User Evaluation of Query Quality

Wan-Ching Wu
SILS
University of North Carolina
Chapel Hill, NC, 27599 USA
wanchinw@email.unc.edu

Diane Kelly
SILS
University of North Carolina
Chapel Hill, NC, 27599 USA
dianek@email.unc.edu

Kun Huang
School of Management
Beijing Normal University
Beijing, China, 100875
huangkun@bnu.edu.cn

ABSTRACT
Although a great deal of research has been conducted about automatic techniques for determining query quality, there have been relatively few studies about how people judge query quality. This study investigated this topic through a laboratory experiment with 40 subjects. Subjects were shown eight information problems (five fact-finding and three exploratory) and asked to evaluate queries for these problems according to several quality attributes. Subjects then evaluated search engine results pages (SERPs) for each query, which were manipulated to exhibit different levels of performance. Following this, subjects reevaluated the queries, were interviewed about their evaluation approaches and repeated the rating procedure for two information problems. Results showed that for fact-finding information problems, longer queries received higher ratings (both initial and post-SERP), and that post-SERP query ratings were more affected by the proportion of relevant documents viewed to all documents viewed rather than the ranks of the relevant documents. For exploratory information problems, subjects' ratings were highly correlated with the number of relevant documents in the SERP as well as the proportion of relevant documents viewed. Subjects adopted several approaches when evaluating query quality, which led to different quality ratings. Finally, during the reliability check subjects' initial evaluations were fairly stable, but their post-SERP evaluations significantly increased.

Categories and Subject Descriptors
H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval - query formulation, search process.

General Terms
Experimentation, Human Factors

Keywords
Query quality, query recommendation, query evaluation

1. INTRODUCTION
Query performance prediction (QPP) is the task of estimating the expected quality of search results for a query in the absence of relevance feedback [4, 8]. The basic goal is to predict when a query will perform poorly so that some intervention can occur before results are returned. For example, additional information might be elicited from the user or term expansion might be used to enhance the query. QPP approaches are classified into two types: pre-retrieval and post-retrieval [4, 8]. Pre-retrieval approaches estimate query performance based on features of the query while post-retrieval approaches consider the results retrieved by the query. Pre-retrieval approaches are further subdivided into those that exploit the linguistic structure of the query, including the morphological, syntactical and semantic properties of the query, and those that use term statistics, including specificity, similarity, coherency and relatedness. Post-retrieval approaches include measures such as clarity and robustness, and score analysis.

Although a great deal of research has been conducted about QPP, there have been relatively few studies about the relationship among QPPs and users' evaluations of query difficulty. Hauff et al. [10] note "while most QPP methods have been motivated and developed based on how a user might rate a query, these intuitions have never been empirically validated" (pg. 980). To address this limitation, Hauff et al. [9, 10] compared the query performance ratings made by humans with performance scores estimated by a suite of QPP methods. Results showed that user ratings and QPPs were mostly uncorrelated, suggesting that QPP methods are not representative of how users evaluate query quality. Lioma et al. [12] found that users could not reliably identify pre-determined query difficulty ratings associated with a set of 420 queries, but were able to identify some features that would make a query difficult for a search system.

While these previous studies provide some insight about the relationship among QPPs and users' evaluations of query difficulty, they do not reveal insight about how people actually judge query quality. In one of the studies reported by Hauff et al. [10], assessors were provided with queries and information need descriptions and asked to judge the queries based on what they expected the results to be if they submitted the queries to a Web search engine. Assessors made their judgments using a 5-point scale, where 1=poor quality query and 5=high quality query. The researchers did not report assessors' experiences using this scale to evaluate query quality, although it was noted that their ratings varied considerably. Lioma et al. [12] asked assessors to rate queries using three categories (easy, medium, hard). In both studies, assessors evaluated queries without inspecting results. Neither study probed people about how they judged query quality.

People rate a variety of objects in daily life (e.g., movies, restaurants, books), but it is unlikely that many people have rated queries. How would people approach this task? What factors would they consider when evaluating query quality? How would they make decisions about which numeric ratings to assign to which queries? In this paper we explore these questions. We are not concerned with the relationship between QPPs and people's evaluations of query quality, but instead seek to address more fundamental questions about how people make evaluations of query quality. Specifically, our research questions are (RQ1)

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.

How do people make judgments about query quality? (RQ2) How are people's judgments related to features of the query, information problem and search results? (RQ3) How reliable are people's judgments?

Understanding how people evaluate query quality is important for several reasons. A better understanding of how people evaluate query quality might provide data on which to model future QPPs, or it might help researchers better understand the differences between automatic QPPs and human QPP. Although QPPs leverage collection-based statistics and should not necessarily be correlated with users' query evaluations, understanding how people view query quality might be helpful in modeling computer-human interaction regarding QPP. Understanding how people evaluate query quality is also important in the context of another popular IR technique: query suggestion. Traditionally, query suggestions are presented to users without any information about their potential goodness or quality. However, it might be useful to allow users to provide query *recommendations* to others, that is, query suggestions that have rating information associated with them in a manner similar to that provided by recommendation services such as Amazon and Netflix. It might also be useful for an information system to elicit query ratings from users so that it can make more personalized recommendations in the future. While *suggestion* and *recommendation* are often used as synonyms in the literature, in this study we distinguish between these two terms to demarcate a difference between the provision of unrated and rated suggestions, respectively. We pose one additional question to better understand what people think of the idea of query recommendation: (RQ4) What are people's perceptions of query recommendations? This is a natural question to ask in the context of an experiment that focuses on how people evaluate query quality, since this would presumably underlie query recommendation.

2. BACKGROUND

This study is related to three major areas of research: query performance predictors (QPPs), query suggestion, and online recommendations. Because this paper is not focused on automatic QPPs, this research is not reviewed (see [4] for an overview).

In a series of studies, Hauff and colleagues [9, 10] explored the relationship between QPPs and users' evaluations of query quality. This work was motivated by the observation that underlying most QPP methods are assumptions about how users evaluate query quality, despite a lack of research about how users make decisions about queries and suggestions. In a series of studies, Hauff and colleagues found few strong correlations between assessors' query quality evaluations and QPPs. The researchers first collected pre-retrieval ratings from 18 assessors for a set of 50 topics. Assessors were shown information need descriptions and queries and asked to indicate their expectations of the quality of the search results.

Hauff et al. [9, 10] found great variability is assessors' query ratings, and few direct correlations between these ratings and a set of QPP pre-retrieval measures. Hauff et al. also compared pre-retrieval and post-retrieval QPP measures with assessors' pre- and post-retrieval query quality ratings of a set of query suggestions. While assessors could distinguish between high and low quality suggestions, their ratings were uncorrelated with the automatic QPP measures. Across all experiments, the QPP measure that was most correlated with assessors' ratings was the pre-retrieval predictor *SumSCQ* which assigns higher quality scores to more specific queries. Typically queries that contain more terms are associated with a higher *SumSCQ*; this suggests that people's pre-retrieval query quality judgments might be influenced by query length. Similarly, Lioma et al [12] found that people were fairly good at identifying features that were correlated with query difficulty, including queries that were "too vague" or "too short," but were unable to reliably assess query difficulty using a rating scale of easy, medium and hard.

In addition to these studies, two studies have examined the relationship between QPPs and user performance [18, 19]. While these studies focused on different research questions than the ones on which we focus, their results showed that user performance was unrelated to several QPP measures. Turpin and Hersh [18] found that the clarity scores of queries created by their users were much lower on average than those reported in system-based QPP evaluations where queries were automatically created. Turpin and Hersh further found that there was no correlation between the *clarity* of a user's query (a post-retrieval QPP) and the user's actual performance on the search task. Zhao and Scholer [19] examined nine pre-retrieval QPPs and found that they were not useful predictors of when users experienced search difficulties.

A portion of Hauff et al.'s [10] work was a secondary analysis of data collected in a study investigating the extent to which users could be induced to take bad query suggestions because they believed many others had taken the suggestions [11]. Kelly et al. [11] did not find an effect for "query popularity," but did find that users selected significantly more high quality than low quality query suggestions. Although this research was not conducted in the context of QPP, it does provide some indirect evidence that people make predictive judgments about query quality when selecting among a set of suggestions and that these judgments are often accurate. Like QPP research, query suggestion research is an area that might benefit from an increased understanding of how people evaluate query quality.

Increasing amounts of research have been published about query suggestion within the last ten years [15] and query suggestion is now a common feature of many information search systems. When users issue queries to most major commercial search services, they receive search results and query suggestions. However, they do not receive information about how the query suggestions were selected. While these query suggestions are likely identified through a combination of techniques based on aggregated query log data, details about the usage and frequency of various queries are not displayed to users. In social and collaborative search systems, query suggestions come from fellow users who belong to a community or specialized group with a common set of information needs [1, 16, 17]. These suggestions are also free-standing in the sense that users only know that they have been suggested or used, but nothing about their potential quality or usefulness. Thus, it seems that an important opportunity might exist in allowing users to associate ratings with query suggestions. In recent work, Baraglia et al. [2] proposed the idea of machine-rated query suggestions where queries would be rated with a 10-point scale (10=positive, 5=neutral and 1=negative), but their research was focused on algorithms for selecting and rating suggestions, rather than user ratings or interpretations of ratings.

In this paper we distinguish between query suggestions and query recommendations. The key difference between the two is the presence (*recommendation*) or absence (*suggestion*) of rating information. While suggestions can be generated by both systems and humans, ratings are, in most cases, associated with humans. The basic act of associating ratings with items is one with which most people are familiar. However, rating behavior, in general, is

not a well-understood topic, even in the recommender system literature [7]. A recent call for a special issue of *ACM Transactions on Interactive Intelligent Systems (TiiS)* states that little research has focused on the decision-making processes of users. Instead, research has focused on algorithms for identifying recommendations and eliciting and modeling users' preferences[1].

Rating behavior has been studied more extensively in the consumer behavior literature. This research has primarily focused on consumer bias, cultural differences, brand loyalty and receiver experience and expertise in relation to commercial products [c.f., 6, 14]. A typical way these studies conceptualize objects that are rated, and people's subsequent use of ratings for making decisions about these objects, is by classifying them into *search* and *experience* goods [13]. *Search goods* are goods that are characterized by product attributes for which full information can be acquired before purchase. Search goods (e.g., dishwasher) are typically evaluated using objective criteria (e.g., capacity, noise level). Institutional-based ratings (e.g., *Consumer Reports*) often guide purchasing behavior. E*xperience goods* (e.g., wine, movies, and books) are goods whose ratings are dominated by subjective attributes. For these types of goods, user ratings (instead of institutional ratings) often guide purchasing behavior.

We submit that query recommendations have more in common with experience goods. As with most experience goods, users need to *experience* the item before rating it; we are not suggesting that people blindly rate other people's queries, but rather rate their own queries after they have finished searching. It is an open question as to how people would approach this task and if they would even find recommendations useful. Furthermore, previous research has found the stability of people's ratings of experience goods somewhat brittle and influenced by the ratings of others [20]; thus, another open question concerns the stability of people's ratings.

3. METHOD

A laboratory experiment was conducted using a classic pre-test/post-test design. The pre-test allowed us to observe how subjects would rate query quality before viewing search results (pre-retrieval), while the post-test allowed us to observe if and how these evaluations changed as a result of viewing search results (post-retrieval). We manipulated search result quality to determine if specific performance levels could be mapped to specific query quality scores.

Subjects were instructed that they would be shown other people's information problems and queries and then provide evaluations of query quality. Subjects completed three major steps: (1) *initial query evaluation* where they were shown an information problem description and query and asked to evaluate the query according to several attributes; (2) *search result evaluation* where they were shown a list of 10 search results and asked to evaluate these results; and (3) *post-SERP query evaluation* where they were once again presented with the information problem description and query and asked to re-evaluate the query. These steps were completed for eight information problems. After the evaluation, subjects were interviewed about the strategies they used to evaluate queries and their opinions about the usefulness of query recommendations. Following this, subjects were asked to repeat the rating procedure for two information problems/queries they had previously seen. This functioned as a reliability check to test the stability of subjects' query quality evaluations.

3.1 Information Problems and Queries

Eight information problems were used in this study. Five of the information problems were fact-finding information problems (FF) and were based on the navigational tasks used in [3]. Three were exploratory information problems (EX) and were created based on topics used in the TREC Aquaint collection (although results were from the Web, not the Aquaint corpus). Each information problem was associated with one query. Query length for FF information problems ranged from one to five words, while for EX information problems, it remained fixed at three words.

Subjects completed the FF and EX problems in blocks (all the FF problems were in one block and all the EX problems were in another). Subjects were randomly assigned to receive either the FF block or EX block first. Within each block, information problems were rotated using a Latin-square. Examples of information problems and queries are shown in Table 1. The full set of queries is shown in Table 2. The full set of information problems and queries can be viewed online at (http://ils.unc.edu/sigir2012queryrating/).

Table 1. Examples of information problems and queries.

Information Problem Description	Query	Task Type
Bob is interested in researching the history of motorcycles. To start his research, he decides to find out when each major motorcycle company was founded. Specifically, he wants to determine the year in which Harley Davidson motorcycles was founded.	harley davidson	FF
Janet is planning a boat convention for her company which will be held next year in Las Vegas and needs to select a hotel. She has heard positive things about the Bellagio hotel, but first wants to find out how many guest rooms are available.	bellagio hotel las vegas rooms	FF
Carol is planning to fly to Amsterdam next month and would like to learn more about the body scanners that are being used in many airports as part of routine security procedures. Specifically, she is interested in gathering a range of information and opinions about these scanners and any privacy or health issues related to their use.	airport body scanners	EX

Table 2. Information problem IDs, queries and length.

Information Problem IDs	Query	Length
FF1	billiards	1
FF2	harley davidson	2
FF3	bank savings rates	3
FF4	statue of liberty spikes	4
FF5	bellagio hotel las vegas rooms	5
EX1	black bear attacks	3
EX2	modern day piracy	3
EX3	airport body scanners	3

[1] http://tiis.acm.org/pdf/human-decision-making-and-recommder-systems.pdf

3.2 Initial Query Evaluation

During the initial query evaluation, subjects were presented with an information problem and query along with four questions about query quality (Table 3). The query quality questions assessed the subject's beliefs about the representativeness of the query, the likelihood it would retrieve useful results, the extent to which the subject would recommend the query to others, and the subject's star-rating of the query. The first three items were measured using a 5–point scale and the star-rating scale ranged from 1 star to 5 stars. This latter type of rating was selected because of its ubiquity in online computing environments.

Table 3. Initial query evaluation items.

Item	Scale
How well do you think the query represents the information problem?	1=not at all; 3=somewhat; 5=very well
How useful do you expect the search to be?	1=not at all; 3=somewhat; 5=very useful
How likely would you be to recommend this query to someone else searching for the same information problem?	1=not at all; 3=somewhat; 5=very likely
Please select the star rating that best reflects your opinion of the potential quality of the query.	★ ★ ★ ★ ★

3.2 Search Result Evaluation

After subjects completed the initial query evaluation, they were directed to a search engine results page (SERP) containing a list of ten search results. We collected 10 search results for each information problem using the Google search engine. Search results were carefully screened until the relevance of each search result was agreed upon by two authors of the study. On the SERP, each result was represented by a title, snippet, and URL (Figure 1). To enhance the realism of the search result environment, the SERP was deliberately designed to mimic Google's SERP. Subjects were debriefed at the end of the experiment about this manipulation. We modified snippets so that they did not contain answers to the information problems. The information problem was visible at the bottom of the page (omitted from screen shot to conserve space).

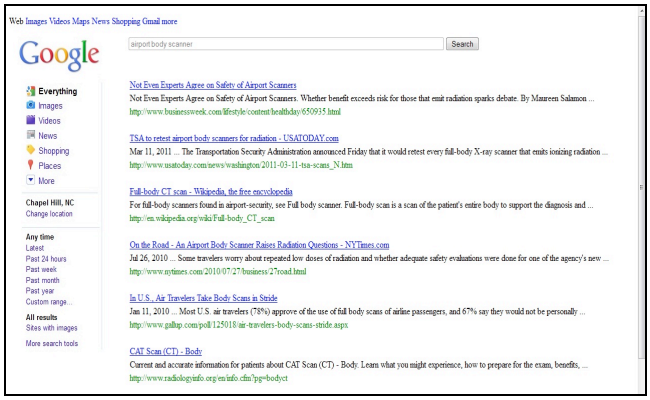

Figure 1. Artificial SERP constructed for study.

When presenting search results of FF information problems, subjects were randomly assigned to a high or a low condition, measured by Reciprocal Rank (RR) – the inverse of the rank of the first relevant search result in a list. For each FF information problem, only one relevant search result appeared in the search results list. For subjects assigned to the high condition, the relevant search result appeared at positions 1 through 5 for FF1-FF5 and for subjects in the low condition the relevant search result appeared at positions 6 through 10. The performance conditions can be viewed in Table 4. Each FF information problem was associated with two performance conditions.

Table 4. Performance conditions for FF problems.

Information Problem	Reciprocal Rank (High Condition)	Reciprocal Rank (Low Condition)
FF1	1.000	0.167
FF2	0.500	0.143
FF3	0.330	0.125
FF4	0.250	0.111
FF5	0.200	0.100

For each EX information problem, subjects were randomly assigned to one of three performance conditions: low, medium, or high. The conditions varied in terms of both number of relevant results presented and the positions of the relevant results, resulting in variations in Normalized Discounted Cumulated Gain (nDCG) from 0.31 to 1. The three conditions for any single EX information problem only differed in terms of where relevant search results were ranked, but different EX information problems differed in the number of relevant results shown on a result list. The performance conditions are shown in Table 5.

Table 5. Performance conditions for EX problems.

Information Problem	Number Relevant	Performance Condition	nDCG
EX1	2	Low	0.31
		Medium	0.51
		High	1.00
EX2	4	Low	0.42
		Medium	0.59
		High	0.80
EX3	6	Low	0.65
		Medium	0.73
		High	0.91

For each search result viewed, subjects were asked two questions: (1) Does the webpage contain the exact information needed? (Yes, No) and (2) Do you think the webpage was useful to the person who typed the query? (1=not at all; 3=somewhat; 5=very useful). Only when both questions were answered could they return to the SERP. Subjects were not required to examine all 10 results and could examine the results in any order. Once they felt they had gathered enough information from the result list to evaluate the query, they proceeded to the post-SERP query evaluation.

3.3 Post-SERP Query Evaluation

During the post-SERP query evaluation, subjects were shown the information problem and query again, and asked the following: (1) How likely would you be to recommend this query to someone else searching for the same information problem? (1=not at all; 3=somewhat; 5=very likely) and (2) Based on the documents you've examined on the search result list, please select the star rating that best reflects your opinion of the actual quality of the query (subjects were presented with the 5-star rating widget). Both of these questions were from the initial query evaluation and functioned as post-test questions. While the first question was identical to one of the initial query evaluation questions, the second contained slight word changes to indicate that subjects should consider their experiences evaluating search results.

3.4 Exit Interview

After subjects completed the evaluations, semi-structured interviews were conducted to obtain qualitative data about the strategies subjects took to evaluate queries and their perceptions of the usefulness of query recommendations. Interviews were tape-recorded and the interview scheme was composed of four parts: questions regarding how subjects judged the quality of queries on their own, questions regarding how search results affected subjects' judgments of query quality, questions regarding evaluation strategies used for different information problem types and questions regarding subjects' opinions of query recommendations in the context of online searching.

3.5 Reliability Check

After the interviews, subjects were asked to re-evaluate the last information problem they were given from the first task block and the first information problem from the second task block. For each problem, subjects repeated the initial query evaluation, search results evaluation and post-SERP query evaluation. Subjects did not have access to their previous ratings.

3.6 Subjects

Subjects were recruited through the staff mailing list at our university. Forty-one subjects participated. From the interview session it was found that one subject misunderstood the study instructions and was therefore excluded from analysis.

Before starting the query evaluations, subjects completed a Demographic Questionnaire. The vast majority of respondents were females (70.73%), 26.83% were males, and 2.44% did not answer this question. Their ages ranged from 18-66 years old (M=39.6, SD=14.0). With respect to the status of participants, 25 (62.5%) were university staff; nine (22.5%) had both student and staff status; two (5%) were students and three (7.5%) were neither university staff nor student, and the status of one subject was unknown. For subjects who were university staff or full-time professionals (37, 92.5%), their occupations ranged from research assistant, project manager/director, programmers/analyst, administrative assistant or manager, financial counselor, business manager, lecturer, editor, teacher, to masseuse.

Subjects' search experience was measured with the Search Self-Efficacy Scale [5]. The Search Self-Efficacy Scale is a 14-item scale used to characterize search expertise. Subjects indicate their confidence in completing a series of activities using a 10 point scale where 1=totally unconfident and 10=totally confident. Subjects scored an average of 7.38 (SD=1.42) on the Search Self-Efficacy Scale, indicating moderate to high search experience. Because we slightly modified the wording of some items from the original scale to make them more contemporary, Cronbach's alpha was computed using subjects' responses to these items. This was found to be 0.947, demonstrating high reliability.

4. RESULTS

The research questions we addressed were (RQ1) How do people make judgments about query quality? (RQ2) How are people's judgments related to features of the query, information problem and search results? (RQ3) How reliable are people's judgments? And (RQ4) What are people's perceptions of query recommendations?

4.1 Initial Query Evaluation

In this section, we investigate how subjects' initial evaluations varied according to information problem type and query length (RQ2). Pearson product-moment correlation coefficients were first computed to assess the relationships among the four initial query evaluation items. Results showed that there was a high correlation among subjects' responses to the items (Table 6).

Table 6. Correlations among initial query evaluation items.

	Rep	Useful	Rec	Rating
Representativeness	-	.869**	.865**	.843**
Useful	.869**	-	.866**	.859**
Recommendation	.865**	.866**	-	.885**
Rating	.843**	.859**	.885**	-

**Correlation is significant at the 0.01 level (2-tailed).

Figures 2 and 3 show subjects' initial query quality ratings for the fact-finding (FF) and exploratory (EX) information problems, respectively. The two figures show that while the initial query evaluations across the four items were virtually identical for different EX information problems, these values increased for FF information problems as query length increased from one word to four words. The figures also show that subjects were generally more conservative with their *recommendation* and *star-ratings* judgments. It was found that for FF information problems there was a significant difference in subjects' judgments of representativeness ($F(4,195)=23.21$, $p<.001$), usefulness ($F(4,195)=22.34$, $p<.001$), recommendation ($F(4,195)=30.11$, $p<.001$) and star-rating ($F(4,195)=22.41$, $p<.001$). Post-hoc tests indicate that the ratings given to FF1 were significantly lower than those given to FF2, FF3, FF4 and FF5; that the ratings given to FF2 were significantly lower than those given to FF4 and FF5; and the ratings given to FF3 were significantly lower than those given to FF4 and FF5. There were no significant differences in subjects' ratings of FF4 and FF5. For EX information problems, there were no significant differences for any of the initial query ratings (representativeness: $F(2, 78)=.27$, $p=.762$; usefulness: $F(2, 78)=.10$, $p=.907$; recommendation: $F(2,78)=.16$, $p=.852$; star-rating: $F(2, 78)=.04$, $p=.959$)

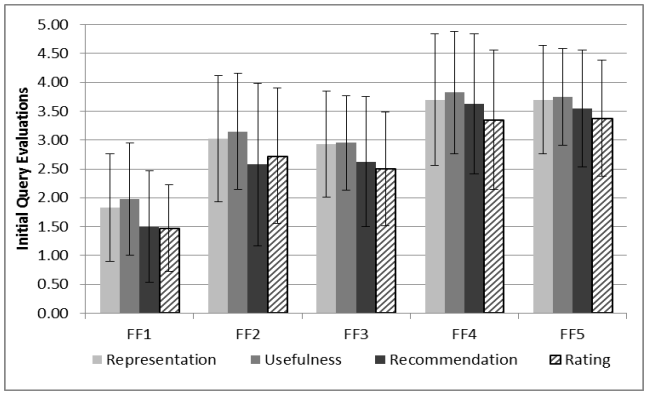

Figure 2. Initial evaluations of FF information problem queries (error bars represent +/-1 standard deviation).

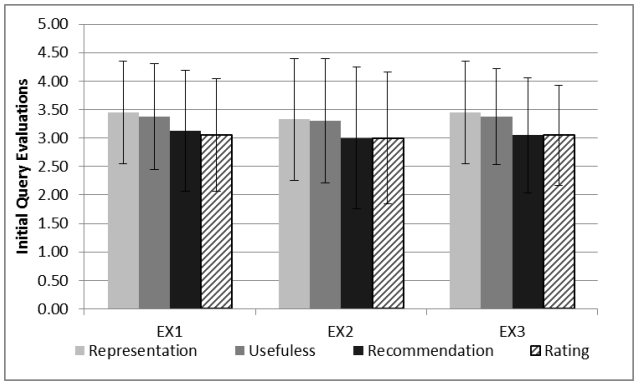

Figure 3. Initial evaluations of EX information problem queries (error bars represent +/-1 standard deviation).

Since all EX information problems were represented by three-word queries, EX information problems were compared to FF3 to examine whether there was a difference in how people rated queries according to information problem type. It was found that overall subjects rated the three-term EX queries higher on all four initial query evaluation items than the three-term FF query. Subjects found EX three-term queries to be more representative (EX: M=3.41, SD=.65; FF3: M=2.93, SD=.92) and the searches to be more useful (EX: M=3.35, SD=.64; FF3: M=2.95, SD=.81); subjects were more likely to recommend three-term EX queries to other people (EX: M=3.06, SD=.76; FF3: M=2.63, SD=1.13) and also assigned more stars to these queries (EX: M=3.03, SD=.70; FF3: M=2.50, SD=.99). Paired-samples t-tests were conducted to compare the means between EX and FF3 on all four items and the results show that three-term queries for EX information problems were rated as significantly more representative ($t(39)=-2.65$, $p=.012$), more useful ($t(39)=-2.53$, $p=.015$) and received significantly more stars than FF3 ($t(39)=-2.85$, $p=.007$).

4.2 Search Result Evaluation

RQ2 also asked about how people's judgments are affected by the quality of the search results. Before we examine this, we first examine the extent to which our performance manipulations worked. During the search result evaluation stage a total of 1479 result clicks were made, which represented 46.22% of the total search results shown to subjects. Figure 4 shows the percentage of times subjects' explicit relevance judgments corresponded with the judgments made by the researchers and the *click rate* for each information problem. *Click rate* is defined as the percentage of web results clicked from a ten-result SERP. If a web result prejudged by the researchers as relevant was clicked and also judged by a subject as relevant, it was counted as an instance of *correspondence*. Note that we are not measuring inter-rater reliability (as is common in the IR literature), but rather whether our experimental manipulations worked. This is why we use the term *correspondence* and report percent agreement. The figure shows that all of the information problems had an 80% or above correspondence rate except for FF3, which had only a 65% correspondence rate. FF3 also received the highest click rate among a range of click rates from 0.40 to 0.57. Even though the information problem had a well-defined goal (locate a bank with a particular savings rate), subjects might have had a difficult time processing the results and understanding if results were relevant.

Figures 5 and 6 compare subjects' *experienced performance* (EP) (according to their clicks and explicit relevance judgments) with the manipulated performance, or the *performance intended* by the researchers (IP). EP for FF information problems, measured by RR, was computed by taking the inverse of the rank of the first clicked result. The greatest disagreement occurred for FF3, which was expected given the results in Figure 4. EP for EX information problems was measured by nDCG of the ten returned results; unclicked results were scored as irrelevant. Despite the disagreement between EP and IP for EX information problems, subjects experienced the intended relative performance conditions except for EX3. For this topic, there were 6 relevant documents in the list and for the high condition this meant that most of these were concentrated near the top of the list. It is likely that subjects stopped reviewing the list once they found a few relevant documents, which would lead to lower EP nDCG scores for high performance conditions. Overall, the mismatch between IP and EP for EX information problems is likely related to the varying number of documents opened by subjects and a discrepancy between their judgments and the researchers' judgments.

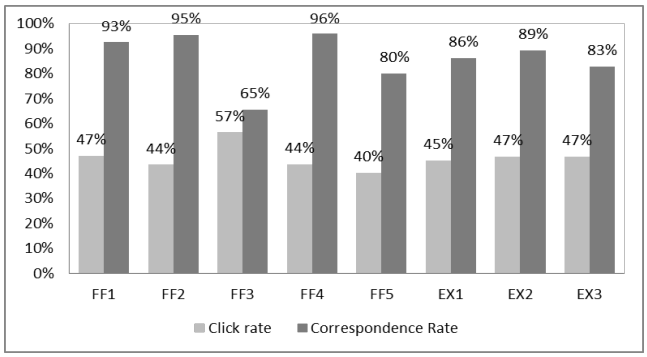

Figure 4. Search results click rate and correspondence rate.

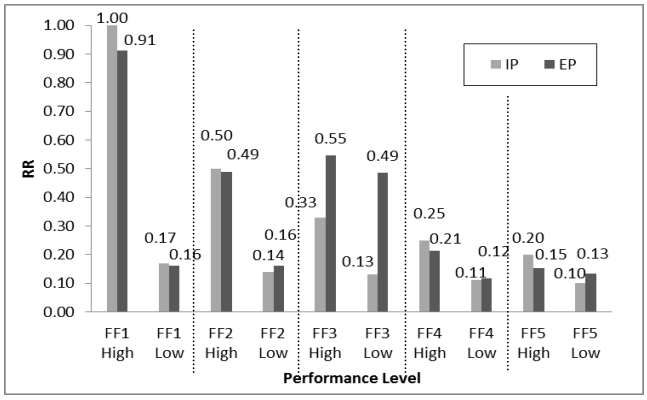

Figure 5. Comparison of intended performance (IP) and experienced performance (EP) for FF information problems.

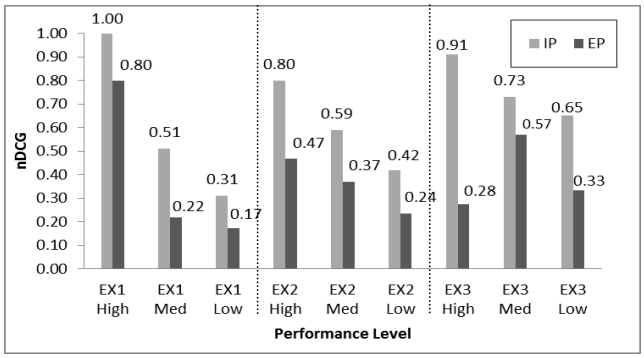

Figure 6. Comparison of intended performance (IP) and experienced performance (EP) for EX information problems.

Results also provide some initial insight into RQ1: How do people make judgments about query quality? In considering subjects' evaluation approaches for each information problem type (Figure 7), we see that for FF information problems most subjects either took an *exhaustive approach*, evaluating all 10 (22%) search results or a *selective approach* only evaluating 1-3 (50%) results. For EX information problems, we see greater diversity in the number of results evaluated. Sixty-one percent of subjects evaluated 1-4 results, while about 18% evaluated all 10 results. In Section 4.4, we examine the relationship between evaluation approach (*exhaustive* vs. *selective*) and subjects' post-SERP query evaluations in more depth.

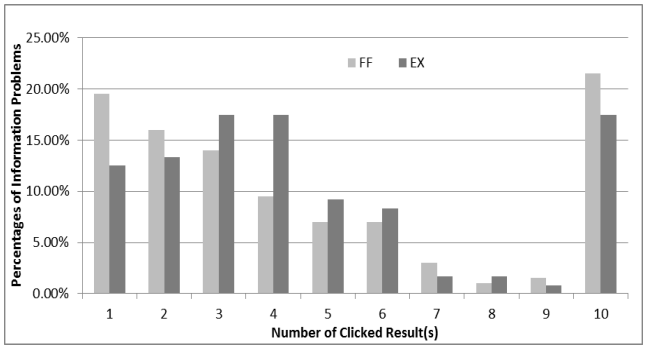

Figure 7. Distribution of number of clicked results according to information problem type.

The order in which search results were clicked was analyzed to investigate whether there was a relationship between click order, rank of result and subjects' relevance judgments. Click order was positively correlated with rank of search result (r=.56, p<.001), and subjects tended to click results ranked higher first. Click order was negatively correlated with perceived usefulness of a search result (r= -.46, p<.01); the search results subjects clicked earlier in time were evaluated as more useful than the ones clicked later.

4.3 Post-SERP Evaluation

After investigating subjects' interactions with the search results, we can now investigate how perceived quality of the search results impacted subjects' ratings (RQ2). During the post-SERP evaluation, subjects were asked again how likely they would be to recommend a query and how many stars they would assign it. The star-ratings assigned and likelihood to recommend the query were positively correlated (r=.88, p<.001). Thus, in subsequent analysis, we only include results related to one of these measures.

4.3.1 FF Information Problem Post-SERP Ratings

Figure 8 shows the relationship between subjects' initial query quality ratings and their post-SERP ratings for the FF information problems. The post-SERP ratings are further divided according to intended performance level (recall that FF information problems were grouped into high and low performance sets). Queries associated with higher performance levels consistently received higher post-SERP ratings than those associated with lower levels. Subjects in the low performance group evaluated an average of 4 documents, while those in the high performance group evaluated an average of 7 documents. Many subjects in the low performance group likely gave-up without finding a relevant document; this might explain the lower post-SERP ratings of this group. Independent-samples t-tests were applied to compare the means in post-SERP ratings between the high (n=21) and low performance groups (n=19) for each FF information problem. None of pairs was significant (FF1: t(38)=-1.96, p=0.056; FF2: t(38)=-0.76, p=0.46; FF3: t(38)=-0.67, p=0.51; FF4: t(38)=-1.80, p=.08; t(38)=-0.32, p=.75).

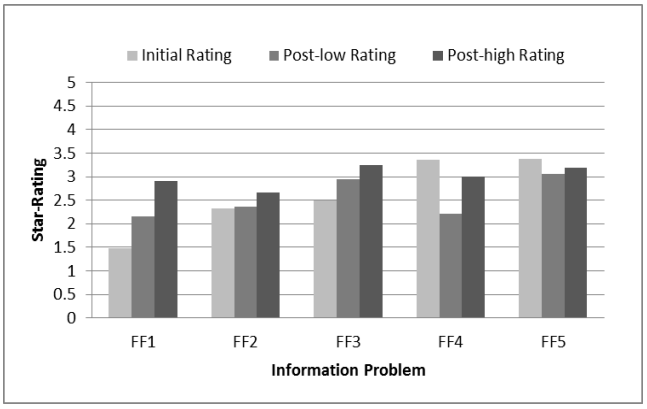

Figure 8. Comparisons of initial and post-SERP ratings for FF information problems according to performance.

The post-SERP ratings were higher than the initial ratings for all information problems except FF4 and FF5. Paired-samples t-tests were conducted to compare initial ratings to post-SERP ratings. Results showed that the only times that initial and post-SERP ratings were significantly different from each other were when FF1 and FF3 information problems were coupled with the high performance levels: FF1-H (t(20)=-5.45, p<.001); FF3-H (t(20)=-2.72, p=.013) and when FF4 was coupled with the low performance level (t(18)=2.79, p=.012).

During the search result evaluations, subjects were allowed to view any number of search results and in any order. To better understand the relationship between subjects' evaluation behaviors and their query quality ratings, we examine *interactive precision,* or the ratio of number of documents marked relevant to number of documents evaluated (Figure 9). In general, post-SERP ratings increased as *interactive precision* went up; the two variables were significantly correlated (r(198)=.57, p<.001).

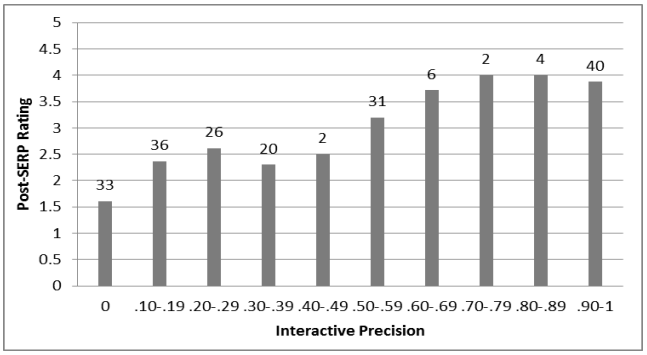

Figure 9. Relationship between interactive precision and post-SERP ratings for FF information problems (data labels represent number of information problems).

Multiple linear regression analysis was used to model the post-SERP ratings given to FF information problems. Since in most cases, the initial and post-SERP ratings did not differ significantly from each other, query length was considered to be a predictor along with number of relevant search results experienced, and *interactive precision*. A stepwise method was used and two models were derived. The first model showed that *interactive precision* was the single best predictor in post-SERP ratings, β = .57, t(198) = 9.72, p < .001. *Interactive precision* also explained a significant proportion of variance in post-SERP rating (adjusted R^2=.32, F (1, 197) = 94.49, p<.001). In the second model both

interactive precision and length significantly predicted post-SERP ratings (interactive precision: $\beta = .58$, $t(198) = 10.00$, $p < .001$; length: $\beta = .16$, $t(197) = 2.75$, $p = .001$); they explained 34.1% of variance in post-SERP rating ($F(2, 197) = 52.58$, $p<.001$).

4.3.2 EX Information Problems Post-SERP Ratings

While in FF information problems the descriptive statistics show that higher post-SERP ratings followed high performance levels, the trend was not observed for EX information problems. When post-ratings were grouped individually by *interactive precision*, post-SERP ratings significantly increased with *interactive precision* (Figure 10) ($r(118)=.61$, $p<.001$).

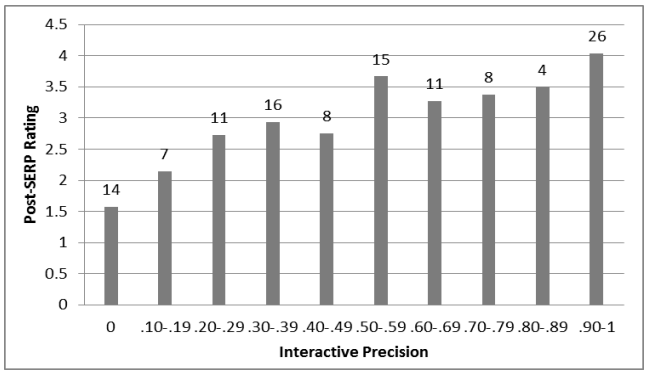

Figure 10. Relationship between interactive precision and post-SERP ratings for EX information problems (data labels represent number of information problems).

Initial ratings and post-SERP ratings of query quality for EX information problems are displayed in Figure 11. Paired-samples t-tests found that the initial and post-SERP ratings were only significantly different for the medium performance level in EX3 ($t(13)=-3.24$, $p=.006$). At first glance it might seem unusual that the difference between the initial and post-SERP ratings were not also significant for the low group for EX3. Paired-sample t-tests were conducted, so there was likely a difference between the initial ratings of those who received the low and medium performance levels for this problem.

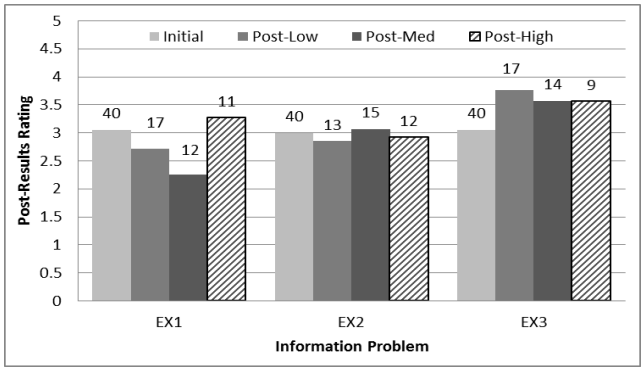

Figure 11. Comparisons of initial and post-SERP ratings for EX information problems by information problem and performance level (data labels represent number of subjects).

Figure 12 shows the differences in subjects' initial ratings and their post-SERP ratings as number of relevant search results experienced increased. Overall, subjects were more consistent in the direction of rating adjustment when they encountered few relevant search results. In cases where subjects identified no relevant documents at all, in most information problems the post-SERP ratings decreased (12 out of 14). In cases where subjects found one or two relevant results, the direction of the differences were not as predictable. When subjects identified more than three relevant search results, in most cases, subjects' post-SERP ratings were higher than their initial ratings.

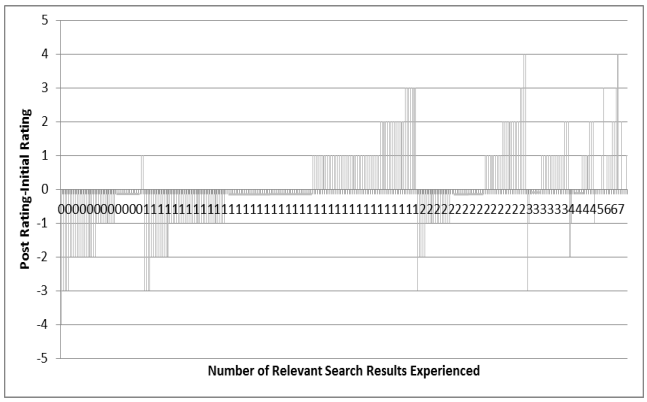

Figure 12. Relationship between number of relevant search results experienced and difference in initial and post-SERP ratings for EX information problems.

A regression analysis was also conducted for EX information problems to better understand which factors most influenced subjects' ratings. *Interactive precision*, number of relevant search results found, and experienced performance were entered into a stepwise regression. Results showed that *interactive precision* was the best predictor of post-SERP ratings, $\beta = .61$, $t(118) = 8.44$, $p<.001$; it also explained a significant proportion of variance in post-SERP ratings (adjusted $R^2=.371$, $F(1, 118)=71.25$, $p<.001$).

4.4 Query Quality Evaluation Approaches

RQ1 asked how people make judgments of query quality. In Section 4.2, it was shown that subjects seemed to take different strategies when evaluating search results. In order to examine whether different approaches had any influence on post-SERP ratings, we classified and compared subjects' evaluation approaches. For FF information problems, we classified subjects' evaluation approaches into two types: *Selective* and *Exhaustive*. Subjects using the *Selective* approach only examined one search result, while those using an *Exhaustive* approach examined all ten search results. For EX information problems subjects took more varied approaches resulting in three groups: *Selective*, *Persistent* and *Exhaustive*. The *Selective* approach represented cases where subjects examined 1-3 results, the *Persistent*, 4 to 7 results, and the *Exhaustive*, 8 to 10 results. These evaluation approaches were examined in the context of subjects' post-SERP ratings.

Results show that for FF information problems, those using the *Exhaustive* evaluation approach gave lower ratings to queries (M=2.14, SD=1.04, n=43) than those using the *Selective* approach (M=3.26, SD=1.27, n=39). An independent-samples t-test indicated this difference was significant ($t(1,80)=4.38$, $p<.001$). For EX information problems, those using the *Exhaustive* approach assigned the highest star-ratings (M=3.17, SD=0.96, n=24), followed by the *Persistent* (M=3.07, SD=1.04, n=44) and the *Selective* (M=3.12, SD=1.34, n=52), but the differences were not significant ($F(2, 117)=.06$, $p=0.95$).

In addition to examining subjects' evaluation behaviors, we asked subjects during the exit interview how they initially evaluated queries and then re-evaluated them after viewing search results (RQ1). With respect to the criteria subjects used to assign the initial query ratings, most stated that the specificity of the query was the most important factor. Some mentioned that they

compared queries to what they would put in a search box; if they could think of better expressions they would rate the queries worse, and vice versa. Subjects took more varied approaches to assigning post-SERP ratings. Some mentioned they based their ratings on the number of relevant results they had found, some relied on the positions of relevant search results, and others considered both factors. Yet rarely did subjects differentiate between approaches taken to evaluate queries for FF and EX information problems. Although some indicated they spent more time on EX information problems because they wished to gather more background information on the topic, many of them said that for FF problems they kept looking for other search results even when they had successfully identified the answer. The motivation for the latter came from the fear that they could miss something, the answer could be wrong, they had personal interests in some topics, or simply out of habit.

We also asked subjects how they felt about using star-ratings during real-world searching (RQ4). When asked whether they would consider using a system which provides rated query suggestions, many said yes because it would save time and benefit people who are not experienced with searching. Some said it depended on who rated the query suggestions and how well they could formulate queries by themselves. People who said they would not use such systems expressed doubt because systems or other people could not predict their information needs, or even if they could some subjects believed star-ratings were very subjective and they would rather have descriptions about why a query could be useful. Subjects constantly associated query star-ratings with hotel and restaurant stars and other product reviews and ratings such as those that appear in Netflix and Amazon. Their familiarity with these common usages resulted in a general agreement on the concept: the more stars there were, the higher the quality. The majority of the subjects believed 5-star query suggestions would lead to the most relevant results, some mentioned that they would expect these queries to also help them find information the quickest. Subjects said they would interpret 3-star queries as those that lead to information that was buried in the search results and would take some time to discover. Finally, subjects said they would interpret 1-star queries as useless and unhelpful; most said they would not use 1-star queries.

4.5 Reliability Check

To examine the consistency of subjects' evaluations (RQ3), subjects were asked to repeat the evaluation procedures for the last information problem they were given from the first task block and the first information problem from the second task block; thus, subjects reevaluated one FF and one EX information problem. Note that they did not have access to their previous ratings in this stage. Paired-samples t-tests were conducted to examine the differences in subjects' evaluations of FF and EX information problems (Table 7). Results demonstrated that subjects' initial ratings in the reliability check did not significantly differ from those in the main study session for both FF and EX information problems, but subjects' post-SERP ratings in the reliability check were significantly higher than those in the main study session for both information problem types. We note that subjects clicked on fewer search results during the reliability check (M_{FF}=3.05, M_{EX}=3.60) than the study session (M_{FF}=4.15, M_{EX}=4.68) and that the distributions of the clicked results were significantly different (FF:Z=-6.50, p<.001; EX:Z=-6.51, p<.001).

Table 7. Relationship between subjects' ratings during the main study and reliability check (*p<.05; **p<.01).

	FF: Initial Rating	FF: Post-SERP Rating	EX: Initial Rating	EX: Post-SERP Rating
Main Study	2.60 (1.32)	2.70 (1.49)	3.10 (1.13)	2.80 (1.29)
Reliability Check	2.80 (1.10)	3.10 (1.34)	3.30 (1.04)	3.23 (1.19)
t-statistic	-1.11	-2.58**	-1.31	-2.21*

5. DISCUSSION

Our study conceptualized query evaluation as a two-stage process: initial evaluation and post-SERP evaluation, a distinction drawn from pre-retrieval and post-retrieval QPP approaches. Queries were first assessed based on impressions of query strings, and further refined after SERPs were examined. We found that by first impression, longer queries led to higher query quality for FF information problems; this was supported by the concept of "specificity" solicited from subjects during the interviews. This result also aligns with Hauff et al. [10] who found that assessors' query quality ratings were moderately correlated with the pre-retrieval predictor *SumSCQ* which assigns higher quality scores to more specific queries and with Lioma et al. [12] whose subjects identified vague and short queries as problematic for systems.

We found that subjects rated queries associated with EX information problems higher in quality compared with FF queries of the same length. Subjects' post-SERP evaluations of queries for EX information problems were also higher than for FF problems. These results might be explained by the vagueness of EX information problems as mentioned by subjects; subjects probably did not pose as strict evaluation standards on open information needs than on more defined information needs. It is also likely that subjects viewed themselves as more open to ideas when searching for EX information problems. This suggests that perhaps query recommendations would have the greatest potential for these types of information problems.

When examining how result quality affected query quality, it appeared that *interactive precision,* the number of relevant search results found to the number of documents viewed was the best predictor of post-SERP ratings of query quality for both types of information problems. This is in contrast with what we expected, which is that reciprocal rank would be a better predictor of subjects' post-SERP evaluations of query quality for FF information problems. Rather, our finding implies that regardless of information problem type, people expected that a good quality query would retrieve more than one relevant result.

For FF information problems, we found that query length was positively correlated with subjects' initial query ratings as well as their post-SERP ratings. We also observed that in most cases, subjects' initial ratings of query quality did not significantly differ from their post-SERP ratings. Only when a relatively long query was followed by a low quality SERP and when a relatively short query was followed by a good SERP did subjects change their ratings of queries. In cases of EX information problems where query length was held constant, people were also more consistent in how they adjusted their ratings when no relevant results or when many relevant results were found.

With respect to evaluation approaches, we found that subjects either took a selective or exhaustive approach when evaluating queries for FF information problems, and a selective, persistent or

exhaustive approach when evaluating queries for EX information problems. This was slightly contrary to what we expected; we expected that most subjects would take a selective approach when evaluating FF information problem queries since the resolution of such problems only requires a single result. Subjects who took an exhaustive approach when evaluating FF problems assigned lower ratings to queries than those who took a selective approach, while the reverse was true for EX problems. In the Exit Interviews, subjects did not differentiate between FF and EX information problems, which was surprising since much is made about the differences in these types of tasks in the research community.

Finally, we found that the initial quality evaluations subjects gave to queries were fairly stable, changing very little during the reliability check. However, their post-SERP evaluations of queries for both types of information problem significantly increased during the reliability check. We found that subjects viewed significantly fewer documents and even clicked on different documents. While this might have been caused by fatigue, it is an open question as to whether query quality evaluations are stable or, if like ratings of other experience goods they are, by nature, brittle.

This study had several limitations; perhaps the most important was the limited number of topics and queries. This study was an initial exploration of this problem space. Using only one query per information problem and holding query length constant for information problems was necessary so that the number of experimental conditions would be manageable. Using a larger combination would have compromised our abilities to collect qualitative data, through interviews, about subjects' ratings. In the future, collecting a larger number of ratings for a larger number of topics and queries would likely enhance our understanding of users' evaluations of query quality.

6. CONCLUSION

Our work explored how people make judgments about query quality; how these judgments are related to features of the query, information problem and search results; how reliable these judgments are; and what people's perceptions are of query recommendation. To our knowledge, this was one of the first systematic studies of how people make query quality evaluations. Our findings provide a useful starting-point for future user-oriented studies of query quality evaluation and recommendation, and might also provide fodder for those working on automatic QPP methods. Future research will examine mechanisms for allowing users to express and share query quality ratings, and develop models of users' decision-making processes regarding query recommendations.

7. ACKNOWLEDGMENTS

We thank Max Felsher for his assistance with data collection.

8. REFERENCES

[1] Amershi, S., & Morris, M. R. (2008). Cosearch: A system for co-located collaborative Web search. *Proceedings of CHI '08*, 1647-1656.

[2] Baraglia, R., Cacheda, F., Carneiro, V., Fernandez, D., Formoso, V., Perego, R., & Silvestri, F. (2009). Search shortcuts: A new approach to recommendation of queries. *Proceedings of RecSys '09*, New York, NY, 77-84.

[3] Buscher, G., Dumais, S. & Cutrell, E. (2010). The good, the bad and the random: An eye-tracking study of ad quality in Web search. *Proceedings of SIGIR '10*, 42-49.

[4] Carmel, D., & Yom-Tov, E. (2010). Estimating the query difficulty for information retrieval. *Synthesis Lectures on Information Concepts, Retrieval and Services, #15*.

[5] Debowski, S., Wood, R. & Bandura, A. (2001). The impact of guided exploration and enactive exploration on self-regulatory mechanisms and information acquisition through electronic enquiry. *Journal of Applied Psychology, 86*(6), 1129-1141.

[6] Duan, W., Gu, B., & Whinston, A. B. (2008). Do online reviews matter? An empirical investigation of panel data. *Decision Support Systems, 45*, 1007-1016.

[7] Ekstrand, M. D., Riedl, J. T., & Konstan, J. A. (2010). Collaborative filtering recommender systems. *Foundations and Trends in Human-Computer Interaction, 4*(2), 81-173.

[8] Hauff, C., Hiemstra, D., & de Jong, F. (2008). A survey of pre-retrieval query performance predictors. *Proceedings of CIKM '08*, 1419-1420.

[9] Hauff, C., Kelly, D., & Azzopardi, L. (2010). Query quality: User ratings and system predictions. *Proceedings of SIGIR'10*, 743-744.

[10] Hauff, C., Kelly, D., & Azzopardi, L. (2010). A comparison of user and system query performance predictors. *Proceedings CIKM '10*, 979-988.

[11] Kelly, D., Cushing, A., Dostert, M., Niu, X., & Gyllstrom, K. (2010). Effects of popularity and quality on the usage of query suggestions during information search. *Proceeding of CHI '10*, 45-54.

[12] Lioma, C., Larsen, B., & Schutze, H. (2011). User perspectives on query difficulty. *Proceedings of ICTIR*, 3-14.

[13] Nelson, P. (1970). Information and consumer behavior. *Journal of Political Economy, 78*(2), 311-329.

[14] Senecal, S. & Nantel, J. (2004). The influence of online product recommendations on consumers' online choices. *Journal of Retailing, 80*(2), 159-169.

[15] Silvestri, F. (2010). Mining query logs: Turning search usage data into knowledge. *Foundations and Trends in Information Retrieval, 4*(1-2), 1-174.

[16] Smyth, B., Balfe, E., Freyne, J., Briggs, P., Coyle, M., & Boydell, O. (2004). Exploiting query repetition and regularity in an adaptive community-based Web search Engine. *User Modeling and User-Adapted Interaction, 14*(5), 382-423.

[17] Smyth, B., Coyle, M. & Briggs, P. (2011). Communities, collaboration, and recommender systems in personalized Web search. *Recommender Systems Handbook*, 579-614.

[18] Turpin, A. & Hersh, W. (2004). Do clarity scores for queries correlate with user performance? *Proceedings of the 15th Australasian Database Conference (ADC '04)*, 85-91.

[19] Zhao, Y., & Scholer, F. (2007). Predicting query performance for user-based search tasks. *Proceedings of the Australasian Database Conference (ADC '07)*, 112-115.

[20] Zhu, H., Huberman, B., & Luon, Y. (2012). To switch or not to switch: Understanding social influence in online choices. *Proceedings of CHI '12*, 2257-2266.