# Effects of Rank and Precision of Search Results on Users' Evaluations of System Performance

Diane Kelly, Xin Fu, Chirag Shah
University of North Carolina
100 Manning Hall, CB#3360
Chapel Hill, NC 27599-3360 USA
+1 919.962.8065
[dianek | fu | chirag] @ email.unc.edu

## ABSTRACT
Previous research has demonstrated that system performance does not always correlate positively with user performance, and that users often assign positive evaluation scores to systems even when they are unable to complete tasks successfully. This paper investigates the relationship between actual system performance and users' perceptions of system performance by manipulating the level of performance experienced by users and measuring users' evaluations of system performance. Eighty-one subjects participated in one of three laboratory studies. The first two studies investigated the impact of the location (or rank order) of five relevant and five non-relevant documents in a search results list containing ten results. The third study investigated the impact of varying levels of precision (.30, .40, .50 and .60) of a search results list containing ten results. Results demonstrate statistically significant relationships between precision and subjects' evaluations of system performance, and ranking and subjects' evaluations of system performance. Of the two, precision explained more variance in subjects' evaluation ratings and was a stronger predictor of subjects' ratings. Finally, the number of documents subjects examined significantly influenced their evaluations, even when the difference was a single document.

## Categories and Subject Descriptors
H.1.2 [**Information Storage and Retrieval**]: Models and Principles – User/Machine Systems – Human factors

## General Terms
Performance, Experimentation, Human Factors

## Keywords
Ranking, precision, user perception, evaluation, search results, search quality, performance

## 1. INTRODUCTION
Several researchers have demonstrated that traditional system performance measures do not always correlate positively with user performance [2, 5, 13, 14]. Hersh, et al. [5] investigated two information retrieval (IR) systems within the context of the TREC-8 Interactive Track which differed only by their term weighting schemes. One of these schemes was demonstrably better when evaluated in a traditional batch-mode framework using average precision. However, when evaluated in an interactive IR framework with users, there were no significant differences in users' performances with each system.

In a follow-up study, Turpin & Hersh [13] conducted analyses using data from the previous study, as well as data from their TREC-9 study where a question-answering task was used, to understand why system performance and user performance were not correlated. Turpin & Hersh evaluated users' queries and the documents retrieved by these queries, rather than the documents saved by users, and found that differences in system performance, as measured by MAP, held. Essentially, users' relevance judgments did not match TREC assessors' relevance judgments, which is why performance differences were not observed. Turpin & Hersh also found that users often retrieved relevant documents, but did not open them, presumably rejecting them based on their titles alone.

Allan, et al. [2] investigated system and user performance at the document passage level by creating artificial answer lists, whose *bpref* scores varied, by blending retrieval results from a state-of-the-art retrieval system and documents marked relevant by assessors. Results demonstrated that differences in *bpref* could result in statistically significant performance differences for users, but only at certain ranges (.50 to .98). These ranges corresponded to recall values of .38 to .56 for users, and even at the highest levels of *bpref*, recall was only about 60%. This result was similar to Turpin & Hersh's [13] in that users' relevance assessments were not entirely aligned with TREC assessors'.

Turpin & Scholer [14] also created artificial lists of retrieval results that corresponded to varying levels of MAP and investigated users' performances with these varying lists when solving a precision-based task (find one relevant document) and a recall-based task (find as many relevant documents as possible within five minutes). Results demonstrated that there was no correlation between system performance and user performance on the precision task, and only a small improvement in performance was detected on the recall task. As in previous studies, TREC relevance assessments were used to evaluate subjects' performances.

While these previous studies provide some evidence that differences in system performance do not always translate into differences in user performance (or that users' relevance judgments are different from TREC assessors'), they do not show how differences in system performance impact users' evaluations of system performance.

Several researchers have observed that system performance does not always correlate positively with users' subjective evaluations of systems, even though users claim that factors such as precision and time influence their ratings of performance [10]. For instance, in a study of *Inquirus*, a Web meta-search tool

developed by the NEC Research Institute, Spink [8] found no correlation between precision and subjective measures of performance, such as those collected via usability questionnaire. However, this study was not designed to explicitly investigate the relationship between actual system performance and users' evaluations of system performance.

Lee, et al. [7] conducted a study to investigate if users could distinguish between document lists that contained varying levels of precision, but did not find any significant results; however, this study had only 10 users and suffered from the same problems as previous studies in that users' relevance assessments did not match the benchmark assessments used to create the experimental lists. Thomas & Hawking [11] found that users could distinguish between sets of high- and low-quality results when presented side-by-side. However, the purpose of this study was to valid an evaluation protocol, rather than to explicitly investigate the relationship between system performance and users' system evaluations. Subjects provided preference data for two lists of search results, which, in most cases, had been created using Google results, rather than systematically manipulated to reflect particular precision values.

In batch-mode IR studies, small differences in performance very often result in significant differences. However, it is unclear if such differences matter to users. The primary purpose of this research is to investigate the relationship between actual system performance and users' evaluations of system performance. A secondary aim is to develop an experimental paradigm that can be used to isolate and study specific aspects of the search process. The goal is to control users' search experiences as much as possible while maintaining a realistic search environment (albeit in the laboratory). One limitation of traditional laboratory-based interactive IR evaluation paradigms, such as those established by the TREC Interactive Track [4] and others [3, 12] is that it is often difficult to isolate and measure individual aspects of the search process and interaction. Other protocols such as the one established by Thomas & Hawking [11] are designed to evaluate systems in more natural environments, and often trade some elements of control for greater ecological validity. As the previous studies showed [2, 13, 14], one difficulty with attempting to control users' search experiences is the subjective nature of relevance. In this study, we attempt to control all aspects of users' search experiences including their relevance assessments through manipulation.

## 2. METHOD

To investigate our research question, we conducted three separate studies. In each study system performance was treated as an independent variable and was manipulated so that subjects experienced particular system performances. System performance was operationalized in two ways: the rank order of search results and the precision of search results. The first two studies examined the impact of rank on subjects' evaluations of system performance and the third examined the impact of precision on subjects' evaluations of system performance.

### 2.1 Instruments and Procedure

During the studies, subjects were instructed that they would be helping the researchers evaluate four search engines. Four topics were provided to subjects, one for each 'search engine,' and subjects were asked to read the topic and pose a single query to a search engine. However, there were no actual search engines used in this study and no real searching took place. We manipulated what was retrieved in response to subjects' queries, so regardless of what query subjects entered, search results were pre-determined. System performance manipulations are described in Section 2.2.

Subjects were told that we were interested in optimizing retrieval to a single query and were not allowed to modify their queries or pose additional queries. Javascript was used to insure that subjects' queries were between 5 and 55 characters. On average, queries were 3.27 words. An examination of subjects' queries did not reveal any cases where subjects entered nonsense queries. All queries were appropriate for each topic and there were no queries that would have caused suspicion over the retrieved search results.

The search interface, search results page and full-text display is shown in Figures 1-3. We modeled the design and layout of these instruments after Google. Search engines were named after four colors (yellow, blue, green and orange) to help subjects distinguish among them. These colors were held constant so that the first search engine that subjects used was always yellow, the second always blue, etc., even though the different levels of system performance were randomized across position.


**Figure 1. Interface for blue (second) search engine**


**Figure 2. Search results list**


**Figure 3. Full-text document display with relevance indicator**

Each search engine returned ten results. Subjects were asked to examine all search results in the order in which they were presented; only one title in the search result list was hyperlinked at any time and subjects had to submit a relevance judgment for a document before a hyperlink to the next result became active. A place for subjects to indicate relevance was found at the bottom of each document. Relevance was binary and subjects were provided

with the choices of "Relevant" and "Not Relevant." We asked subjects to evaluate ten search results because studies have shown that Web users typically only examine the first page of search results, which usually contain ten results [9].

After subjects evaluated documents for the first search engine, they were asked to evaluate the performance of the search engine. Subjects were instructed, "Now that you have used the first engine, please rate the quality of the search results" and provided with a 10-point scale, with the anchors "very poor" and "very good." For subsequent search engines, subjects were presented with their previous ratings as well as a place to evaluate the current search engine. Subjects were allowed to change their previous ratings if they wished. The interface for evaluating the third search engine is presented in Figure 4. The purposes of allowing subjects to change their ratings of previous search engines were to capture relative ratings and to allow subjects to calibrate the scale according to their experiences. We also believed that viewing these ratings helped remind subjects of their previous experiences. Background colors of each box were identical to search engine colors.

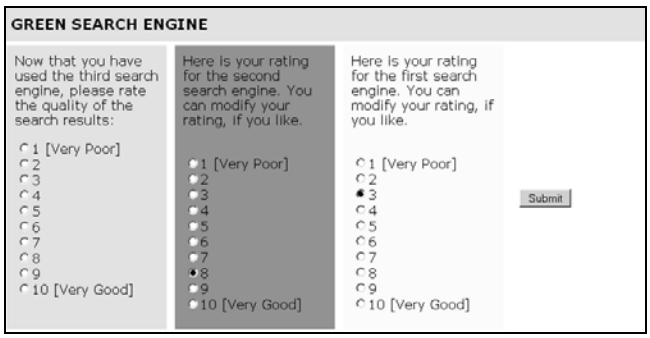


**Figure 4. Interface for evaluating the performance of the green (third) search engine**

## 2.2 System Performance Manipulation

In Studies 1 and 2, system performance was operationalized as the rank order of search results; this variable had four levels and was a within subjects variable. The levels of system performance were presented randomly to subjects and each subject experienced all four performance levels. We held the precision of the search results constant, and set this value to .50. Thus, for each performance level, there were a total of five relevant documents and five non-relevant documents in the search results list. Documents were identical across each level. The only difference was the location of documents in the search results list. Each 'search engine,' corresponded to one of the four performance levels in Table 1. The precision value represents a medium-level precision and allowed us to have symmetrical lists, so that we could investigate the impact of having five relevant documents first (positions 1-5 – the theoretically best performance) and last (positions 6-10 – the theoretically worst performance). Levels 3 and 4 are inverses of each other; we did not want to alternate relevant and non-relevant documents since subjects may have detected this pattern. Finally, for each topic, relevant and non-relevant documents (see Section 2.3) were held constant, so that the document represented by "R" at Performance Level 1, Position 1 was identical to the "R" document at Performance Level 2, Position 6, Performance Level 3, Position 1, and Performance Level 4, Position 2.

**Table 1. Performance levels for Studies 1 and 2 (R: relevant and NR: non-relevant)**

| | | Performance Level | | | |
|---|---|---|---|---|---|
| | | **1** | **2** | **3** | **4** |
| **Position in Search Results List** | **1** | R | NR | R | NR |
| | **2** | R | NR | NR | R |
| | **3** | R | NR | NR | R |
| | **4** | R | NR | R | NR |
| | **5** | R | NR | R | NR |
| | **6** | NR | R | NR | R |
| | **7** | NR | R | R | NR |
| | **8** | NR | R | NR | R |
| | **9** | NR | R | R | NR |
| | **10** | NR | R | NR | R |

The only difference between Studies 1 and 2 was the instructions provided to subjects. In Study 1, subjects were instructed to examine each of the 10 documents without any indication of how many relevant documents were in the set. In Study 2, subjects were instructed to find five relevant documents. Subjects were still required to examine documents in the order in which they were presented, but once subjects marked the fifth relevant document, they were taken to the search engine evaluation page. For instance, in the Level 1 performance condition, subjects would theoretically only need to examine five documents, while in Level 2 subjects would need to examine all ten. The first set of instructions holds time constant (to a certain extent) since subjects have to examine all 10 documents for all 4 search engines. In the second set of instructions, time varies; subjects' experiences with some search engines will last longer than their experiences with others. We hoped that this would allow us to investigate the impact of time on subjects' evaluations of systems.

In Study 3, system performance was operationalized as the precision of the search results. This variable was also a within subjects variable with four levels, corresponding to the following precision scores: .30, .40, .50 and .60. We selected these precision scores because they represent a mid-range of scores. We wanted to use four sequential scores and did not want to select scores that were too poor or too good. We believe that these precision scores represent a realistic range since many state-of-the-art search systems yield precision scores in this range [2].

Manipulating the precision of search results was a challenge because the only way to manipulate the precision of a list of documents where the total number retrieved remains constant is to add or subtract relevant documents. This not only changes the precision value, but also the total number of relevant documents in the list. Making a decision about the position of documents in the search results lists also presented some problems. We were interested in holding the relevance of the document at position 1 constant to minimize any potential bias [6] and we needed to identify a way to order the documents to minimize the impact of the presentation order (or rank) of relevant and non-relevant documents, while varying precision.

To accomplish this, we first created a table that contained all possible document rankings for each of the four precision values. Next, for each list we computed the precision at each document (i.e., precision at 1, precision at 2, precision at 3, etc.), and then

averaged these values (10 total values). We were unable to use a strict average precision measure since this measure assumes a fixed number of relevant documents for any given topic, so instead we call this measure mean average precision at 10, or MAP@10. This measure also differs from average precision in that precision values are computed at each document, whether the document is relevant or not relevant. The rank lists that we used in this study are displayed in Table 2. The corresponding precision and MAP@10 scores are also listed in this Table. The maximum difference between the MAP@10 scores of any two lists was .0006, while the difference between the best and worst precision scores was .0001. Our goal in selecting these lists was to minimize differences in MAP@10.

**Table 2. Performance levels, precision and MAP@10 scores for Study 3 (R: relevant, NR: non-relevant)**

| | | Performance Level | | | |
|---|---|---|---|---|---|
| | | **1** | **2** | **3** | **4** |
| **Position in Search Results List** | **1** | R | R | R | R |
| | **2** | NR | R | R | R |
| | **3** | R | NR | NR | R |
| | **4** | NR | NR | R | NR |
| | **5** | R | R | NR | NR |
| | **6** | R | NR | NR | NR |
| | **7** | NR | R | NR | NR |
| | **8** | R | NR | R | NR |
| | **9** | NR | NR | NR | NR |
| | **10** | R | R | NR | NR |
| **Precision** | | .60 | .50 | .40 | .30 |
| **MAP@10** | | .6285 | .6283 | .6289 | .6285 |

## 2.3 Topics and Documents

The selection of topics and documents for this study was an extremely important step. The entire experiment was contingent on subjects experiencing particular performance levels. This, of course, was contingent on subjects making particular relevance assessments of individual documents. Therefore, it was necessary for the researchers to select documents that had a very high probability of being judged relevant or not relevant for a given topic by a large number of people. In order for this experimental paradigm to succeed, it was critical for subjects to make relevance assessments that were identical to the TREC assessments.

To select topics and documents for this study, we started with the TREC[1] HARD 2005 collection [1]. This collection consists of 50 topics and a 3GB corpus of newswire text, drawn from three sources: Xinhua News Service (1996-2000), New York Times News Service (1998-2000), and Associated Press Worldstream News Service (1998-2000). We selected newspaper articles because they are generally easy to understand and read, and are relatively short in length. Subjects were told that they would use our search engines to search a collection of newspaper articles.

Topics were in the standard TREC format with a title, description and narrative. Although topics were assigned to subjects, we selected general topics that we believed would interest our target subjects and that had relevance to current events. Another criterion for selecting topics was that each topic needed at least six relevant documents in the corpus (to fulfill the requirements of the Study 3 condition where precision=.60). The topics we selected are displayed in Table 3. The narrative, which is not displayed below provided further specification about what constituted a relevant document. For instance, the narrative for the first topic was, "Any document identifying an instance where a journalist or correspondent has been killed, arrested or taken hostage in the performance of his work is relevant."

**Table 3. Topics used in the study, with TREC topic number, title and description**

| Topic | Title | Description |
|---|---|---|
| 354 | Journalist Risks | Identify instances where a journalist has been put at risk (e.g., killed, arrested or taken hostage) in the performance of his work. |
| 374 | Nobel Prize Winners | Identify and provide background information on Nobel prize winners. |
| 408 | Tropical Storms | What tropical storms (hurricanes and typhoons) have caused significant property damage and loss of life? |
| 448 | Ship Losses | Identify instances in which weather was a main or contributing factor in the loss of a ship at sea. |

The TREC collection includes a set of relevance judgments that have been made by TREC assessors for each topic. The assessments are made on approximately 1000 documents retrieved by various systems that were part of the original TREC experiment. We used these assessments to identify candidate relevant and non-relevant documents for each topic. The three researchers examined lists of relevant and non-relevant documents independently for each topic to identify documents that were believed to be unambiguously relevant and not relevant. That is, documents that we felt were obviously relevant or not relevant. We also wanted to select non-relevant documents that were realistic; that is, documents that subjects could believe may have been retrieved in response to their queries, but that were still not relevant. We took the overlap of our lists as the set of documents we used in this study. This proved to be a very challenging process: we went through several iterations, with multiple lists and topics before we arrived at the set of topics and documents we used in this study. These topics and documents were piloted during a test of Study 1 and the subject assessed the relevance of each document as we anticipated and afterwards did not indicate confusion with any topic, document or relevance judgment.

## 2.4 Subjects

Subjects were recruited by sending an email solicitation to undergraduate students at our university. Subjects participated in 1 of 16 study sessions. The one-hour sessions typically contained around 5-8 subjects and were conducted in a computer laboratory with 30 workstations. In each session, subjects participated in the same study (Study 1, 2 or 3) and were compensated with $10.00 USD. Studies were assigned to sessions so that an approximately equal number of subjects would participate in each study.

Eighty-one subjects participated in this experiment (27 per study). Fifty-nine subjects were female and 21 were male (one subject skipped this question). Subjects had a mean age of 20 years. Twelve percent of the subjects were humanities majors, 29% were

[1] Text Retrieval Conference (http://trec.nist.gov)

social science majors, 22% were science majors, 33% were in a professional school (e.g., business) and 4% were undecided. Subjects' mean search experience was 3.31 (where 1=very inexperienced and 4=very experienced) and subjects indicated that they searched the Web for information very frequently (mean=3.91, where 1=less than monthly and 4=daily). There were no significant differences in age, sex, search experience and search frequency among subjects in Studies 1, 2 and 3.

# 3. RESULTS

Results are reported in two sections. The first section describes subjects' relevance assessments of documents and the extent to which our experimental manipulations worked. The second section describes the relationship between actual system performance and subjects' evaluations of system performance.

## 3.1 Relevance Assessments

Table 4 shows the amount of agreement we observed in subjects' relevance assessments and the benchmark relevance assessments. This Table displays an agreement fraction, the corresponding number of misjudged documents, and the frequency of these agreement fractions for subjects in each study. The computations for Studies 1 and 3 are based on subjects' evaluations of 40 documents, while the base for Study 2 varies, depending on the number of documents subjects evaluated. On average, 90% agreement was observed between subjects' assessments and the actual relevance assessments. The highest levels of agreement were observed in Study 3.

**Table 4. Relationship between subjects' relevance assessments and benchmark relevance assessments**

| | | No. of Subjects | | |
|---|---|---|---|---|
| **Agreement** | **No. Missed** | **Study 1 (n=27)** | **Study 2 (n=27)** | **Study 3 (n=27)** |
| 1.0 | 0 | 1 | 2 | 1 |
| .95-.99 | 1-2 | 6 | 5 | 11 |
| .90-.94 | 3-4 | 9 | 8 | 10 |
| .85-.89 | 5-6 | 7 | 4 | 2 |
| .80-.84 | 7-8 | 4 | 4 | 3 |
| .75-.79 | 9-10 | 0 | 2 | 0 |
| .70-.74 | 11-12 | 0 | 2 | 0 |
| **Mean Agreement** (Stand. Deviation) | | .90 (.10) | .89 (.12) | .92 (.11) |

Was there less agreement in one study condition than another? Table 5 shows the agreement fractions for each condition. These conditions correspond to those listed in Tables 1 and 2 (note that Study 3's conditions were different than Studies 1 and 2's). Overall, the agreement fractions among conditions were similar.

Was there less agreement for one topic than another? Table 6 shows the agreement fractions for each topic. Within each study, the agreement fractions for Topics 354, 374 and 408 were similar. The most agreement was observed with Topic 448, "Ship Losses," which indicates that there was less ambiguity in the documents that subjected evaluated for this topic. The least agreement was with Topic 354, "Journalist Risks," indicating perhaps more ambiguity in one or more of the documents for this topic.

**Table 5. Mean (standard deviation) agreement for conditions**

| Condition | Study 1 | Study 2 | Study 3 |
|---|---|---|---|
| 1 | .90 (.10) | .88 (.14) | .90 (.12) |
| 2 | .89 (.11) | .87 (.11) | .92 (.13) |
| 3 | .90 (.09) | .89 (.12) | .94 (.08) |
| 4 | .91 (.10) | .91 (.12) | .93 (.10) |

**Table 6. Mean (standard deviation) agreement for topics**

| Topic | Study 1 | Study 2 | Study 3 |
|---|---|---|---|
| 354 | .88 (.08) | .86 (.13) | .88 (.11) |
| 374 | .88 (.13) | .88 (.12) | .92 (.11) |
| 408 | .90 (.11) | .88 (.13) | .93 (.12) |
| 448 | .95 (.06) | .95 (.07) | .95 (.07) |

To better understand these disagreements, we computed agreement fractions for each document. Table 7 shows the agreement fractions that were observed for the 52 documents used in this study. An agreement of 1.0 was observed for 15 (29%) of the documents (i.e., these documents were marked correctly by all subjects). Table 7 shows that certain documents were problematic with respect to our experimental manipulations – in particular, 6 documents had agreement fractions of less that .80. Out of these 6 documents, two were associated with Topic 354 and two with Topic 374, which explains some of the differences observed in Table 6. A thorough failure analysis of why these discrepancies occurred is not possible here, but a preliminary examination showed that the frequently misjudged documents for Topic 354 were about journalists who committed crimes (which we suppose puts them at risk) and those for Topic 374 were of a non-standard genre or format.

**Table 7. Agreement fractions for documents**

| Agreement | | | | | | |
|---|---|---|---|---|---|---|
| **1.0** | **.99-.95** | **.94-.90** | **.89-.85** | **.84-.80** | **.79-.75** | **<.75** |
| 15 | 12 | 10 | 4 | 5 | 3 | 3 |

Before we analyzed subjects' evaluations of system performance, we wanted to look at subjects' individual relevance assessments for each condition to see if some subjects' assessment behaviors were anomalous. We looked at agreement per condition (i.e., 'search') rather than overall agreement since it provides a more refined view of subjects' assessment behavior. For any given condition, the majority of subjects misjudged 0-2 documents. There were 12 instances where subjects misjudged 3 documents and 12 where subjects misjudged 4-5 documents. Because we were concerned about the integrity of the experiment, we decided to exclude subjects who misjudged 4 or more documents in any one condition from subsequent analyses since the experimental manipulations clearly did not work for these subjects, or they did not take the experiment seriously. We also looked at the amount of time these subjects spent completing the experiment and their system evaluations to confirm the appropriateness of their exclusion. In most cases, we found that these subjects completed the study much faster than others and/or had system evaluations that were questionable. For example, one subject in Study 2, Condition 2 marked 5 of the first 6 documents relevant. Not only were his relevance assessments incongruent with the actual assessments, he assigned the lowest possible score to this system

despite its high precision. Based on our elimination criteria, 3 subjects were eliminated from Study 1, 4 from Study 2 and 4 from Study 3.

## 3.2 Evaluations of System Performance

Recall that after using each 'search engine' subjects were asked to evaluate the performance of the system. Subjects were also given the opportunity to modify their previous evaluations. Table 8 shows the number of subjects who changed their ratings 0, 1, 2, 3, 4 or 5 times. Overall, most subjects did not change their ratings. In Studies 1, 2, 3 the percentages of subjects who did not change their ratings were: 58%, 57% and 44%, respectively. In Studies 1 and 2, about 20% of the subjects changed one rating. In Study 3, there were slightly more changes, with one person making 5 changes. We used the last evaluation score assigned by subjects to each 'search engine' in the analyses reported in this section.

**Table 8. Frequency with which subjects changed their evaluation ratings (numbers in cells represent subjects)**

| Freq. | Study 1 | Study 2 | Study 3 | Total |
|---|---|---|---|---|
| 0 | 14 (58%) | 13 (57%) | 10 (44%) | 37 (53%) |
| 1 | 5 (21%) | 5 (21%) | 6 (26%) | 16 (23%) |
| 2 | 3 (17%) | 2 (9%) | 4 (17%) | 10 (14%) |
| 3 | 1 (4%) | 2 (9%) | 2 (9%) | 5 (7%) |
| 4 | - | 1 (4%) | - | 1 (1.5%) |
| 5 | - | - | 1 (4%) | 1 (1.5%) |

The primary focus of Studies 1 and 2 was to examine the impact of the rank order of documents on subjects' evaluations of system performance while holding precision constant at .50. The primary focus of Study 3 was to examine the impact of precision of search results lists on subjects' evaluations of system performance, while holding rank as constant as possible. Because the experimental manipulations were not 100% effective and some subjects experienced different levels of precision and ranking than intended, we could not simply use condition as an independent variable. Instead, we computed the precision and MAP@10 that each subject experienced based on his relevance assessments, categorized these values into bins (for example, a value of .28 would be categorized into the .20 bin) and used these categorical variables as independent variables. Table 9 shows the overall precision values that were observed in each study, the number of times these values were observed and the mean evaluation rating assigned by subjects. Shaded cells correspond to target precision values in each study (those associated with the four original conditions). Figure 5 shows the relationship between precision and mean evaluation rating.

The general trend is that as precision increases, evaluation scores increase. There are a few anomalies in the data, which we believe are in part due to the small cell sizes. For instance, in Study 1 the mean evaluation rating for the two subjects who experienced a precision of .20 was 4.00, which is higher than those subjects experiencing a precision of .30. The standard deviation at precision .20 indicates that these two subjects' evaluations ranged from 2 to 6; there is a good chance that more data points would have smoothed this distribution. What is clear from Table 9 is that evaluations corresponding to a precision value of .50 vary between 5 and 6, which represent mid-range scores, while evaluation scores associated with precision values less than .50 range from 2 to 5 and those associated with precision values above .50 range from 6 to 9.

**Table 9. Mean (standard deviation) evaluation according to experienced precision**

| | | Study 1 | | Study 2 | | Study 3 | |
|---|---|---|---|---|---|---|---|
| | | N | Mean (SD) | N | Mean (SD) | N | Mean (SD) |
| Experienced Precision | .1 | 0 | - | 0 | - | 1 | 3.00 (-) |
| | .2 | 2 | 4.00 (2.83) | 1 | 4.00 (-) | 3 | 2.67 (.58) |
| | .3 | 6 | 2.83 (.98) | 9 | 4.67 (1.66) | 18 | 3.67 (1.57) |
| | .4 | 24 | 4.29 (1.30) | 21 | 4.86 (1.59) | 21 | 4.86 (1.65) |
| | .5 | 40 | 5.35 (1.59) | 23 | 5.78 (1.59) | 24 | 6.33 (1.61) |
| | .6 | 15 | 6.87 (1.51) | 15 | 7.00 (1.65) | 17 | 7.59 (1.50) |
| | .7 | 7 | 6.57 (1.72) | 6 | 6.83 (1.84) | 5 | 8.20 (.84) |
| | .8 | 2 | 7.50 (.71) | 3 | 9.67 (.58) | 2 | 8.50 (.71) |
| | .9 | 0 | - | 0 | - | 0 | - |
| | 1 | 0 | - | 11 | 9.00 (1.01) | 0 | - |
| Total | | 96 | 5.27 (1.84) | 92 | 6.29 (2.12) | 91 | 5.69 (2.21) |

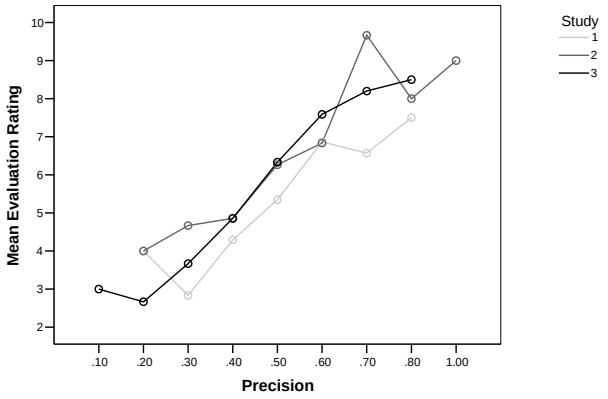


**Figure 5. Mean evaluation according to experienced precision**

One-way analysis of variance tests demonstrated that results of all three studies were statistically significant: Study 1 [$F(6, 89) = 9.02$, $p<.000$]; Study 2 [$F(7, 91) = 11.22$, $p<.000$]; and Study 3 [$F(7, 83) = 14.60$, $p<.000$]. Scheffe's follow-up tests were conducted to evaluate pair-wise differences among the means. Table 10 displays results of these tests. These results show that scores clustered into two groups in all three studies. For instance, in Study 1, no significant differences were observed among precision scores of .20-.50 or .50-.80, but these groups of scores were significantly different from one another. For Studies 1 and 2, evaluations scores associated with precision of .50 and .60, respectively, could be placed in either group. That is, there were no significant differences between evaluation scores associated with these precision values and those associated with other precision values.

**Table 10. Results of Scheffe's post-hoc tests for precision**

| Study | Group 1 | Group 2 |
|---|---|---|
| 1 | .20 - .50 | .50 - .80 |
| 2 | .30 - .60 | .60 – 1.0 |
| 3 | .20 - .40 | .50 - .80 |

The other variable that we explored in relation to subjects' ratings was the rank order of the search results. Table 11 shows the overall MAP@10 values that were observed in each study, the number of times these values were observed and the mean evaluation rating assigned by subjects. Shaded cells correspond to

MAP@10 for the original conditions. Figure 6 illustrates the relationship between MAP@10 and mean evaluation rating.

**Table 11. Mean (standard deviation) evaluation according to MAP@10**

| | | Study 1 | | Study 2 | | Study 3 | |
|---|---|---|---|---|---|---|---|
| | | N | Mean (SD) | N | Mean (SD) | N | Mean (SD) |
| MAP@10 | .1 | 14 | 3.86 (1.41) | 15 | 4.73 (1.71) | 0 | - |
| | .2 | 6 | 4.33 (.82) | 4 | 4.75 (1.26) | 0 | - |
| | .3 | 6 | 5.17 (1.94) | 7 | 4.71 (1.60) | 5 | 4.00 (1.41) |
| | .4 | 19 | 4.95 (1.68) | 23 | 6.00 (1.68) | 7 | 3.71 (1.11) |
| | .5 | 19 | 4.89 (1.63) | 18 | 6.44 (1.92) | 5 | 4.80 (2.28) |
| | .6 | 8 | 5.13 (1.46) | 6 | 7.17 (1.84) | 51 | 5.69 (2.20) |
| | .7 | 6 | 7.00 (1.20) | 3 | 6.67 (3.05) | 16 | 6.31 (1.99) |
| | .8 | 16 | 6.75 (1.69) | 0 | - | 4 | 8.25 (.96) |
| | .9 | 2 | 8.50 (.71) | 5 | 8.00 (1.58) | 3 | 8.00 (1.00) |
| | 1 | 0 | - | 11 | 9.00 (1.01) | 0 | - |
| **Total** | | 96 | 5.27 (1.84) | 92 | 6.29 (2.12) | 91 | 5.69 (2.21) |

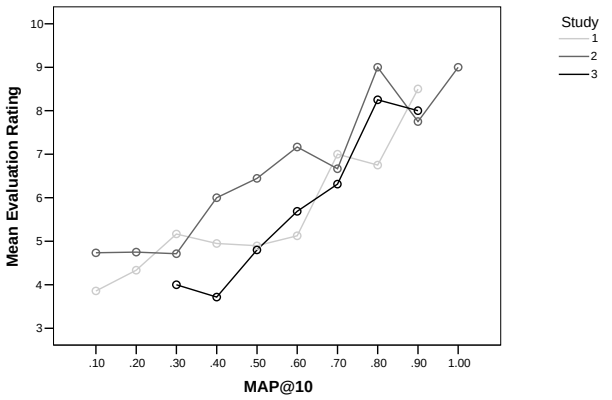


**Figure 6. Mean evaluation according to MAP@10**

This data displays the same general trend as the previous data: as MAP@10 increases, evaluation scores increase. However, this relationship does not appear as strong. One-way ANOVA tests demonstrated that the differences were statistically significant: Study 1 [$F(8, 87) = 5.77$, $p<.000$]; Study 2 [$F(8, 83) = 7.09$, $p<.000$]; and Study 3 [$F(6, 84) = 3.83$, $p<.002$]. Note that the F-values were lower, which again indicates that the strength of the relationship between MAP@10 and evaluation ratings was not as strong as that between precision and evaluation ratings. Table 12 displays results of Scheffe's tests that were conducted to evaluate pair-wise differences among the means. Similarly to the previous results, these scores also clustered into two groups; however, the divisions differ from those in Table 10. In Study 2, in particular, more scores clustered in Group 2 than Group 1.

**Table 12. Results of Scheffe's post-hoc tests for MAP@10**

| Study | Group 1 | Group 2 |
|---|---|---|
| 1 | .01-.06 | .06-.09 |
| 2 | .01-.40 | .40-1.00 |
| 3 | .03-.05 | .05-.09 |

So, which independent variable – precision or rank – explains more of the variance in subjects' ratings? *Eta* squared (*eta*$^2$) was computed for each relationship. *Eta*$^2$ indicates how much of the variance that exists within the dependent variable can be explained by differences within the independent variable. These figures are shown in Table 13. Overall, these values are high for both independent variables, but in each case precision explains more of the variability in subjects' ratings than MAP@10. In each study, precision explained 38%, 48% and 55% of the variance in subjects' ratings and exhibited a stronger influence on subjects' evaluations.

**Table 13. Results of *eta*$^2$ to determine strength of relationships**

| Study | Precision | MAP@10 |
|---|---|---|
| 1 | .38 | .35 |
| 2 | .48 | .41 |
| 3 | .55 | .21 |

Finally, recall that Studies 1 and 2 were identical expect for the instructions provided to subjects. Subjects in Study 1 were required to evaluate all 10 documents, while subjects in Study 2 were able to stop once they found 5 relevant documents. The purpose behind this manipulation was to see if time spent using the 'search engine' had any measurable effects on subjects' evaluation ratings (time is operationalized as the number of documents subjects had to examine). To compare the ratings between these two studies, we analyzed only those cases where our experimental manipulations worked. Although this reduced our sample size, we believe that it was the most valid way to perform this analysis.

Figure 7 displays the mean evaluation ratings (error bars represent +/- one standard deviation) for each condition for each study. Table 14 displays the cell sizes, the expected relationships between the means given the time differential that subjects experienced (assuming less time leads to better ratings), and t-test results. These results demonstrate that subjects in Condition 1 in Study 2 rated the system significantly higher than subjects in Condition 1 in Study 1. The amount of time it took subjects to find the relevant documents in Study 2 seemed to matter greatly; even in the case of Condition 3, where subjects examined one less document in Study 2 than Study 1, there was still a statistically significant difference in subjects' evaluation ratings. Subjects in Conditions 2 and 4 had to examine 10 documents regardless, and results show no significant difference in these means.

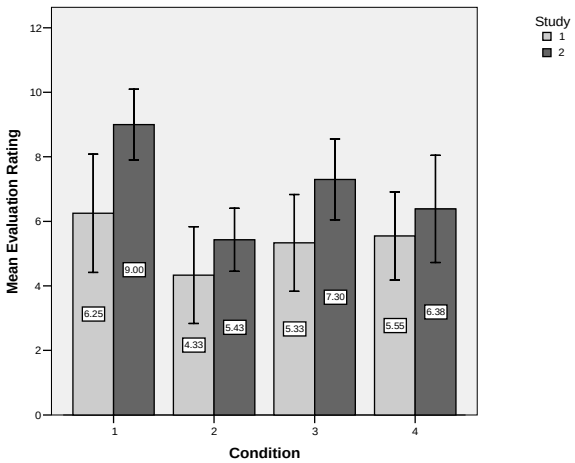


**Figure 7. Mean evaluation for condition and study (error bars represent +/- one standard deviation)**

**Table 14. Cell sizes, expected relationships and t-test results**

| Cond. | Study 1 (n) | Study 2 (n) | Exp. Rel. | T-test Results |
|---|---|---|---|---|
| 1 | 8 | 11 | S1 < S2 | $t(17) = -4.09$, $p=.001$ |
| 2 | 9 | 7 | S1 = S2 | $t(14) = -1.67$, $p=.117$ |
| 3 | 9 | 10 | S1 < S2 | $t(17) = -3.12$, $p=.006$ |
| 4 | 11 | 13 | S1 = S2 | $t(22) = -1.34$, $p=.196$ |

# 4. CONCLUSION

Previous research has demonstrated that system performance does not always correlate positively with user performance, and that users often assign positive evaluation scores to systems even when they are unable to use systems effectively. In this study we investigated the relationship between actual system performance (operationalized as precision and rank) and users' perceptions of system performance. We found statistically significant relationships between precision and subjects' evaluations of system performance, and ranking and subjects' evaluations of system performance. In both cases, post-hoc tests showed that the relationship was not linear, but that evaluation scores clustered into two groups that were split along midrange values of precision and MAP@10. An examination of the strength of these relationships showed that precision influenced subjects' evaluations more strongly than ranking. We also found that the number of documents subjects examined influenced their evaluations, even when the difference was a single document.

The second aim of this study was to develop an experimental paradigm that could be used to isolate and study specific aspects of the search process. Overall, the experimental paradigm developed in this study proved to be an effective method for investigating our research questions. Although our manipulations were only 90% effective, in general the agreement in relevance assessments was much higher than that reported in previous studies [2, 7, 13, 14]. We believe our efforts at selecting relevant and non-relevant documents provided us with greater levels of control and our experiment with greater levels of internal validity. In previous studies, fatigue may have impacted user performance since experimental sessions were about 8 hours in length and users were often presented with result lists containing 100 or more results. In this study, we only required subjects to evaluate a maximum of 40 documents, which allowed for one hour sessions. We believe that this experimental paradigm is extensible and can be used to investigate a number of other issues such as the impact of particular interface features and other behaviors related to system output and performance. We will continue to work on the experimental collection by replacing documents where disagreements occurred and including more topic-document sets.

This study had several limitations. This was a laboratory study and like all laboratory studies, some ecological validity is sacrificed for increased control. Although we tried to simulate Web-based searching, the documents were from newspapers and did not contain familiar Web elements such as navigation bars and images. Our sample was a convenience sample and only consisted of undergraduates. Furthermore, many subjects were probably primarily motivated by the compensation rather than an interest in IR. We were able to identify a few subjects who did not seem to take the study seriously, but there may have been others. Finally, this study only examined two factors which impact users' evaluations of search systems. There are likely other factors that impact users' evaluations of search systems and additional studies are needed to understand their potential impact.

# 5. REFERENCES


[1] Allan, J. (2006). HARD Track overview in TREC 2005 high accuracy retrieval from documents. In E. M. Voorhees & L. P. Buckland (Eds.), *TREC-2005, Proc. of the Text Retrieval Conference*. Washington, D.C.: Government Printing Office.

[2] Allan, J., Carterette, B., & Lewis, J. (2005). When will information retrieval be 'good enough'? *Proc. of ACM SIGIR*, Salvador, Brazil, 433-440.

[3] Borlund, P. (2003). The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research*, *8*(3), no. 152.

[4] Dumais, S. T., & Belkin, N. J. (2005). The TREC Interactive Tracks: Putting the user into search. In E. M. Voorhees & D. K. Harman (Eds.) *TREC: Experiment and Evaluation in Information Retrieval* (pp. 123-153), MIT Press.

[5] Hersh, W., Turpin, A., Price, S., Chan, B., Kraemer, D., Sacherek, L., & Olson, D. (2000). Do batch and user evaluations give the same results? *Proc. of ACM SIGIR*, Athens, Greece, 17-24.

[6] Joachims, T., Granka, L. A., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. *Proc. of ACM SIGIR*, Salvador, Brazil, 154-161.

[7] Lee, H.-J., Belkin, N. J., & Krovetz, B. (2006). Rutgers information retrieval performance evaluation project. *Journal of the Korean Society for Information Management, 23*(2), 98-111.

[8] Spink, A. (2002). A user-centered approach to evaluating human interaction with Web search engines: An exploratory study. *Information Processing & Management, 38*, 401-426.

[9] Spink, A., & Jansen, B. J. (2004). *Web search: Public searching of the Web*. Kluwer Academic Publishers.

[10] Su, L. T. (2003). A comprehensive and systematic model of user evaluation of Web search engines: II. An evaluation by undergraduates. *Journal of the American Society for Information Science & Technology, 54*(13), 1193-1223.

[11] Thomas, P., & Hawking, D. (2006). Evaluation by comparing result sets in context. *Proc. of CIKM*, Arlington, VA.

[12] Toms, E. G., Freund, L., & Li, C. (2004). WiIRE: The Web interactive information retrieval experimentation system prototype. *Information Processing & Management, 40*(4), 655-675.

[13] Turpin, A., & Hersh, W. (2001). Why batch and user evaluations do not give the same results. *Proc. of ACM SIGIR*, New Orleans, LA, 225-231.

[14] Turpin, A., & Scholer, F. (2006). User performance versus precision measures for simple search tasks. *Proc. of ACM SIGIR*, Seattle, WA, 11-18.