

Human Performance and Retrieval Precision Revisited

Mark D. Smucker
Department of Management Sciences
University of Waterloo
msmucker@uwaterloo.ca

Chandra Prakash Jethani
David R. Cheriton School of Computer Science
University of Waterloo
cpjethan@cs.uwaterloo.ca

ABSTRACT

Several studies have found that the Cranfield approach to evaluation can report significant performance differences between retrieval systems for which little to no performance difference is found for humans completing tasks with these systems. We revisit the relationship between precision and performance by measuring human performance on tightly controlled search tasks and with user interfaces offering limited interaction. We find that human performance and retrieval precision are strongly related. We also find that users change their relevance judging behavior based on the precision of the results. This change in behavior coupled with the well-known lack of perfect inter-assessor agreement can reduce the measured performance gains predicted by increased precision.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]

General Terms: Experimentation, Human Factors, Performance

Keywords: Precision, user studies, human performance, Cranfield, evaluation metrics, interaction

1. INTRODUCTION

The Cranfield approach to information retrieval (IR) evaluation measures the ranking quality of a retrieval system given a test collection of documents, search topics, and relevance judgments [4]. Popular evaluation metrics used as part of the Cranfield-style of evaluation are typically some measure reflecting the precision or recall of the ranking produced by the retrieval system.

In the past decade, various studies have offered conflicting reports on the value of the Cranfield approach and its associated metrics for users of interactive retrieval systems. Most notable is perhaps the work of Hersch et al. [7] who asked “Do improvements in system performance demonstrated by batch evaluations confer the same benefit for real users?” Hersch et al. found no statistically significant gains for users

of systems with better batch evaluation performance as measured by mean average precision (MAP). In a followup work, Turpin and Hersch [13] reached a similar conclusion. Subsequent studies have offered conflicting reports on the value of batch evaluation for predicting human performance or satisfaction [1, 2, 11, 12].

In all of these studies, human performance on some task has been compared to measures of precision. From the beginning, it is clear that the attention to recall and precision has met with resistance. Responding to criticism, Cleverdon et al. [5] wrote about recall and precision saying “The unarguable fact, however, is that they are fundamental requirements of the users, and it is quite unrealistic to try to measure how effectively a system or a subsystem is operating without bringing in recall and precision.” Voorhees [14] stresses that performing well on the abstract task represented by a Cranfield-style evaluation is “assumed to be a necessary *but not sufficient* prerequisite for performing well on real search tasks.”

In this paper, we revisit the connection between retrieval precision and human performance by first noting that most Cranfield-style retrieval metrics actually make little to no attempt to include a model of a user or user interface (UI). At most, the precision based metrics have a model of a user as someone who reads the ranked documents, in order, one after the other. As this hypothetical user reads documents, the user takes the time to make a relevance judgment for each document. This user takes a constant amount of time to evaluate the relevance of each document regardless of the document’s length, style, or the search topic. Given what most evaluation metrics do not attempt to do, we do not find it too surprising that they lack strong predictive power for human performance with interactive retrieval systems.

If we are to create evaluation metrics that are predictive of human performance, we must first better understand the relationship between retrieval precision and human performance. Our approach is to first try to understand the simplicity of systems. We can then gradually increase the complexity of the systems studied as our understanding grows.

As such, we conducted a two phase user study that carefully controlled the precision of ranked lists and also the level of user interaction allowed. For both phases, we investigated human performance at two different levels of precision — low and high. In the first phase of the study, the users alternated between judging the relevance of full documents and summaries of documents. This interface forced users to proceed down the ranked list of documents one at a time. This first phase of the experiment models the common as-

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

SIGIR’10, July 19–23, 2010, Geneva, Switzerland.

Copyright 2010 ACM 978-1-60558-896-4/10/07 ...\$10.00.

sumptions of Cranfield-style precision-based metrics minus the assumption that documents take a uniform amount of time to judge. We alternated between summaries and documents to obtain performance data on both while avoiding a learning effect that could be caused by doing either all documents or all summaries first for a given topic. Our interest in summaries comes from their use in modern search interfaces and their use in phase 2.

In the second phase of the study, the interface showed users the ranked list of documents in a fashion similar to modern web search engines that show 10 query-biased document summaries per page. Clicking on a summary allowed the user to save the document if the user thought it was relevant to the given search topic. While similar to web search interfaces, this interface did not provide any means for the user to reformulate queries, or even enter a query. Here the user had to search the provided results list. While we still limited interaction in phase 2, we hypothesized that any difference in human performance at the two levels of precision will be reduced compared to phase 1. In other words, precision as a metric should not predict human performance in phase 2 as well as in phase 1. Phase 2 offered considerably more freedom for behavior outside of what traditional Cranfield-style metrics attempt to predict.

From this experiment, we found that:

- For both interfaces, human performance was significantly greater with the higher level of retrieval precision.
- The difference in human performance between the two levels of retrieval precision was greater in the phase 1 interface that more closely matched Cranfield-style evaluation assumptions.
- Users change their behavior depending on the precision of the results list:
 - When judging high precision lists, users were less likely to judge NIST relevant documents as relevant compared to when judging the lower precision lists. This effect was stronger for phase 2 than for phase 1.
 - In phase 2, when searching low precision lists, users take more care to avoid clicking on summaries of NIST non-relevant documents as compared to when searching the higher precision lists.

We next describe the experiment details.

2. METHODS AND MATERIALS

We created a two phase, within-subjects study with each phase utilizing a different user interface. The first phase tightly controlled the behavior of the study participants while the second phase offered more freedom. Participants in both phases were the same. For each phase, the participants performed four tasks. Two of the tasks were with *good* retrieval results and two were with *bad* retrieval results. We next describe the experiment in more detail.

2.1 Collection and Search Topics

We selected 8 topics from the 2005 TREC Robust track. Table 1 shows the topics and their titles. Our criteria for the topics were that they be something that was not too dated

Number	Topic Title	Relevant
310	Radio Waves and Brain Cancer	65
336	Black Bear Attacks	42
362	Human Smuggling	175
367	Piracy	95
383	Mental Illness Drugs	137
426	Law Enforcement, Dogs	177
427	UV Damage, Eyes	58
436	Railway Accidents	356

Table 1: Topics used in the study and the number of NIST relevant documents for each topic.

and that would be of possible interest to the the study participants. For each topic, we took the topic’s description and narrative and rewrote them as a single description of what the participants should consider to be a relevant document.

The 2005 TREC Robust track used the AQUAINT document collection. This collection contains 1,033,461 newswire documents from the New York Times, Associated Press, and Xinhua News Agency.

For the 2005 TREC Robust track, NIST assessors judged documents as non-relevant, relevant, and highly relevant. To simplify analysis, we treat relevant and highly relevant documents as simply relevant documents.

2.2 Construction of Results Lists

The primary goal of this study is to measure how human performance changes with different levels of precision. To achieve this goal, we had to carefully control the precision of the results.

In both phases of the study, participants view an ordered list of documents. To understand how different levels of precision affect human behavior, it was important that the precision of this list be uniform. In other words, we decided to create an artificial, ranked list that controlled precision so that precision at rank N would remain approximately constant. This is in contrast to operational retrieval systems where it is well-known that, on average, precision decreases with increasing rank. Understanding how humans respond to ranked lists with decreasing levels of precision is a more difficult task that we leave for future work.

Even though our ranked lists are artificial, we carefully constructed the lists with two additional goals in mind. First, the ranking needed to seem plausible to the participants, i.e. the ranking should look like one produced by a real retrieval system. Second, choosing between relevant and non-relevant documents should remain approximately as difficult as it is for users of real systems. We believe that we achieved both of these goals as we next describe.

2.2.1 Uniform Precision

We used the following procedure for each topic to create ranked results of near uniform precision at rank N , where precision is one of the possible precisions at 10 (P10) greater than zero, i.e. 0.1, 0.2, ..., 1.0.

We first created a ranking of documents by performing a reciprocal rank fusion [6] on all of the runs submitted to the 2005 TREC Robust track minus the 4 lowest performing runs. From this ranking, we separated the known relevant documents from the non-relevant documents while maintaining the order of each set of documents. We used the NIST

provided relevance judgments (qrels). We treated unjudged documents as non-relevant. Unjudged documents made up a small fraction of the seen documents in both phases (2% in phase 1 and 6% in phase 2).

With a ranked list of relevant documents, and a ranked list of non-relevant documents in hand, we then recombined them as follows to achieve near-uniform precision. For a given P10 value, from the top of the ranked lists we remove the number relevant and non-relevant documents needed to obtain the desired P10. For example, if P10 is 0.3, we remove the top 3 ranked relevant documents and the top 7 non-relevant documents. We then randomly permute these 10 documents and append them to a final ranked list. We repeat this process 10 documents at time until we have a ranked list of 1000 documents.

2.3 Good and Bad Levels of Precision

Given our limited resources, we decided to only investigate two levels of precision, and thus we needed to create, for each search topic, 2 lists of ranked documents. We refer to the lists with higher precision as having a quality of *good* and the other as having a quality of *bad*. While our *good* and *bad* labels refer to specific precision levels, Al-Maskari et al.’s good and bad labels referred to composite retrieval systems [1].

In choosing the two levels of precision, we wanted to have the levels be different enough that if precision does affect human performance, we could be expected to see the difference given a reasonable number of participants. At the same time, the levels of precision had to be within of the range of plausibility for real retrieval systems. We believe we achieved these goals as follows.

To determine what the *good* and *bad* precision levels should be, we took the automatic, title-only runs from the 2005 TREC Robust track and measured their precision at 10 (P10). Over all 49 topics, the best system had an average P10 of 0.592 and the bottom of the “okay” runs was 0.304. Three runs were considered non-okay with P10’s of 0.120 and lower; these results are typically buggy or the result of operational mistakes. We decided that a P10 of 0.6 was to be the *good* quality and that 0.3 was to be the *bad* quality.

Because our constructed ranked lists have a near uniform precision at N , for mean average precision (MAP) and R-precision (precision at the number of known relevant documents, R), our *good* results are approximately 0.6 for both of these metrics (0.3 for *bad* results). P10, P20, and so forth are all exactly 0.6 and 0.3 up till the number of relevant documents is exhausted.

While a precision of 0.6 is a 100% relative improvement over 0.3, a uniform precision of 0.3 produces a mean average precision of approximately 0.3. The best performing automatic title-only run in the 2005 TREC Robust track had a MAP of 0.332. A uniform precision of 0.3 is by no means a ranked list of such low quality that it would prevent participants from finding many relevant documents.

These precision levels are similar to, but also different from previous investigations. For example, our *bad* level of precision at 0.3 is similar to the systems Hersh et al. investigated [7] while our *good* level of precision at 0.6 is more similar to the work of Turpin and Scholer [12]. Hersh et al.’s two systems had MAPs of 0.2753 and 0.3239 on the 6 topics from the TREC 8 interactive track used in their study [7]. Turpin and Scholer [12] used controlled result

lists with APs of 0.55 to 0.95. In contrast to Hersh et al. and Turpin and Scholer, our precision levels are such that one has a majority of non-relevant documents in the ranked list and the other has a majority of relevant documents.

2.4 Performance Measures

The aspect of human performance that we focus on in this paper is the number of relevant documents that a human can find using a retrieval system in a given amount of time. This is similar to but not the same as instance recall, which was used by Hersh et al. [7]. Instance recall requires that the *instances* found be novel. The example that Hersh et al. give is a search for Hubble telescope discoveries. If two documents report the same discovery, both count towards number of relevant documents found but only the instance of the specific discovery counts towards instance recall.

We are not measuring document recall. As Moffat and Zobel [9] have pointed out, there are numerous issues with attempts to measure recall since it requires the correct estimation of the number of relevant documents in the collection. We are in effect measuring cumulated gain [8] at a time t rather than a rank N .

To measure the number of relevant documents found, we need to determine when we consider a document relevant or not. Cleverdon et al. list four choices [5], of which the first three appear relevant to our study. Relevance can be determined:

1. “By the questioner.”
2. “By the consensus of opinion of a group of people.”
3. “By an individual, not the questioner.”

In this paper we measure relevance both by items 1 and 3. In future work, we plan to look and see if item 2 can be used for our study by using the judgments of our participants to create consensus relevance judgments.

Item 1 corresponds to equating the NIST relevance judgments (qrels) to be the relevance judgments of the questioner. Measuring performance by the qrels means that documents judged relevant by a participant only count toward the number found if the NIST assessor said that the documents are relevant. Item 3 corresponds to accepting on a per participant basis the participant’s judgments as correct.

With our experimental design, it is important to report measures based on item 3 as well as item 1. We report using item 1 since it has been common practice, but if a participant decides to behave in a random fashion or decides to judge everything relevant, the participant will show gains when documents are counted based only on NIST relevance. The same random participant will not show gains when the participant’s judgments are accepted as correct (item 3).

2.5 Study Design

As mentioned above, we conducted a two phase, within-subjects study. Each phase of the study utilized a different user interface, but the underlying result lists (Section 2.2) were the same for both phases.

Phase 1 started with participants filling out consent forms and answering a basic demographic and experience questionnaire. For both phases, participants then completed a 10-15 minute tutorial describing the phase’s tasks and giving participants practice with the user interfaces.

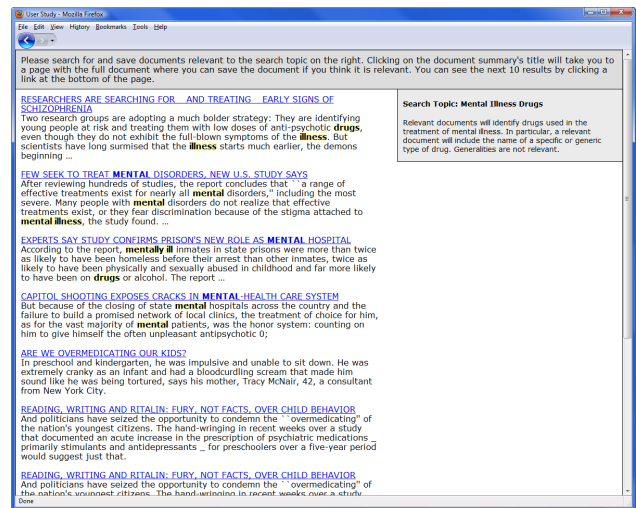
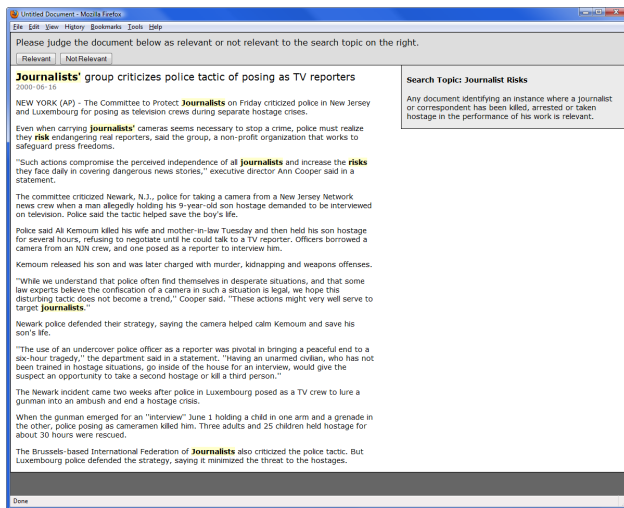


Figure 1: The left screenshot shows the user interface (UI) for phase 1 with a full document. Participants in phase 1 also judged document summaries in the same way. The right screenshot shows the phase 2 UI with query-biased document summaries shown. Clicking on a summary took the user to page with the full document. This page, similar to the phase 1 document judging UI on the left, allowed the user to save the document as relevant, but did not require a relevance judgment be made.

In each phase, a participant completed 4 search tasks. Each search task lasted 10 minutes. Two of the four search tasks had *good* result lists and two had *bad* result lists. Each search task corresponded to 1 of the 8 TREC search topics.

Before and after completing a search task, participants answered a short questionnaire on the topic. We used the same 4 pre- and post-task questions as Bailey et al. [3]. On the post-task questionnaire, we also provided the participants with a free-form textbox to allow them to alert us to any issues they encountered during the task. One possible effect of the pre-task questionnaire is that it may have forced the participant to study the search topic and its description of what is and is not relevant before beginning the actual search task.

We offered no incentives to participants to encourage them to work fast or with high accuracy. From a participant perspective there was no end to each task except for an eventual notice that 10 minutes had past and the task was over.

We did adopt language similar to that used by Smith and Kantor [11] and instructed the participants in phase 1 to “try to find as many relevant documents as possible in the 10 minutes while still making as few mistakes in judging the documents’ relevance as possible.” In phase 2, we asked participants to “try to find and save as many relevant documents as possible in the 10 minutes while saving as few non-relevant documents as possible.”

In phase 1, we received several questions from participants about seeing duplicate documents and summaries. In phase 2, we explicitly instructed participants to judge duplicate documents the same and that the duplicate documents were there by design.

2.5.1 User Interfaces

In this section, we describe the user interfaces for phases 1 and 2 and the associated tasks for each phase. Figure 1 shows both interfaces.

The phase 1 interface asks users to make binary relevance judgments. The interface shows the user one document or one document summary at a time. To see the next document or summary, the user must make a relevance judgment. The interface alternates between summaries and full documents. The interface instructs users to judge summaries based on what they think the full document’s relevance to be. Both summaries and documents have the search topic’s title terms highlighted. The search topic’s title and description remained visible for the entire task.

For the document summary, we displayed the document’s title, if known, as well as a query-biased snippet. To construct the query-biased snippets for each document, we used the topic’s title as a bag-of-words query and then retrieved the top two scoring sentences from the document. Snippets had a maximum length of 50 words.

All participants in phase 1 worked down their given ranked list. For example, a participant would start with judging the document or summary for the document at rank 1 of the assigned results list. If given a full document to judge, the participant would next judge a document summary of the document at rank 2 of the result list. We hid the navigation bar (back button, etc.) for phase 1. For both phase 1 and 2 we disabled the ability to right click on the web page.

We modeled the phase 2 interface after the standard style of modern web search engines. We presented the user with 10 document summaries per page. Clicking on a summary allowed the user to view the full document. The user could then decide to save a document if the user considered the document to be relevant. The user could undo their save operation if needed. Both saving and unsaving were implemented with client side scripts and dynamic HTML, which did not require a page reload. To go back to the list of search results, the user could use the browser’s back button. At the bottom of the page of search results were links to take the user to the next 10 results or the previous 10 results. As

with the phase 1 UI, the search topic’s title and description remained visible on every page for the entire task.

Note that the phase 2 interface lacks a query search box. There was no way for users to reformulate the query and produce a new ranked list. Again, like phase 1, the idea was to minimize the flexibility in the interface to control the experiment and determine how precision affects human performance.

In phase 2, when viewing the list of search results, the participants viewed 10 query-biased document summaries per page. With the uniform precision result lists, each page has the same number of relevant documents – 6 for *good* and 3 for *bad* results.

2.5.2 Balanced Design, Averages, and Significance

We built Graeco-Latin squares to balance topics and precision quality across the four tasks of each phase. A participant worked on 4 of the 8 topics in phase 1 and the remaining 4 topics in phase 2. A balanced block including whether participants started with a summary or a document in phase 1 required 16 participants. For each block of 16, we randomly permuted the topics and assigned the first four to one phase and the last four to the other phase. Half of the participants used one half of the topics for phase 1 while the other participants used the other half of the topics for phase 1. We randomly permuted the rows and columns of the squares and then randomly assigned the treatments to the identifiers used for building the squares.

In both phases, we have balanced topics across users and tasks. The goal of the balanced design is to eliminate the effect of topic on the results. We have utilized a within-subjects experiment so that we can block the data by user. This design allows us to measure statistical significance with matched pairs of data at the user level to compare the *good* and *bad* precision levels.

For all measures, we produce an average for each participant’s *good* tasks and for their *bad* tasks. We take the mean of these averages across all participants to obtain overall averages for the *good* and *bad* levels of precision. In the few cases where we could not compute a measure for a participant, we excluded the participant from the average. In addition to the mean, we report the standard deviation of the mean (SDOM, also known as the standard error, or standard error of the mean). For example, an average of 7.4 with a SDOM of 0.6 is reported as 7.4 ± 0.6 .

We measure statistical significance with the paired Student’s t-test in all cases. Pairing is by participant. In the case of the measures for which a priori we predict that increased precision will result in increased human performance (Table 2), we report the one-sided t-test’s p-value as well as the two-sided p-value. In all other cases (Table 3) we report only the two-sided p-value.

2.6 Cleaning of Data

At the initial close of the experiment, we had filled three blocks of 16 participants and had started to fill a fourth block of 16. From this data, we eliminated cases where we had technical issues affecting data collection, or where it was apparent that the participant did not attempt to accurately complete the task, or where we had other concerns about the data collection that would affect the integrity of the experiment. We then recruited additional participants to replace the deleted slots. We successfully filled 3 blocks of

16. Our results are based on this fully balanced set of 48 participants.

2.7 Participants

After having our study approved by our university’s office of research ethics, we recruited participants via posters and an email announcement to a university-wide graduate student email list. While we did not keep records on how participants learned of the study, the posters appeared to attract a diversified set of participants.

We paid participants \$10 for each phase of the study. We did not require participants to do both phases. For the majority of participants, phase 1 was conducted several weeks prior to phase 2. Other participants had several hours or days between phase 1 and phase 2. Only a few participants did phase 1 and phase 2 back to back. Some participants found phase 1 too boring and tedious and dropped out of the study after completing phase 1. We replaced participants who did not return for phase 2 with new participants.

Our results are based on the 48 participants that remained after data cleaning (Section 2.6).

All participants had used web search engines before. The majority reported using search engines several times a day, enjoying using search engines, and felt that they were experts who did not have trouble finding information on the Internet. Three participants reported receiving training in searching or information retrieval (none of these participants were known to us as IR students or researchers).

Hersh et al. [7] note that there have been unpublished observations that performance on recall tasks is related to reading speed. As such, we also asked participants about their self-perceived reading speed. 21 of the participants felt they were fast readers, 23 were neutral, and only 4 felt they were not.

The participants consisted of 23 females and 25 males. All but 2 of the participants were students. Of the students, 12 were undergraduates and 34 were graduate students. For the students, 34 (27 grad, 7 ugrad) were science, technology, engineering, or mathematics students. The other 12 students identified themselves as “arts” or “other”. The median age was 24.5, minimum was 18, and maximum was 56.

All participants considered themselves fluent speakers of English, but for many participants we observed that they did not seem to be native speakers of English, but we did not collect survey data on native language.

3. RESULTS AND DISCUSSION

Table 2 shows the main results of the study. For both phase 1 and 2, the *good* level of precision produced statistically significant improvements in human performance over the *bad* level of precision. We set *good* precision to 0.6 and *bad* precision to 0.3 (Section 2.2). Figure 2 graphically shows the results on a per-participant and per-topic basis. Table 3 shows additional measures of interest.

As explained in Section 2.4, we use two ways to count the number of relevant documents found by users in 10 minutes. The first way takes a user’s judgment as truth. When a user judges a document as relevant, we count that as finding a relevant document regardless of the NIST assessor’s judgment. The second way only counts a relevant document as found by the user when the user judges the document to be relevant and the NIST assessor agrees that the document is relevant.

Phase	Average Measure	Precision Bad	Quality Good	Diff.	Rel. Gain	1-sided p-value	2-sided p-value
1	Summaries Judged Relevant by User	3.6 ± 0.4	5.6 ± 0.7	2.1	59%	0.004	0.008
1	Summaries Judged Relevant by User and NIST	2.3 ± 0.3	4.6 ± 0.6	2.3	101%	< 0.001	0.001
1	Documents Judged Relevant by User	6.0 ± 0.6	8.8 ± 1.0	2.8	46%	< 0.001	0.001
1	Documents Judged Relevant by User and NIST	3.9 ± 0.4	7.4 ± 0.8	3.4	86%	< 0.001	< 0.001
2	Documents Saved as Relevant by User	10.8 ± 0.9	12.6 ± 1.1	1.9	17%	0.042	0.083
2	Documents Saved as Relevant by User and NIST	7.4 ± 0.6	10.9 ± 1.0	3.6	48%	< 0.001	0.001

Table 2: Main Results. Section 2.5.2 explains the calculation of averages and statistical significance.

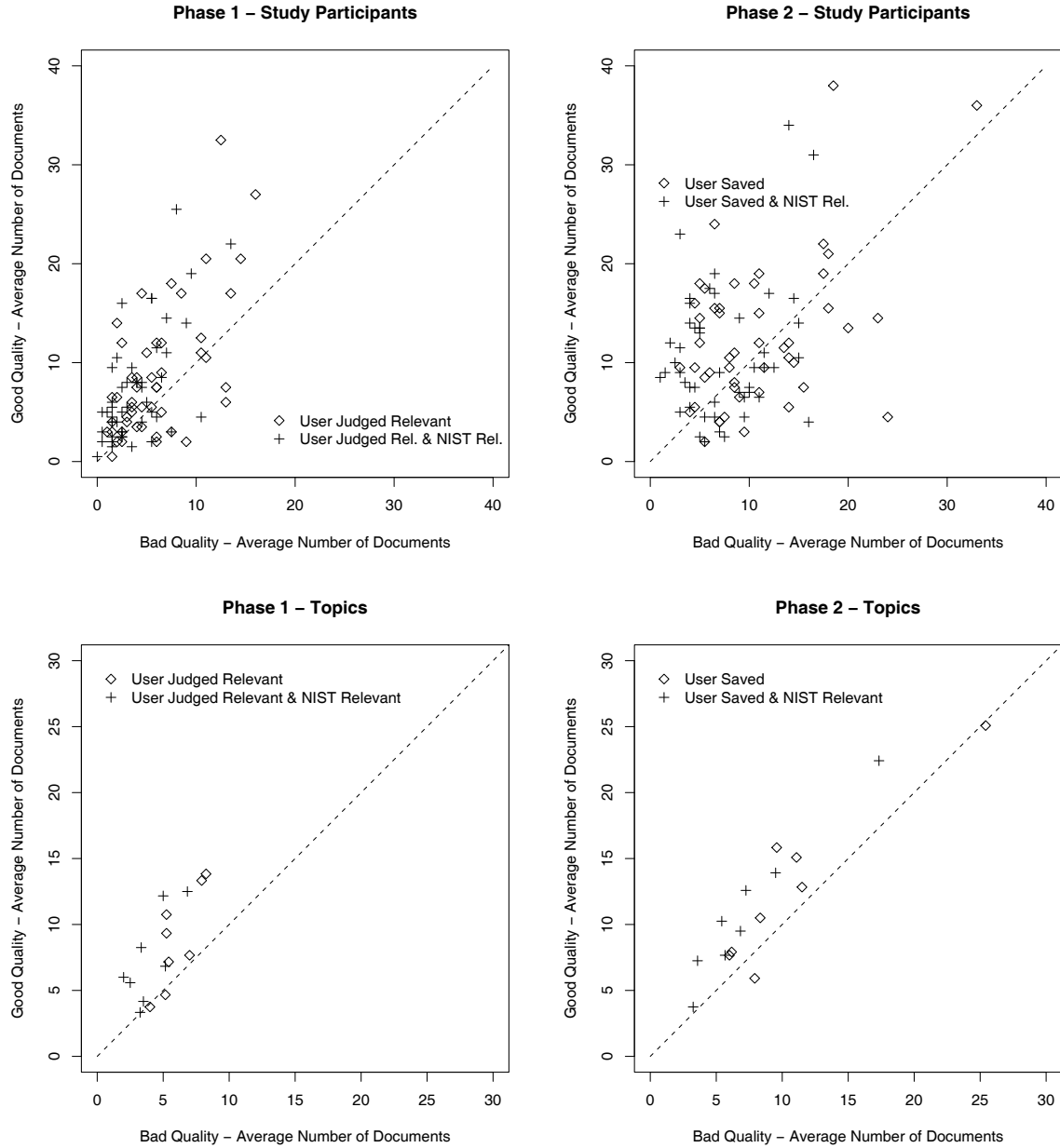


Figure 2: Average number of documents judged relevant (phase 1) and saved as relevant (phase 2) for *good* vs. *bad* precision levels. The top two plots show per-user averages. For these plots, points above the dashed line are study participants (users) who had better performance with the *good* level of precision. The bottom two plots show per-topic averages. For these two plots, each point is a topic and is the average of 12 study participants (users) for *bad* and 12 other participants for *good*.

Phase	Average Measure	Precision Quality		Rel. Change	t-test p-value
		Bad	Good		
1	Summaries Judged by User	16.6 ± 2	17.6 ± 2	6%	0.463
1	Documents Judged by User	16.4 ± 2	17.4 ± 2	6%	0.461
1	Time in Seconds to Judge Summary	15.6 ± 1	15.4 ± 1	-1%	0.906
1	Time in Seconds to Judge Document	49 ± 6	49 ± 7	-1%	0.950
1	User Summary Accuracy (Agreement with NIST)	0.72 ± 0.02	0.62 ± 0.02	-14%	0.007
1	User Document Accuracy (Agreement with NIST)	0.76 ± 0.03	0.75 ± 0.02	-1%	0.913
1	P(user judges rel NIST non-rel, summary)	0.17 ± 0.03	0.18 ± 0.03	7%	0.779
1	P(user judges rel NIST rel, summary)	0.47 ± 0.05	0.47 ± 0.03	0%	0.997
1	P(user judges rel NIST non-rel, document)	0.24 ± 0.03	0.23 ± 0.03	-3%	0.899
1	P(user judges rel NIST rel, document)	0.78 ± 0.03	0.73 ± 0.03	-6%	0.295
2	P(user clicks sum NIST non-rel, first 10 results)	0.58 ± 0.04	0.66 ± 0.04	14%	0.019
2	P(user clicks sum NIST rel, first 10 results)	0.84 ± 0.03	0.80 ± 0.03	-4%	0.307
2	P(user saves doc NIST non-rel, first 10 results)	0.20 ± 0.03	0.23 ± 0.03	11%	0.520
2	P(user saves doc NIST rel, first 10 results)	0.70 ± 0.04	0.63 ± 0.04	-10%	0.117
2	Fraction of Documents Viewed that are NIST Non-Relevant	0.56 ± 0.02	0.32 ± 0.01	-43%	< 0.001
2	Fraction of Documents Viewed that are NIST Relevant	0.44 ± 0.02	0.68 ± 0.01	54%	< 0.001
2	Fraction of Viewed NIST Non-Rel. Documents that are Saved	0.29 ± 0.03	0.26 ± 0.03	-11%	0.390
2	Fraction of Viewed NIST Relevant Documents that are Saved	0.77 ± 0.03	0.70 ± 0.03	-9%	0.080
2	Time in Seconds Viewing Summaries before Action	10.1 ± 1	8.0 ± 0.8	-21%	0.007
2	Time in Seconds Viewing Document before Action	27 ± 2	26 ± 2	-4%	0.730
2	Maximum Rank Viewed	50 ± 6	35 ± 3	-29%	0.009
2	Number of Documents Viewed	20.5 ± 1	21.5 ± 2	5%	0.412

Table 3: Additional Measures of Interest. The t-test is paired by user and is two-sided.

In phase 1, when user judgments are counted as truth, the relative gain of *good* over *bad* is 59% for summaries and 46% for documents. When we only count documents judged relevant by both the user and NIST, the relative gain is 101% for summaries and 86% for documents. More documents are judged as relevant than summaries, for users are less likely to judge summaries as relevant. It seems that if in doubt, users make a decision of non-relevant.

In phase 2, the relative gains are less than in phase 1; the gap in human performance between the *good* and *bad* lists has shrunk. The relative gain for documents saved is 17% and it is 48% for documents saved that are NIST relevant. The one-sided t-test is appropriate for all the results in Table 2, and for these results, all are statistically significant ($p < 0.05$). Figure 2 shows that for the majority of topics, users performed better with the higher precision results. On a per-topic basis, we obtain statistically significant results for a one-sided, paired t-test with only 8 topics ($p < 0.05$).

While we measured statistically significant increases in human performance, the relative gain from *bad* to *good* varies considerably. An obvious question to ask is why are so many of the human performance improvements so much less than the 100% difference in precision?

There are several possibilities that need to be investigated to understand these differences. One possibility is that users may take more time to judge a document relevant than they do to judge a document non-relevant. With a low precision list of documents, the user will encounter many more non-relevant documents than when the user views a higher precision list. The difference in judging times would allow for more documents to be judged in the low precision list than in the high precision list in a given amount of time.

We found no support for users taking different amounts of time for the judging of summaries or documents in phase 1.

Nor did we find any evidence that users were able to judge more documents when judging the lower precision list in phase 1. For phase 2, users did spend less time viewing summaries when the list quality was *good* – 8.0 seconds vs. 10.1 seconds. This difference in time did not result in a statistically significant difference in the number of documents viewed (20.5 vs. 21.5 documents).

Another possibility is that users may change how they judge relevance if they detect that the results are good or bad on average. Scholer and Turpin have shown users can vary significantly in their relevance criteria [10].

In effect, the user could learn a prior probability of relevance for the results list either consciously or subconsciously. This added knowledge, or something else related to the precision of the list, could lead the user to change what they consider to be relevant to the search topic. The user might change their relevance criteria by either tightening or loosening it. Changing the relevance criteria does not change the rate of judging but it does change the rate at which documents are judged to be relevant by the user.

We do see some evidence that the users change their relevance criteria given the precision of the results list. Table 3 shows the measured probabilities that a user would judge a document as relevant. For documents judged by NIST to be non-relevant documents, we see no evidence of users changing their relevance criteria given the precision of the results list. In both phases, users save as relevant 20 to 24 percent of documents judged by the NIST assessor to be non-relevant.

On the other hand, for documents judged relevant by NIST, it appears that users change their relevance criteria. While there is only suggestive evidence for phase 1, for phase 2 the evidence is stronger. In phase 1, when precision is *bad*, the users have a probability of 0.78 of judging NIST

relevant documents as relevant, but when precision is *good*, the probability drops to 0.73 ($p = 0.295$). For phase 2, we have two measures of how users judge NIST relevant documents. The first measure estimates this probability based on the user behavior for the first page of 10 results under the assumption that the users view all 10 of these documents. In this case, for *bad* the probability is 0.70, and for *good* the probability is 0.63 ($p = 0.117$). The second measure is the fraction of NIST relevant documents viewed, i.e. the user clicks on a summary and views the full document, that are saved as relevant. For *bad* this fraction is 0.77, and for *good* it is 0.70 ($p = 0.080$). Taken together, it appears that when users are searching a low precision list, they are more likely to save a NIST relevant document as relevant than when they are viewing a high precision list.

How does this affect the measured human performance? The reduction in predicted performance for documents saved as relevant comes about from users saving NIST non-relevant documents as relevant and also from the difference in the probability of saving relevant documents. When users view NIST non-relevant documents in both the *bad* and *good* results, they will save them as relevant about 20 to 24 percent of the time, but when viewing a low precision list, the user sees many more NIST non-relevant documents. For relevant documents, the story is different. The user viewing the low precision list is more likely to save the NIST relevant documents as relevant than when the user views the high precision list.

In phase 2, there is another way that users change their behavior given the precision of the results. Users are less likely to click on summaries of documents judged by NIST to be non-relevant. For *bad*, users click on summaries of non-relevant documents with a probability of 0.58 while for *good*, the probability is 0.66 ($p = 0.019$). This goes in hand with users taking longer to view summaries when the list precision is *bad*. The end result is that users go considerably deeper into the *bad* ranked list than they do in the *good* list. For *bad*, users on average reach rank 50, while for *good*, they reach rank 35 ($p = 0.009$). With what appears to be a small overhead of about 2 seconds per view of the query-biased summaries, users of the lower precision results can avoid wasting time viewing full documents not likely to be relevant.

4. CONCLUSION

We ran a two phase, within-subjects user study in which we controlled both the retrieval precision and the amount of interaction allowed in the search interface. By doing so, we were able to observe that the users change their behavior given the precision of the results list. When viewing lower precision results, users are more careful to avoid clicking on non-relevant summaries and are more likely to judge relevant documents as relevant. When viewing higher precision results, the same users spend less time viewing summaries, are more likely to click on non-relevant summaries, and are less likely to judge relevant documents as relevant. These behaviors, combined with the lack of perfect inter-assessor agreement, reduce the measured gains expected from increases in precision.

For both phases of the study, we found that retrieval precision and human performance are strongly related. The higher precision result lists produced statistically significant gains in human performance. We found that when the user

task and user interface better match the Cranfield-style evaluation metric, the metric better predicts human performance.

5. ACKNOWLEDGMENTS

Special thanks to Gordon Cormack for his helpful advice, and thanks to the anonymous reviewers for their helpful feedback. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), in part by an Amazon Web Services in Education Research Grant, and in part by the University of Waterloo. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

6. REFERENCES

- [1] A. Al-Maskari, M. Sanderson, P. Clough, and E. Airio. The good and the bad system: does the test collection predict users' effectiveness? In *SIGIR'08*, pages 59–66. ACM, 2008.
- [2] J. Allan, B. Carterette, and J. Lewis. When will information retrieval be “good enough”? In *SIGIR'05*, pages 433–440. ACM, 2005.
- [3] E. W. Bailey, D. Kelly, and K. Gyllstrom. Undergraduates' evaluations of assigned search topics. In *SIGIR'09*, pages 812–813. ACM, 2009.
- [4] C. Cleverdon. The Cranfield tests on index language devices. In *Aslib Proceedings*, volume 19, pages 172–192, 1967.
- [5] C. Cleverdon, J. Mills, and M. Keen. Aslib Cranfield research project - factors determining the performance of indexing systems; volume 1, design; part 1, text. Technical report, Cranfield University, 1966. URI: <http://hdl.handle.net/1826/861>.
- [6] G. V. Cormack, C. L. A. Clarke, and S. Buettcher. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *SIGIR'09*, pages 758–759. ACM, 2009.
- [7] W. Hersh, A. Turpin, S. Price, B. Chan, D. Kramer, L. Sacherek, and D. Olson. Do batch and user evaluations give the same results? In *SIGIR'00*, pages 17–24. ACM, 2000.
- [8] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *TOIS*, 20(4):422–446, 2002.
- [9] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *TOIS*, 27(1):1–27, 2008.
- [10] F. Scholer and A. Turpin. Relevance thresholds in system evaluations. In *SIGIR'08*, pages 693–694. ACM, 2008.
- [11] C. L. Smith and P. B. Kantor. User adaptation: good results from poor systems. In *SIGIR'08*, pages 147–154. ACM, 2008.
- [12] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *SIGIR'06*, pages 11–18. ACM, 2006.
- [13] A. H. Turpin and W. Hersh. Why batch and user evaluations do not give the same results. In *SIGIR'01*, pages 225–231. ACM, 2001.
- [14] E. M. Voorhees. I come not to bury Cranfield, but to praise it. In *HCIR'09*, pages 13–16, 2009.