a certain parameter is significantly smaller and they show a higher probability to correctly judge the parameter. Table 5 shows the judging performance for all users for each specific task. Also here the division in terms of performance between simple (S) and complex (C) tasks is clearly visible.

RQ4: Does the users’ ability to judge if the information they have found is correct or not depend on task complexity?

We compared users’ subjective values for the question whether they thought they could/would find the correct information (yes or no) with the objectively graded outcome (by building the difference of the two submitted values). The objective result is a manual review of all the answers given by the participants of our study. In this evaluation we skipped the tasks where a user had not delivered any result, as we did not know what was the reason for not delivering (could be not being able, found wrong results and did not want to submit, or simply forgot to submit).

We also analyzed, whether the users were able to judge the correctness of their found results after having finished

Task	difficulty (%)	time effort (%)	query effort (%)	task outcome (%)
1 (S) (n=51)	92±4	90±4	88±5	98±2
2 (S) (n=48)	81±6	85±5	73±6	88±5
3 (S) (n=47)	89±5	91±4	87±5	96±3
4 (S) (n=51)	98±2	100±0	98±3	96±3
5 (S) (n=49)	84±5	82±6	82±6	94±3
6 (S) (n=49)	96±3	96±3	94±3	96±3
7 (C) (n=47)	62±7	51±7	60±7	85±5
8 (C) (n=48)	67±7	69±7	71±7	77±6
9 (C) (n=48)	60±7	81±6	73±6	81±6
10 (C) (n=46)	72±7	67±7	72±7	52±7
11 (C) (n=49)	65±7	67±7	65±7	61±7
12 (C) (n=47)	74±6	55±7	68±7	79±6

Table 5: Fraction of users correctly judging task parameters per task

Task type	Correctly estimated tasks (%)
simple (n=259)	87±2
complex (n=233)	52±3
p-value	<0.001

Table 6: Judgments of expected search outcome (in pre-task questionnaire) compared to correctness of manually evaluated search results (mean values over tasks)

the search task. These evaluations gives us an estimate of how well users can judge that a result they found on the Internet is actually correct and how well they can judge in advance, if they will be able to find the correct result.

Table 6 shows that the ability to predict, whether it is possible to find the correct information, is significantly higher for simple search tasks than for complex search tasks, 87% versus 52% in case of complex tasks. We used paired sample t-tests to compare the results from simple and complex tasks. We paired average correctness for simple tasks and average correctness for complex tasks.

Table 7 depicts that the ability to judge whether a found information is correct or not is significantly higher in case of simple tasks than it is in case of complex tasks. Here the difference is also significant, 88% versus 60%.

The difference of observations between simple and complex tasks is due to the fact that not all users always correctly entered their rating into our system and therefore those tasks had to be omitted.

RQ5: Is there a correlation between the overall search performance (ranking in the experiment) and the ability to assess difficulty, effort and outcome for complex tasks?

As also mentioned in the methods section, to understand the relation between search performance (ranking of the user in the experiment) and the ability to estimate task difficulty, effort, and outcome, we ordered the users according to their

Task type	Correctly estimated tasks (%)
simple (n=259)	88±2
complex (n=230)	60±3
p-value	<0.001

Table 7: Assessments of self-judged search results (in post-task questionnaire) compared to correctness of manually evaluated search results (mean values over tasks)

	Avg. difficulty in %	Avg. time effort in %	Avg. query effort in %	Avg. task outcome in %
1. quartile (n=67)	67±6	64±6	67±6	85±4
4. quartile (n=59)	73±6	75±6	73±6	64±6
p-value	n.s.	n.s.	n.s.	<0.05

Table 8: Correct estimations of best and worst quartile for expected and experienced task parameters

ranking in the experiment. We ranked the users first by the number of correct answers given and then, in cases of users with the same number of correct answers, by answers with right elements (simple and complex tasks).

Then we compared the complex tasks of the first quartile of ranked users (n=67 tasks) with the complex tasks of the fourth quartile of ranked users (n=59 tasks) as outlined in Table 8. The number of tasks is different due to the fact that not all users always correctly entered their estimates into the system and therefore those tasks had to be omitted.

The results show that good searchers are not significantly better at judging difficulty and effort for complex search tasks. However, they are significantly better at judging the task outcome.

RQ6: Does the judging performance depend on task complexity or only on the individual user?

In this subsection we examine whether the judging performance depends on task complexity or simply on the individual ability to make those judgments. It could for example be that some users are very good at judging simple as well as complex tasks while others perform badly for both task groups. In an analysis that takes the average over all simple tasks and compares them with the average over all complex tasks (independent of the user e.g. as done for research question 3) this fact would not show up.

Table 9 shows the results for users judging the task difficulty for simple and complex tasks at the same time. 3 out of 53 users (6%) were able to judge the difficulty totally right for simple and complex tasks. 30 out of 53 users (56%) managed to judge the difficulty for simple tasks right and at the same time were not totally correct for the complex tasks. Three users (6%) were able to correctly judge all complex

user	simple	complex	user	simple	complex	user	simple	complex
44	100%	100%	65	100%	67%	87	83%	83%
58	100%	100%	71	100%	67%	55	83%	75%
92	100%	100%	78	100%	67%	82	83%	67%
28	100%	83%	90	100%	67%	39	83%	50%
42	100%	83%	48	100%	60%	70	83%	33%
46	100%	83%	37	100%	50%	77	83%	33%
61	100%	83%	45	100%	50%	81	80%	100%
63	100%	83%	57	100%	50%	59	80%	80%
66	100%	83%	67	100%	50%	72	80%	50%
89	100%	83%	68	100%	50%	75	67%	100%
34	100%	80%	73	100%	50%	47	67%	83%
43	100%	80%	88	100%	50%	64	67%	50%
54	100%	80%	32	100%	40%	79	67%	50%
93	100%	80%	84	100%	40%	83	67%	50%
24	100%	67%	52	100%	20%	33	60%	60%
38	100%	67%	76	83%	40%	74	50%	60%
51	100%	67%	41	83%	83%	69	40%	100%
62	100%	67%	60	83%	83%			

Table 9: Users judging the task difficulty

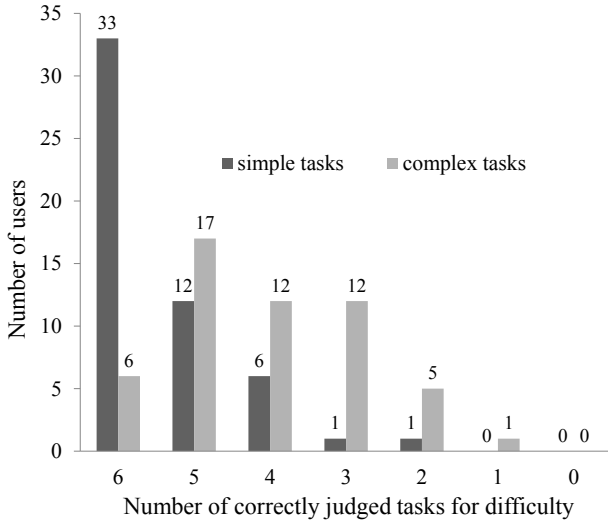


Figure 1: Users judging the task difficulty

tasks while at the same time wrongly judging a number of simple tasks.

Figure 1 illustrates a histogram of users versus correctly judged task difficulty. The x-axis shows the number of correctly carried out tasks from 6 (all) to 0 (none). The y-axis shows number of users (black bar simple tasks, grey bar complex). From this figure it is also evident that users are far better at judging the difficulty for simple tasks. 33 users of 53 (62%) have managed to correctly judge the difficulty of all simple tasks, but only 6 out 53 (11%) have managed to correctly judge the difficulty of all complex tasks.

Table 10 shows the results for users judging the time effort for simple and complex tasks. 4 out of 53 users (8%) were able to judge the time effort right for simple and complex tasks. 31 out of 53 users (58%) managed to judge the time effort for simple tasks right and at the same time were not totally correct for the complex tasks. One user was able to correctly judge all complex tasks while wrongly judging a

user	simple	complex	user	simple	complex	user	simple	complex
61	100%	100%	63	100%	67%	59	83%	83%
69	100%	100%	82	100%	67%	72	83%	83%
71	100%	100%	92	100%	67%	24	83%	80%
83	100%	100%	33	100%	60%	75	83%	67%
37	100%	83%	38	100%	60%	48	83%	50%
42	100%	83%	64	100%	60%	41	83%	40%
44	100%	83%	66	100%	60%	47	83%	33%
57	100%	83%	28	100%	50%	78	83%	25%
65	100%	83%	34	100%	50%	46	80%	80%
68	100%	83%	51	100%	50%	55	80%	75%
79	100%	83%	73	100%	50%	54	75%	60%
89	100%	83%	76	100%	50%	67	67%	80%
39	100%	80%	87	100%	50%	58	67%	40%
81	100%	75%	74	100%	33%	60	60%	100%
88	100%	75%	84	100%	33%	62	60%	83%
32	100%	67%	93	100%	33%	77	40%	60%
43	100%	67%	90	100%	17%	70	33%	50%
45	100%	67%	52	83%	83%			

Table 10: Users judging the time effort

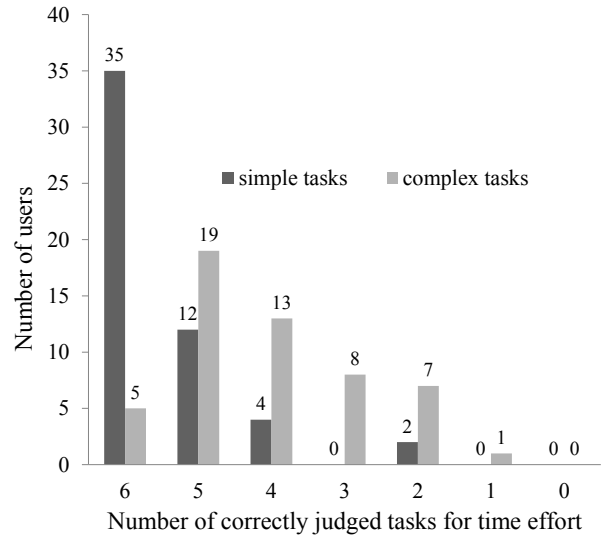


Figure 2: Users judging the time effort

number of simple tasks.

Figure 2 illustrates a histogram of users versus correctly judged time effort. From this figure it is also evident that users are far better at judging the time effort for simple tasks. 35 users of 53 (66%) have managed to correctly judge the time effort of all simple tasks, but only 5 out 53 (9%) have managed to correctly judge the difficulty of all complex tasks.

Table 11 shows the results for users judging the query effort for simple and complex tasks. 4 out of 53 users (8%) were able to judge the time effort right for simple and complex tasks. 22 out of 53 users (42%) managed to judge the query effort for simple tasks right and at the same time were not totally correct for the complex tasks. Five users (9%) were able to correctly judge all complex tasks while wrongly judging a number of simple tasks.

Figure 3 illustrates a histogram of users versus correctly

user	simple	complex	user	simple	complex	user	simple	complex
39	100%	100%	78	100%	50%	90	83%	67%
88	100%	100%	92	100%	50%	79	83%	67%
47	100%	100%	61	100%	50%	41	83%	60%
69	100%	100%	59	100%	50%	44	83%	50%
93	100%	83%	24	100%	40%	48	83%	50%
42	100%	83%	84	100%	33%	75	83%	17%
65	100%	83%	82	100%	33%	62	80%	83%
63	100%	83%	74	100%	17%	46	80%	80%
34	100%	67%	43	97%	67%	55	80%	75%
89	100%	67%	54	88%	40%	68	67%	83%
51	100%	67%	28	83%	100%	76	67%	83%
87	100%	67%	83	83%	100%	67	67%	80%
45	100%	67%	72	83%	100%	38	60%	40%
73	100%	67%	52	83%	83%	70	50%	100%
32	100%	67%	37	83%	83%	58	50%	20%
64	100%	60%	57	83%	83%	60	40%	100%
66	100%	60%	71	83%	75%	77	40%	60%
33	100%	60%	81	83%	75%			

Table 11: Users judging the query effort

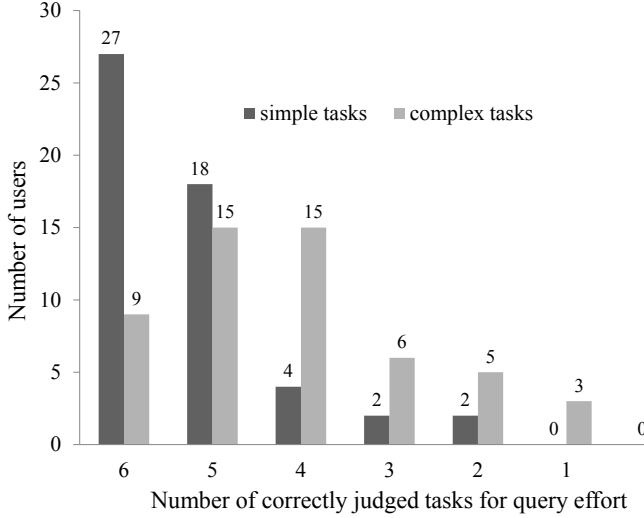


Figure 3: Users judging the query effort

judged query effort. From this figure it is also evident that users are far better at judging the query effort for simple tasks. 27 users of 53 (51%) have managed to correctly judge the query effort of all simple tasks, but only 9 out 53 (17%) have managed to correctly judge the difficulty of all complex tasks.

Table 12 shows the results for users judging the task outcome for simple and complex tasks. 14 out of 53 users (26%) were able to judge the time effort right for simple and complex tasks at the same time. 27 out of 53 users (51%) managed to judge the task outcome for simple tasks right and at the same time were not totally correct for the complex tasks. One user was able to correctly judge all complex tasks while wrongly judging a number of simple tasks.

Figure 4 illustrates a histogram of users versus correctly judged task outcome. From this figure it is also evident that users are far better at judging the query effort for simple

user	simple	complex	user	simple	complex	user	simple	complex
24	100%	100%	73	100%	83%	82	100%	50%
32	100%	100%	90	100%	83%	87	100%	50%
33	100%	100%	48	100%	80%	93	100%	40%
34	100%	100%	54	100%	80%	46	100%	33%
39	100%	100%	74	100%	80%	63	100%	33%
43	100%	100%	76	100%	78%	60	83%	83%
44	100%	100%	55	100%	75%	89	83%	83%
52	100%	100%	58	100%	75%	64	83%	67%
57	100%	100%	28	100%	67%	70	83%	67%
71	100%	100%	37	100%	67%	75	83%	63%
72	100%	100%	38	100%	67%	41	83%	50%
77	100%	100%	45	100%	67%	47	83%	33%
84	100%	100%	51	100%	67%	83	83%	33%
92	100%	100%	88	100%	67%	81	80%	67%
42	100%	83%	61	100%	50%	59	80%	20%
62	100%	83%	66	100%	50%	79	50%	67%
65	100%	83%	68	100%	50%	69	40%	100%
67	100%	83%	78	100%	50%			

Table 12: Users judging the task outcome

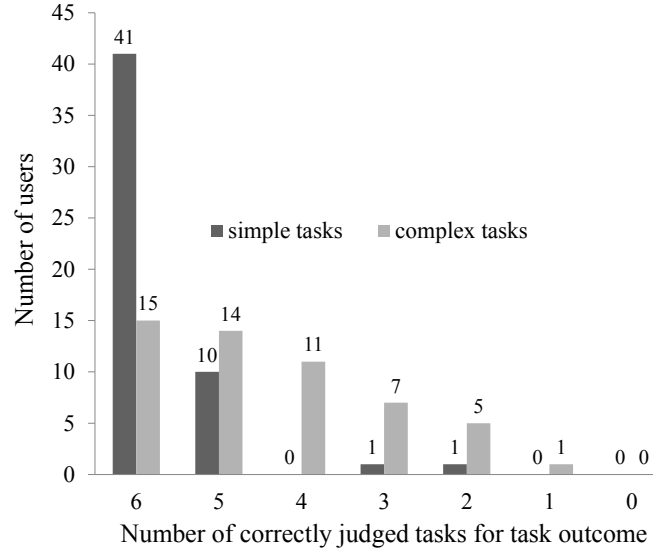


Figure 4: Users judging the task outcome

tasks. 41 users of 53 (77%) have managed to correctly judge the task outcome of all simple tasks, but only 15 out 53 (28%) have managed to correctly judge the difficulty of all complex tasks.

7. DISCUSSION

In this section, we first discuss our findings in the context of our six research questions. Then, we discuss the performance of the users for selected individual search tasks that brought some interesting results in terms of judgment capabilities.

It is obvious that in the case of simple tasks people are very well capable of assessing, how difficult a simple search task would be. For 90% of the study participants the estimated and experienced difficulties were in line.

This might be due to the fact, that most people, even or-

dinary Web users, have sufficient experience with carrying out simple search tasks on the Internet. Therefore they know what to expect. This could also be interpreted in a way that users also know for which tasks search engines could be helpful to them. When it comes to judging the search outcome and whether users would be able to find the correct results, 95% of the study participants correctly assess their ability to find the correct result. Ability here needs to be understood as comprising: as well understanding the problem as carrying out the task with a search tool.

When examining users' ability to judge the aforementioned parameters for complex search tasks, their ability to judge ability goes down in comparison to simple tasks. However, the fact that two thirds of the users are still able to sufficiently judge the subjective difficulty was a bit surprising to observe. It needs to be kept in mind that the experiment was carried out in a laboratory environment and probably the participants would judge differently in a real life scenario, where they were more emotionally involved in the study. In addition, especially the high factor of 73% claiming to have found the correct results is not in line with our manual evaluation of their results. Only 47% (158 out of 336 carried out tasks) of the results that were submitted for complex search tasks were correct. This may indicate that the problem with complex web searching might not be users finding no results, but the results found only seemingly being correct. This may explain why users are generally satisfied with their web search outcomes.

As expected we have observed significant differences between users judging simple tasks and users judging complex tasks. Users are significantly better at judging simple tasks than at judging complex tasks.

When examining users' ability to judge whether they had carried out a complex search task correctly or not, as in the previous section, the difference is significant. It is interesting to observe that in case of simple tasks, the users' judgment ability regarding the correctness of the task outcome is over 3 times better than in case of complex tasks (10% error rate versus 33% error rate). This could either be due to the fact that the complex tasks were quite open (not specific enough) as described by White and Iivonen [22] and therefore the users did not really have a sense for correct and incorrect. The other explanation would be that users got less support from the search engine side than expected.

When it comes to search capabilities, one would expect that better searchers would also be better at judging difficulty, effort and task outcome for complex search tasks. As the results shows, only for the task outcome, users who perform better in the whole experiment are also significantly better at judging the outcome of the task. For difficulty and effort, the differences are insignificant.

Regarding the question whether the judging performance is independent of the task type (simple/complex) but depends on the user the answer is as follows: There are some users who are able to correctly judge the task parameters like task outcome (26% of all users) both for simple and complex tasks. Yet the number of users who managed to correctly judge those parameters for simple tasks (and were wrong for all complex tasks) is much bigger (51% in case of task outcome) than the number of users who correctly judged the parameters for complex tasks and at the same time were wrong with their judgments for simple tasks (only one user in case of task outcome). Although numbers vary, this re-

lationship also holds true for task complexity, time effort, and query effort. Together with the results from research questions RQ1 to RQ5 it seems that task complexity indeed impacts the judging performance of users.

As far as preparing the work tasks and controlling the work task complexity is concerned, we expected Task 7 "Italy" to be simple when setting up the experiment. Due to the study participants clearly showing complex search behavior for this task, we reclassified it as a complex task for our analysis.

8. CONCLUSIONS AND FUTURE WORK

In this paper we have examined how ordinary Web search engine users manage to judge the three parameters task difficulty, task effort and task outcome for search tasks. We have compared according judgments for simple tasks and for complex tasks and also investigated, whether better searchers are also better judges. In addition we have investigated, whether the judging performance depends on task complexity or simply on the individual searcher.

Our results confirm that people are very well able to judge difficulty, effort and task outcome for simple tasks. They are significantly less good when they are asked to do the same for complex search tasks. Users tend to over estimate their own search capabilities in case of complex search tasks. We also observed, that better searchers are also better at judging whether they would be able to find the correct information than worse performing searchers. Regarding the hypothesis that the judging performance might depend on the users themselves (and not the task complexity), we can conclude from our results that task complexity is the main impacting factor for judging performance.

We also analyzed, how search engine operators could use the results of this paper to offer better support for search engine users to assess tasks. As expected in case of simple search tasks users are quite well able to judge task parameters like difficulty and effort. We have identified little need to offer better support for this kind of tasks. Yet when it comes to complex search tasks, we think it would help that search engines would at least build awareness about task complexity. Users need to know that tasks are different and that their expectations need to be in line with task complexity. If a task is complex, a user has to know that she needs to put in more cognitive effort as stated by Gwizdka [6]. We assume that more awareness would lower dissatisfaction. In addition it is also thinkable that search engine operators identify when people e.g. work on a task over a longer time. They could then assist those users by offering estimates for task effort and task time based on similar tasks carried out by other searchers.

As far as limitations of our study are concerned, we think that the sample used in our study was a bit wide. Users with very different backgrounds (from the house wife to the university student) participated. While this was of course intended to get realistic outcomes, it also resulted in sometimes high standard errors of mean for certain indicators. Some of the studies mentioned in the related work section (that only work with e.g. university students) might have brought clearer results. Yet a wide validity of results of those studies (that worked with less representative user samples) towards drawing conclusions for mainstream users remains questionable. Another limitation is that the study was carried out in laboratory environment and people were only

given a limited amount of time to carry out the tasks. We assume that taking away the time limitation would lead to slightly different results as published by Singer et al. [17].

In future work it would be interesting to not only analyze if study participants correctly or incorrectly judged tasks but also investigate to what extent the users tend to over- and underestimate the task parameters. Regarding sample size we are planning to run experiments with bigger sample sizes. This will enable us to get more correct statistics with more significant features. In addition we are planning to conduct studies with study participants from certain professional domains like teachers or blue collar workers only.

Acknowledgments

This paper was supported by the European Union Regional Development Fund through the Estonian Centre of Excellence in Computer Science and by the target funding scheme SF0180008s12. In addition the research was supported by the Estonian Information Technology Foundation (EITSA) and the Tiger University program, as well as by Archimedes Estonia.

9. REFERENCES

- [1] N. J. Belkin, R. Oddy, and H. Brooks. Ask for information retrieval: Part i. background and theory. *Journal of Documentation*, 38(2):61–71, Dec. 1982.
- [2] D. Bell and I. Ruthven. Searcher’s assessments of task complexity for web searching. In *Proceedings of the 26th European Conference on Information Retrieval (ECIR)*, pages 57–71. Springer Berlin / Heidelberg, 2004.
- [3] H. Bruce, R. Fidel, P. Yngversen, and P. Vakkari. *CoLIS 4: Proceedings of the Fourth International Conference on Conceptions of Library and Information Science*, Seattle, WA, USA, July 21-25, 2002. Libraries Unlimited, Greenwood Village, 2002.
- [4] K. Byström and P. Hansen. Conceptual framework for tasks in information studies. *Journal of the American Society for Information Science and Technology*, 56(10):1050–1061, 2005.
- [5] K. Byström and K. Järvelin. Task complexity affects information seeking and use. *Information Processing & Management*, 31(2):191–213, Apr. 1995.
- [6] J. Gwizdka. Assessing cognitive load on web search tasks. (arXiv:1001.1685), Jan. 2010. Comments: Published in the special issue Hot Topic: Cognition and the Web.
- [7] J. Gwizdka and I. Spence. What can searching behavior tell us about the difficulty of information tasks? a study of web navigation. *Proceedings of the American Society for Information Science and Technology*, 43(1):1–22, Jan. 2006.
- [8] P. Ingwersen. *Information retrieval interaction*. Taylor Graham Publishing, London, UK, UK, 1992.
- [9] P. Ingwersen and K. Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer Netherlands, 1 edition, Aug. 2005.
- [10] B. Kules and R. Capra. Designing exploratory search tasks for user studies of information seeking support systems. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pages 419–420, Austin, TX, USA, 2009. ACM.
- [11] Y. Li and N. J. Belkin. A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management*, 44(6):1822–1837, Nov. 2008.
- [12] Y. Li and N. J. Belkin. An exploration of the relationships between work task and interactive information search behavior. *Journal of the American Society for Information Science and Technology*, 61(9):1771–1789, 2010.
- [13] Y. Li, Y. Chen, J. Liu, Y. Cheng, X. Wang, P. Chen, and Q. Wang. Measuring task complexity in information search from user’s perspective. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–8, 2011.
- [14] G. Marchionini. Exploratory search: from finding to understanding. *Communications of the ACM*, Volume 49(4):41–46, 2006.
- [15] G. Singer, U. Norbistrath, and D. Danilov. Complex search: Aggregation, discovery, and synthesis. *Proceedings of the Estonian Academy of Sciences*, 61(2):89–106, 2012.
- [16] G. Singer, U. Norbistrath, and D. Lewandowski. Ordinary search engine users carrying out complex search tasks. <http://arxiv.org/abs/1206.1492>, 2012. Manuscript submitted for publication.
- [17] G. Singer, U. Norbistrath, E. Vainikko, H. Kikkas, and D. Lewandowski. Search-Logger analyzing exploratory search tasks. In *Proceedings of the 2011 ACM Symposium on Applied Computing, SAC ’11*, pages 751–756, New York, NY, USA, 2011. ACM.
- [18] G. Singer, P. Pruulmann-Vengerfeldt, U. Norbistrath, and D. Lewandowski. The relationship between internet user type and user performance when carrying out simple vs. complex search tasks. *First Monday*, 17(6), June 2012.
- [19] P. Vakkari. Task based information searching. *Annual Review of Information Science and Technology*, 37(1):413–464, Jan. 2003.
- [20] P. Vakkari and S. Huuskonen. Search effort degrades search output but improves task outcome. *Journal of the American Society for Information Science and Technology*, 63(4):657–670, Apr. 2012.
- [21] S. Weitz. Search isn’t search. Microsoft company report, Search Marketing Expo SMX, Munich, 2009.
- [22] M. D. White and M. Iivonen. Assessing level of difficulty in web search questions. *The Library Quarterly*, 72(2):205–233, Apr. 2002.
- [23] R. W. White, G. Marchionini, and G. Muresan. Evaluating exploratory search systems: Introduction to special topic issue of information processing and management. *Information Processing & Management*, 44(2):433–436, Mar. 2008.
- [24] R. W. White and R. A. Roth. Exploratory search: Beyond the Query-Response paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1):1–98, 2009.