Method Bias? The Effects of Performance Feedback on Users' Evaluations of an Interactive IR System

Diane Kelly, Chirag Shah, Cassidy R. Sugimoto, Earl W. Bailey, Rachel A. Clemens, Ann K. Irvine, Nicholas A. Johnson, Weimao Ke, Sanghee Oh, Anezka Poljakova, Marcos A. Rodriguez, Megan G. van Noord, & Yan Zhang

University of North Carolina
100 Manning Hall, CB#3360
Chapel Hill, NC 27599-3360 USA
+1 919.962.8065

dianek@email.unc.edu

ABSTRACT

In this study, we seek to understand how providing feedback to users about their performances with an interactive information retrieval (IIR) system impacts their evaluations of that system. Sixty subjects completed three recall-based searching tasks with an experimental IIR system and were asked to evaluate the system after each task and after finishing all three tasks. Before completing the final evaluation, three-fourths of the subjects were provided with feedback about their performances. Subjects were assigned randomly to one of four feedback conditions: a baseline condition where no feedback was provided; an actual feedback condition where subjects were provided with their real performances; and two conditions where subjects were deceived and told that they performed very well or very poorly. Results show that the type of feedback provided significantly affected subjects' system evaluations; most importantly there was a significant difference in subjects' satisfaction ratings before and after feedback was provided in the actual feedback condition. These results suggest that researchers should provide users with feedback about their performances when this information is available in order to elicit the most valid evaluation data.

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, Measurement

Keywords

Evaluation methods, performance feedback, user satisfaction, performance ratings, usability, rating bias

1. INTRODUCTION

Evaluation is a core part of interactive information retrieval (IIR) and IR more generally. Good evaluation is dependent on good evaluation methods and metrics. In IR, the evaluation method has changed little from the Cranfield [4] model, which was later refined and made popular by TREC [21]. One of the most well-established methods for evaluating experimental IIR systems was also developed as part of TREC within the Interactive Track [5]. This method prescribed a protocol for conducting IIR evaluations, along with specific instruments and measures. A number of studies have investigated some of the assumptions about, and limitations of, this evaluation method, especially with respect to how relevance is conceptualized and measured [c.f., 18] and the appropriateness of traditional TREC performance metrics [c.f., 12] and search tasks [3]. Borlund [2] summarizes this work and presents an alternative evaluation method for experimental IIR systems. Toms, et al [20] and Thomas & Hawking [19] present alternatives for Web IIR. In this study, we are concerned with evaluation methods modeled after the TREC Interactive Track and evaluation situations for which these methods are appropriate.

In the typical evaluation scenario for experimental interactive information retrieval (IIR) systems, users are asked to search for information about particular topics and then evaluate one or more systems using 5-point Likert-type scales. Users are not provided with any objective indicator of how well they performed and instead rely on their *perceptions* of how well they performed. However, in many evaluation situations, especially those involving standard TREC collections and fixed relevance judgments, information about performance is available, but it is not usually provided to users. Instead, users are asked to evaluate a system without actually knowing anything about the outcome of their interactions with the system. For experiments involving high-precision tasks that only require a single document for resolution, users' perceptions of performance are likely to be somewhat accurate. However, for studies involving recall-oriented tasks where users have no knowledge of how many relevant documents are actually available, it is unlikely that users can adequately judge system performance based solely on their perceptions. Instead, it would seem that providing actual performance information would help users make more informed and accurate system evaluations.

Evidence exists that shows that users' evaluations of systems are generally favorable, even when they do not perform well according to objective performance measures, and that objective performance metrics do not always correlate positively with users' satisfaction ratings [7, 10, 14, 16]. In a recent study, Hornbæk and Law [10] conducted a meta-analysis of data sets from 73 published human-computer interaction (HCI) studies and found that the correlation between effectiveness and satisfaction was a paltry .164. In these studies the most common measure of effectiveness was error rate instead of recall or precision, but it is

peculiar that users would be satisfied with systems that they could not use effectively. In an earlier study, Nielsen and Levy [16] conducted a meta-analysis of 57 HCI studies and found that users' ratings of the performances of 101 of the 127 systems described in these studies were higher than the neutral points on the scales used for evaluation. One possible (but unlikely) explanation for this finding is that all of the experimental systems were above-average. An alternative explanation, which we examine in this paper, is that users' evaluations of systems are somehow biased by the method, which may lead them to make more positive ratings.

The studies investigated by Hornbæk and Law [10] and Nielsen and Levy [16] were primarily from HCI publications. While it is likely that the publications included some studies of IIR systems, it is probably the case that these types of studies were in the minority. In the area of IIR, recent studies have provided evidence that seems to differ slightly from the results of these HCI studies [1, 11]. These studies have found that user satisfaction is positively related to at least some objective system performance measures. However, it is difficult to compare the findings of at least one of these studies [11] to the previous studies since no frequency or descriptive information was given about users' satisfaction ratings, nor were readers told what type of instrument was used to elicit these ratings. It may be the case that users' ratings were still relatively positive and above the scale mid-point regardless of performance. These studies were also in the context of Web searching and used high-precision tasks where users' perceptions of how well they performed were likely accurate.

In this paper, we investigate the following research question, "How does providing feedback to users about their search performance for high-recall tasks affect their system evaluations?" We were interested in investigating the impact of four types of feedback which we present below. These feedback conditions are discussed in more detail in the Method section, but are presented here so that we can present our hypothesis. This feedback was provided to users before they completed their final system evaluations. The feedback conditions were:

1. No Feedback: users were not provided with any feedback about their performances. This represents what typically occurs in IIR studies and functions as a baseline.
2. Actual Feedback: users were told their real performance.
3. Good Feedback: users were deceived and told that they performed very well.
4. Poor Feedback: users were deceived and told that they performed very poorly.

We hypothesized that users in the poor feedback (PF) condition would rate the system the lowest, followed by users in the actual feedback (AF), no feedback (NF) and good feedback (GF) conditions [PF < AF < NF < GF].

The potential difference between the no feedback and actual feedback conditions is of greatest interest since this difference will provide evidence that the evaluations users give when they know their performance differs from those they give when they do not know, which is what generally happens in IIR study. Such evidence would suggest that the evaluation method used in traditional IIR experiments can generate biased user ratings; user ratings which may be more positive than they would be otherwise.

2. METHOD

We conducted a controlled, laboratory study with a single experimental IIR system. This study was specifically designed to investigate the research question posed above, so only a single system was used. This system will not be described in detail, but it has been used in two other evaluations and interested readers can see [8, 13] for more information. The system was a basic search system that was built on top of Lemur[1] and used BM25 for retrieval. The interface provided a standard IIR interaction: users could query, review result lists and the full text of documents, and save documents. The experimental part of the system was that users could organize saved documents in different ways.

2.1 Subjects

Subjects were recruited by sending email solicitations to undergraduate students at our university. Subjects participated in private, 1 hour sessions and were compensated with $10.00 USD. Sixty subjects participated in this experiment (15 per condition). Thirty-eight subjects were female and 22 were male. Subjects' mean age was 20 years (SD=1.14). Seventeen percent of the subjects were humanities majors, 23% were social science majors, 33% were science majors, and 27% were in a professional school. Ninety-five percent of the subjects said they were fairly or very experienced searchers and 92% said they search the Web daily.

2.2 Feedback Conditions

This study had one independent variable – feedback condition – which had four levels: no feedback (baseline) (NF), actual feedback (AF), good feedback (GF) and poor feedback (PF). Subjects were randomly assigned to experimental condition. For the three conditions where subjects received feedback, we presented this feedback to them in the form of *percent of relevant documents saved*. For instance, if a subject found and saved 10 of the 30 relevant documents for a particular topic, then his feedback score would be 33%. We wanted the system to give subjects feedback about their performances instead of the experimenter, so selected a measure that we felt could be easily communicated to subjects via computer screen and that did not require explanation.

To determine values for the poor and good feedback conditions, we first examined results from the two previous evaluations of the experimental system which involved a combined 72 users [8, 13]. In these two studies, most users found about 20-30% of the relevant documents. This suggested the kind of feedback subjects in the actual feedback condition would likely receive and helped us identify appropriate values for the poor and good feedback conditions. For the poor feedback condition, we selected a value of 12%, which was about half of the average expected performance. For the good feedback condition, we selected a value of 92%, which was more than three times greater than the average expected performance.

2.3 Topics and Documents

Since we used the performance of subjects from two previous studies as a benchmark for identifying values for our feedback conditions, we used the same search tasks and corpus from these previous studies to increase the chances that these values would be observed again. The search tasks, corpus and relevance judgments were from the TREC-8 Interactive Track collection

[1] http://www.lemurproject.org/

[9]. This collection consists of a corpus of 210,158 articles from the Financial Times of London 1991-1994 and a set of topics that focused on aspectual recall. The original aspectual recall task required users to find documents that discuss different aspects or instances of a topic, instead of finding all relevant documents about a particular topic. For example, a user might be interested in finding out the different treatments for high blood pressure, rather than finding all documents that discuss all such treatments.

Modified versions of four aspectual recall topics from this collection were used in the previous studies to emphasize both comprehensiveness (find documents covering as many aspects as possible) and exhaustiveness (find as many documents as possible relating to each aspect). These topics were also modified in this previous study to include work task scenarios [2]. The four topics we used in this study were: 408i (tropical storms), 428i (declining birth rates), 431i (robotic technology) and 446i (tourist, violence) (used for tutorial), all of which had been used successfully in the previous studies. The modified version of the tropical storms topic is displayed in Figure 1.

Topic: 408 **Title**: Tropical Storms

Description: Imagine that you are enrolled in an environmental science course and you are interested in learning more about tropical storms (hurricanes and typhoons). It seems that tropical storms are becoming more destructive and you decide to investigate past storms. Your instructor asks you to prepare a short, 5-page paper investigating past tropical storms. Your instructor asks you to use historical newspaper articles as your sources of information and to collect comprehensive information on different tropical storms and impacts of each storm. Specifically, your goal is to identify different tropical storms that have caused property damage and/or loss of life, and to find as much information as possible about each storm.

Figure 1. Modified version of TREC-8 Interactive Track Topic used in this study (from [20]).

2.4 System Evaluations

In the typical IIR experiment, researchers elicit system evaluations from users via questionnaires. We used two questionnaires to elicit subjects' evaluations, the Post-Task and Exit Questionnaires. These questionnaires contained items that are commonly used to evaluate IIR systems. Subjects completed three Post-Task Questionnaires, one after each task. Subjects completed the Exit Questionnaire after they finished all searching or after they received feedback about their performance.

The Post-Task Questionnaire contained a series of 6 items. Subjects responded using a 7-point Likert-type scale, where the labels were 1=not at all, 4=somewhat, and 7=extremely. The text of these items can be seen in Figure 2 in the Results section. The Post-Task items allowed us to gauge subjects' impressions of the system before they were provided with feedback. The last item, in particular, functioned as a pre-feedback measure of satisfaction which we could directly compare to a near identical satisfaction item on the Exit Questionnaire. The remaining items did not have counterparts on the Exit Questionnaire because we did not want to make subjects suspicious with seemingly repetitive items.

The Exit Questionnaire contained a series of 13 items and subjects responded with a 7-point Likert-type scale, where the labels were 1=strongly disagree and 7=strongly agree. We changed the scale type so that subjects would not feel as if they were answering the same items. The text of these items can be seen in Figure 3 in the Results section. Item 13 provided a post-feedback measure of satisfaction that was a counterpart to the last satisfaction item on the Post-Task Questionnaire.

2.5 Procedure

Subjects were instructed that they would be helping the researchers evaluate an experimental information retrieval system. When they arrived to the lab, they completed a consent form and a short demographic questionnaire to collect background data. Next, subjects watched a video tutorial of the experimental system. Following this, they were presented with the first search task. Subjects were given up to 15 minutes to complete each search task. In total, subjects completed 3 search tasks, which were rotated using a Latin-Square. After each search task, subjects completed the Post-Task Questionnaire which asked them to evaluate the system with respect to each task. After subjects finished the last Post-Task Questionnaire, but before they began the Exit Questionnaire, the system provided them with feedback about their performances (subjects in the no feedback condition went straight to the Exit Questionnaire). Feedback was accessed via hypertext link on the computer screen which read, "Compute my overall performance." Subjects were instructed to click this link. The system then displayed the appropriate feedback nested in the following sentence, "You found N% of the relevant documents," where N was equal to whatever feedback they were meant to receive. After completing the Exit Questionnaire, subjects were debriefed about the experiment and the deception (when appropriate).

3. RESULTS

We start with some benchmark data about how subjects actually performed during the study. The next section presents results of the Post-Task Questionnaire, which provides information about subjects' evaluation of the system before feedback was provided. Results from the Exit Questionnaire are presented in the next section. Finally, the pre- and post-feedback measures of satisfaction are compared directly.

3.1 Actual Performance

On average, subjects in this study performed similarly to those in the previous two studies with a mean recall of .23 (SD=.11) across all three tasks. The minimum performance score was .13 and the maximum was .52. No subject actually performed equal to or worse than our poor feedback condition (12%).

Subjects' performances according to condition and task are presented in Table 1. There were no statistically significant differences in subjects' performances according to condition, which was expected since subjects were randomly assigned to condition [$F(3,59)=.860$, $p=.467$]. There were significant differences in how subjects performed according to task, but these differences were consistent across condition [$F(2,79)=5.30$, $p=.006$]. These results show that subjects' performances were roughly equal in each condition.

Table 1. Mean (standard deviation) performance according to condition and task.

		Topic			Total
		408	428	431	
Condition	NF	.20 (.08)	.19 (.07)	.22 (.09)	.20 (.08)
	AF	.22 (.15)	.24 (.14)	.27 (.09)	.24 (.13)
	GF	.24 (.13)	.20 (.07)	.30 (.09)	.25 (.10)
	PF	.24 (.08)	.21 (.07)	.28 (.15)	.24 (.11)
Total		.23 (.11)	.21 (.09)	.27 (.11)	.23 (.11)

3.2 Pre-Feedback Evaluation Items

Results from the Post-Task Questionnaire provide a baseline indication of subjects' impressions of the system before any feedback was provided. Subjects' mean ratings for each item according to feedback condition are displayed in Table 2. ANOVA results are presented in the last two columns of the table. As expected, there were no statistically significant differences for any of these measures according to condition.

1. Before your search, how familiar were you with this topic?
2. How easy was it to search on this topic?
3. How satisfied are you with the search results?
4. How confident are you that you identified all the relevant documents for this topic?
5. Did you have enough time to do an effective search?
6. How satisfied are you with your performance?

Figure 2. Post-Task Items

The familiarity item (1) is of little interest to us in this study, but it is noted that subjects' familiarities with tasks were quite low. These values are similar to those observed in the two previous studies [8, 13]. As can be seen from subjects' responses to all other items, overall, things seem to be going okay for most subjects – the mean values for the other five items were all above the scale mid-point.

Table 2. Means, standard deviations and ANOVA [*df* = (3, 179] results for Post-Task items.

	Condition				Total	*F*	*p*
	NF	AF	GF	PF			
1	3.31 (1.56)	2.91 (1.43)	2.80 (1.46)	3.11 (1.54)	3.03 (1.49)	1.03	.38
2	4.87 (1.24)	4.84 (1.22)	4.91 (1.22)	4.96 (1.09)	4.89 (1.18)	.077	.97
3	4.78 (1.41)	4.51 (1.38)	4.51 (1.33)	4.56 (1.41)	4.59 (1.37)	.385	.76
4	4.31 (1.22)	4.31 (1.29)	4.42 (1.47)	4.24 (1.54)	4.32 (1.38)	.127	.94
5	5.36 (1.26)	5.53 (1.36)	5.42 (1.36)	5.33 (1.41)	5.41 (1.34)	.199	.90
6	4.89 (1.01)	5.22 (1.19)	4.82 (1.28)	4.69 (1.20)	4.91 (1.18)	1.68	.17

Pearson correlation coefficients were computed between subjects' objective performance (i.e., recall) and each Post-Task item. These results are shown in Table 3. There were two statistically significant correlations at the .05 level between recall and easy to do search and recall and satisfaction with search results. However, these correlation coefficients are quite small and not particularly meaningful. There was no correlation between recall and subjects' evaluations of performance (Items 4 and 5) or satisfaction with performance (Item 6).

Table 3. Correlation coefficients between performance and Post-Task items. *p<.05

	Post-Task Item					
	(1)	(2)	(3)	(4)	(5)	(6)
Recall	.005	.164*	.166*	-.055	-.093	.003

Overall, results in this section show that subjects' evaluations were, in general, positive before they received any performance feedback and that, for the most part, these evaluations did not correlate with their actual performance.

3.3 Post-Feedback Evaluation Items

Results from the Exit Questionnaire provide an indication of how performance feedback affected subjects' system evaluations. Subjects' mean ratings for each item according to feedback condition are displayed in Table 4. ANOVA results are presented in the last two columns of the table. Overall, feedback condition had a statistically significant impact on how subjects' responded to 5 of the 13 items. Scheffe's post-hoc tests were done to examine between which pairs of conditions significant differences existed. These results are displayed in Table 5. Scheffe's is a conservative test and as can be seen from the results, there were no significant pair-wise relationships found for two items, despite the significant ANOVA results.

Table 5 also displays eta^2 values indicating effect size, which shows how much variance in a particular item (e.g., satisfaction) is explained by differences the treatment (i.e., feedback condition). Overall, the effect sizes were large, with that for satisfaction being the greatest. Despite the weak Scheffe's results, the eta^2 values indicate that feedback condition explains quite a bit of the variability in each of these items.

1. The system was easy to learn to use.
2. I didn't notice any inconsistencies when I used the system.
3. It was easy to pose queries to the system.
4. It was easy to navigate the search results.
5. The search methods I used in this study were similar to those I use when I normally search the Web.
6. The color-coding of the piles made sense to me.
7. Overall, the system was easy to use.
8. The system made things I wanted to accomplish easy to do.
9. In general, it was easy to find relevant documents with the system.
10. It was easy to understand why documents were retrieved in response to my query.
11. The various functions of the system were well integrated.
12. Overall, the system was effective in helping me complete search tasks.
13. Overall, I am satisfied with my performance.

Figure 3. Exit Questionnaire

Table 4. Means, standard deviations and ANOVA [*df* = (3, 59)] results for Exit items. **p*<.05; ***p*<.01

	Condition				Total	*F*	*p*
	NF	AF	GF	PF			
1	5.87 (.92)	5.40 (1.18)	6.00 (1.00)	5.73 (1.10)	5.75 (1.05)	.89	.45
2	5.93 (1.03)	4.47 (1.64)	5.47 (1.30)	4.73 (1.43)	5.15 (1.46)	3.60	.02*
3	6.00 (1.07)	5.13 (1.13)	5.73 (1.62)	4.47 (1.91)	5.40 (1.48)	4.50	.01**
4	5.47 (1.25)	5.07 (1.34)	5.20 (1.90)	4.20 (1.74)	5.12 (1.57)	2.99	.04*
5	4.80 (1.90)	4.53 (1.64)	6.40 (1.18)	4.87 (1.96)	4.85 (1.82)	.33	.81
6	6.13 (1.06)	5.93 (1.10)	5.93 (1.28)	5.87 (1.45)	6.08 (1.20)	.59	.63
7	5.93 (1.10)	5.67 (.82)	5.93 (1.28)	5.07 (1.75)	5.65 (1.30)	1.52	.22
8	5.27 (1.39)	5.00 (1.25)	5.47 (1.55)	4.33 (1.63)	5.02 (1.49)	1.71	.18
9	4.93 (1.16)	4.27 (1.58)	4.73 (1.58)	4.13 (1.69)	4.52 (1.51)	.94	.43
10	5.07 (1.39)	3.87 (1.92)	4.60 (1.24)	3.60 (1.60)	4.28 (1.63)	2.80	.05*
11	5.53 (.83)	4.87 (1.36)	5.47 (.92)	4.93 (1.16)	5.20 (1.10)	1.54	.21
12	5.40 (1.06)	4.53 (1.77)	5.33 (1.18)	4.87 (1.73)	5.03 (1.47)	1.17	.33
13	5.27 (1.03)	4.40 (1.64)	5.93 (.88)	3.60 (1.40)	4.80 (1.53)	9.54	.00**

Table 5. Results of Scheffe's post-hoc tests and effect sizes (ns=not significant)

	Exit Item				
	2	3	4	10	13
Scheffe's Results	AF < NF	PF < NF PF < GF	ns	ns	PF < NF AF < GF PF < GF
Effect Size (Eta^2)	.16	.19	.14	.13	.34

Figure 4 illustrates the differences among mean scores for each Exit item according to condition. These differences are presented in relation to the NF condition (baseline) whose mean for each item is set to 0. For only about half of the items, subjects' ratings in the GF condition were the highest, which is somewhat surprising. For the other half of the items, subjects' ratings in this condition were equal to or less than subjects' ratings in the NF condition. Although subjects' ratings in the PF condition were always lower than those in the GF condition, they were not always lowest overall. Instead, for 5 of the items, subjects' ratings in the AF condition were the lowest, although the differences were only statistically reliable for 1 of the 5 items. Ratings by those in the NF condition were never the lowest and in fact were usually the second or first highest ratings and were always higher than those in the AF condition. Taken together, these results seem to suggest that subjects exhibit some type of response bias when evaluating systems and that informing subjects of their actual performance before asking them to make such evaluations can change the nature of these evaluations.

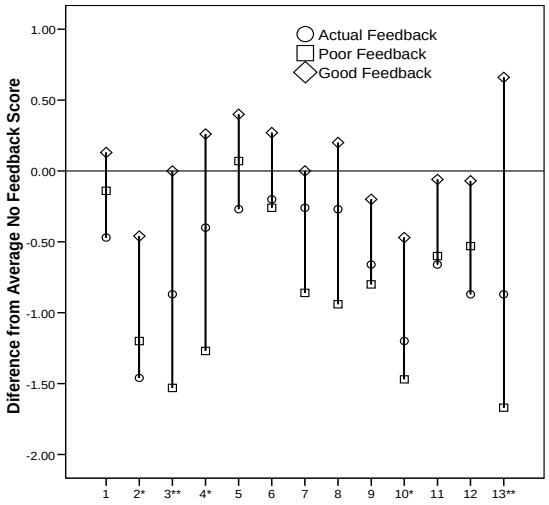

Figure 4. Differences among subjects' mean ratings for each Exit item according to condition. Subjects' ratings in the NF condition appear as the 0 reference line. **p*<.05; ***p*<.01

The strongest ANOVA result was associated with the final satisfaction item. The impact of condition on this item is best illustrated by Figure 5. Note the differences in both means and variability for scores in each condition. Subjects' ratings in the AF condition were most variable, while those in the GF condition were least variable. It is also clear from this Figure that subjects in the NF condition rated their satisfaction nearly one point higher than those in the AF condition and that their scores were closer to those of subjects in the GF condition.

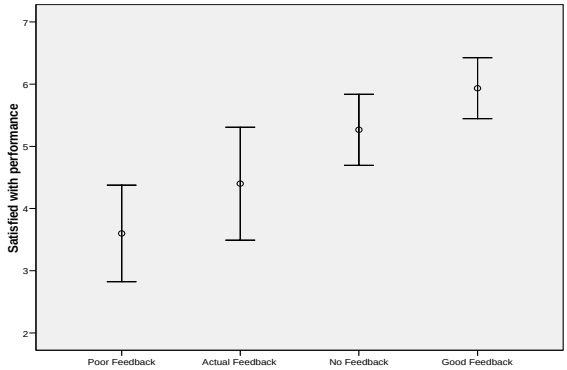

Figure 5. Mean (and variance) satisfaction ratings according to feedback condition.

One possible factor that may have impacted the differences in subject's ratings in the PF and GF conditions is their actual performance. For instance, if a subject in the PF condition actually performed well and had a 'good feeling' about the search was the effect of feedback on that person's ratings greater than for a person who didn't try hard to begin with and therefore was unsurprised by the poor feedback? To investigate this question, we grouped subjects into four performance classes which were defined by frequency quartiles. The definitions of these classes and distributions across conditions are displayed in Table 6.

Table 6. Distribution of performance groups across condition

	Performance Group				
	1 (.13-.17)	**2 (.18-.22)**	**3 (.23-.27)**	**4 (.28-.52)**	**Total**
NF	7	2	5	1	15
AF	4	4	2	5	15
GF	4	3	4	4	15
PF	2	5	4	4	15
Total	17	14	15	14	60

The distribution of performance groups across the NF condition was somewhat uneven, with nearly half of the subjects performing in the lowest group. Recall that subjects in this condition gave the system some of the most positive evaluations. To see if there was an interaction effect between actual performance and condition, we conducted a multivariate ANOVA using the items from the Exit Questionnaire. We note that this was an exploratory analysis; cell sizes were small for the interaction so these results are not definitive. Results confirmed the earlier findings with respect to main effects for condition (these statistics were reported in Table 4), but there was no significant main effect for performance group and there were no significant interaction effects (these 26 non-significant statistics will not be presented to conserve space). Thus, the exploratory analysis suggests that subjects' actual performances did not impact their system evaluations and that there was no interaction between actual performance and condition.

3.4 Pre- and Post-Feedback Satisfaction

The preceding analyses showed that performance feedback can impact subjects' system evaluations, at least for some items. The strongest result was related to satisfaction. In this section, we compare pre- and post-feedback satisfaction ratings. These results are displayed in Figure 6.

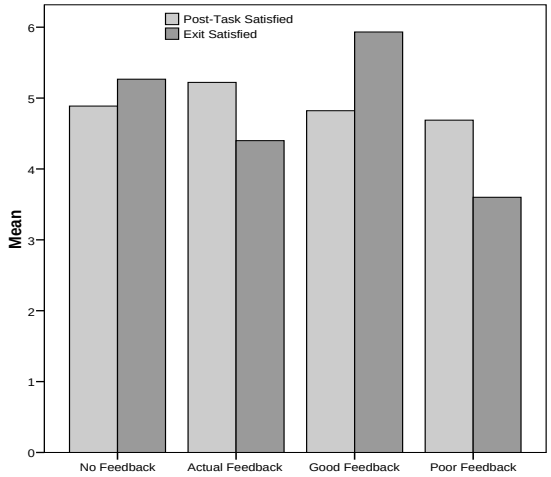

Figure 6. Paired comparisons of pre- and post-feedback mean satisfaction ratings according to condition.

In general, the results are as expected. Satisfaction stayed about the same in the NF condition, increased in the GF condition and decreased in the AF and PF conditions. Paired-sample t-tests within each condition demonstrated that the difference between the pre- and post-feedback measures were not statistically significant in the NF condition, but were statistically significant in the other three conditions in the anticipated directions: AF [$t(14) = -3.14$, $p=.007$], GF [$t(14) = 6.29$, $p=.000$] and PF [$t(14) = 3.58$, $p=.003$]. While results in Section 3.3 show that subjects' final evaluations were affected by the type of feedback they received, these results show that providing some type of feedback – whether actual, good or bad – changes the way subjects respond to the same evaluation item.

4. DISCUSSION

Overall, results of this study provide evidence that performance feedback impacts how users' respond to questionnaire items designed to elicit evaluative data about an IIR system. This impact was most pronounced for items regarding satisfaction. We originally hypothesized that feedback condition would impact the ratings in the following way: PF < AF < NF < GF. Although this was the relationship we found with respect to satisfaction, it was not consistent across all results, even for those that were statistically significant. However, in the majority of cases, subjects' ratings in the NF condition were more like those in the GF condition, while subjects' ratings in the AF condition were more like those in the PF condition.

We explore three possible explanations for our findings: (1) subjects' perceptions of how well they performed were inaccurate and providing them with information about this helped them to make more accurate evaluations; (2) subjects' initial ratings were inflated and providing them with information about their performances motivated them to make more critical evaluations; and (3) subjects exhibited an attribution bias as a result of the feedback which resulted in less favorable evaluations. These three explanations are explored in more detail below and will be referred to as the perception, inflation and attribution explanations, respectively.

The perception explanation is supported by a number of our findings. First, it is important to note that this explanation rests on the assumption that actual performance should be correlated positively with system evaluations. Although previous research [10] has shown that these things are not always positively correlated, it is unclear if this is because subjects are unaware of their actual performances when they complete their evaluations or because there is a measurement problem. With respect to our results, we did not find any strong correlations between objective and subjective performance measures on the Post-Task Questionnaire. It is also the case that subjects in the NF group – half of whom were classified in the low performing group (.13-.17 recall) – gave the system some of the highest system evaluations, which would seem to suggest that their performance perceptions were inaccurate. If their perceptions were accurate, one would expect their scores to be more similar to those of subjects in the PF condition rather than the GF condition since their actual performances were more like the performance information that those in the PF condition received. In fact, the ratings of those who received false performance feedback were consistent with the nature of this feedback – those who received PF rated the system less positively than those who received GF.

Another finding in support of the perception explanation is the significant pair-wise difference that was detected between ratings in the AF and NF conditions for the item regarding inconsistencies in the system. Subjects in the AF condition rated the system significantly lower than subjects in the NF condition. This may have been a result of the inconsistencies between

subjects' perceptions of how well they performed and the reality of how well they performed. Overall, subjects in the NF condition provided the most positive rating to this item, perhaps because they did not receive information (i.e., feedback) that conflicted with their perceptions of the interaction.

Many results of this study also support the inflation explanation. One finding that seems to suggest that the feedback corrected an inflation bias was that several of the items affected by feedback condition were about interface features. While items related to system performance and satisfaction measured things that differed for subjects since subjects posed different queries and the system retrieved different documents, the interface items addressed things that everyone experienced in the same way (e.g., posing queries and navigating). If the feedback changed users' perceptions of their performances and therefore allowed them to make more accurate evaluations, then one would expect changes in performance and satisfaction items, but not interface items. If, however, the feedback motivated users to be more critical of the entire system, then one would expect changes in all ratings, even interface ratings, which is what happened in this study.

Another finding that supports the inflation explanation is that although ratings differed according to feedback condition and changed across two points in time for the same item, overall, most ratings regardless of condition were still above the scale mid-point, which is consistent with Nielsen and Levy's [16] findings. Again, we can assume that the system was actually above average, but even we do not believe this. This behavior poses some hard questions about rating scales and about how subjects interpret and use these types of scales. For instance, what is average? What is below average? What would it take for a subject to actually use a 1 or 2 rating? Does a system necessarily receive some points just for accepting a query and returning documents even if most of the documents are not relevant? One way to extend this study is to investigate feedback condition in relation to two systems: one that performs poorly and the other that performs well. If users continue to provide above average ratings when the system performs poorly and they receive poor feedback, then this would provide strong evidence for the inflation explanation.

The inflation explanation is also theoretically sound as a lot has been written about it in the measurement literature. It is well-known that people exhibit a number of biases when responding to questionnaire items, including social desirability responding and acquiescence, and are sensitive to characteristics of the measurement tool and context [17]. There is no reason to believe that subjects of IIR studies do not exhibit such biases.

Attribution theory, which explains how people attribute causality to events, also provides a useful lens for interpreting the results of this study [6]. Attribution theory is often used to explain how people perceive the causes of success and failure, and how people attribute blame in such situations. For instance, when a student performs poorly on an exam does he attribute responsibility to himself or his teacher? This theory has also been used to explore how people attribute blame when using computers to accomplish tasks [15]. Although the relationship can be quite complex, the basic finding is that when computer failures occur, users are quite willing to blame the computer. There are two key players required for IIR – the system and the user – and each bear some responsibility for the outcome of the interaction. Attribution theory predicts that when the outcome of this interaction is bad, users will attribute more of the blame to the computer, instead of themselves. In the current study, blaming the computer would translate into lower system evaluations in cases where subjects believed the interaction was a failure and higher system evaluations in cases where subjects believed the interaction was a success. Many of the results discussed above can just as easily be explained by attribution theory.

Despite finding a number of interesting and significant results, the findings were not as consistent as we would have liked. One possible explanation for not finding more consistent results lies with subjects' interpretations of the feedback. We had lengthy discussions about how best to present this to subjects and decided that percent of relevant documents found was the easiest measure to communicate and the most appropriate for the task. However, it is possible that some subjects did not understand what this meant or how to interpret this score. Even subjects who performed very well (e.g., .52) may have thought they did poorly if they interpreted this score in the context of a traditional grading scale, where 90-100% equals an 'A' and 52% equals an 'F.' We did not tell subjects that most people only find about 20-30% of the documents until after the experiment was over, so it could be that subjects' needed help interpreting the performance scores before they completed the evaluations.

Another limitation of this research is that the questionnaire items we used were not from an instrument with established validity and reliability. Instead, we used items that were representative of the types of questions commonly found on IIR evaluation questionnaires and note that historically IIR researchers have not been as concerned about measurement validity and reliability as other researchers who rely on questionnaires to elicit data from human subjects. Thus, there may have been some measurement error caused by the items themselves. Lack of valid and reliable measures for evaluating IIR systems is a perennial problem. Without such measures, we are likely to continue to have a difficult time separating variance caused by the method from variance caused by the system.

5. CONCLUSION

In this study, we have shown that providing feedback to users about their performances changes how they respond to evaluation questionnaires. Most notably, the satisfaction ratings of subjects who were provided with their actual performances before completing an Exit Questionnaire significantly decreased from the satisfaction ratings they provided on Post-Task Questionnaires. Differences were also found between subjects who received no performance feedback and those who received feedback about their actual performance. These results suggest that researchers should provide users with feedback about their performances when this information is available in order to elicit the most valid evaluation data. We note again that providing feedback to users about their performances is not possible or appropriate in all types of IIR studies and these finding primarily inform studies of experimental IIR systems that are designed to support high-recall tasks.

We explored three possible explanations of the data related to user perceptions, inflation and attribution. Each offers reasonable interpretations of the data and it would be difficult to design an experiment that tested all three of these theories simultaneously. It is probably the case that a combination of the three offers the

best explanation anyway. However, it is clear that users' evaluation behaviors are brittle and susceptible to slight changes in the evaluation method. This suggests that more work needs to be done to understand these behaviors before we can adequately evaluate IIR systems.

6. ACKNOWLEDGEMENTS

We'd like to thank Dean José-Marie Griffiths and Eleanor Kilgour for their support of this research through the Fred and Eleanor Kilgour Faculty Development Fund at SILS. We'd also like to thank Professor Barbara Wildemuth for her feedback about this project and manuscript. This study was done as part of the Seminar in Interactive Information Retrieval taught by the first author.

7. REFERENCES

[1] Al-Maskari, A., Sanderson, M. & Clough, P. (2007). The relationship between IR effectiveness measures and user satisfaction. *Proceedings of 30th Annual ACM International Conference on Research and Development in Information Retrieval*, Amsterdam, The Netherlands, 773-774.

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

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

[4] Cleverdon. C. W. (1997/1967). The Cranfield tests on index language devices. In K. Spark Jones & P. Willett (Eds.), *Readings in Information Retrieval*, San Francisco: Morgan Kaufman Publishers. (Reprinted from *Aslib Proceedings*, 173-192.)

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

[6] Försterling, F. (2001). *Attribution: An Introduction to Theories, Research and Applications*. East Sussex, UK: Psychology Press, Ltd.

[7] Frokjær, E., Hertzum, M., & Hornbæk, K. (2000). Measuring usability: Are effectiveness, efficiency and satisfaction really correlated? *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, The Hague, The Netherlands, 345-352.

[8] Harper, D. J. & Kelly, D. (2006). Contextual relevance feedback. *Proceedings of the 1st Symposium on Information Interaction in Context (IIiX)*, Copenhagen, Denmark, 218-234.

[9] Hersh, W., & Over, P. (1999). TREC-8 interactive track report. In D. Harman and E. M. Voorhees (Eds.), *The Eighth Text Retrieval Conference (TREC-8)*, 57-64.

[10] Hornbæk, K. & Law, E. L.-C. (2007). Meta-analysis of correlations among usability measures. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, San Jose, CA, 617-626.

[11] Huffman, S. B. & Hochster, M. (2007). How well does result relevance predict session satisfaction? *Proceedings of 30th Annual ACM International Conference on Research and Development in Information Retrieval*, Amsterdam, The Netherlands, 567-574.

[12] Järvelin, K. & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems, 20*(4), 422-446.

[13] Kelly, D., Harper, D. J., & Landau, B. (2008). Questionnaire mode effects in interactive information retrieval experiments. *Information Processing & Management, 44*(1), 122-141.

[14] Kissel, G.V. (1995). The effect of computer experience on subjective and objective software usability measures. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems Conference Companion*, Denver, CO, 284-285.

[15] Moon, Y., & Nass, C. (1998). Are computers scapegoats? Attributions of responsibility in human-computer interaction. *International Journal of Human-Computer Studies, 49*(1), 79-94.

[16] Nielsen, J., & Levy, J. (1994). Measuring usability – preference vs. performance. *Communications of the ACM, 37*(4), 66-75.

[17] Podsakoff, P. M., MacKenzie, S. B., Lee, J,-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879-903.

[18] Saracevic, T. *(1995.)* Evaluation of evaluation in information retrieval. *Proceedings of the 18th Annual ACM SIGIR Conference on Research and Development of Information Retrieval.* Seattle, WA, 138-146.

[19] Thomas, P. & Hawking, D. (2006). Evaluation by comparing result sets in context. *Proceedings of the Conference on Information and Knowledge Management (CIKM '06)*, Arlington, VA, 94-101.

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

[21] Voorhees, E. M. & Harman, D. K. (2005). *TREC: Experiment and Evaluation in Information Retrieval*, Cambridge, MA: MIT Press.