User Adaptation: Good Results from Poor Systems

Catherine L. Smith & Paul B. Kantor
Rutgers University, SCILS
4 Huntington Street
New Brunswick, NJ

csmith, kantor @scils.rutgers.edu

ABSTRACT

Several recent studies have found a weak relationship between system performance and search "success". We hypothesize that searchers are successful because they alter their search behavior. To clarify the relation between system performance and search behavior, we designed an experiment in which system performance is controlled in order to elicit adaptive search behaviors. The study includes 36 subjects, each of whom completed 12 searches using either a standard system or one of two degraded systems. Using a general linear model, we isolated the main effect of system performance, by measuring and removing main effects due to searcher variation, topic difficulty, and the position of each search in the time series. We find that searchers using our degraded systems were *as successful* as those using the standard system, but that in achieving this success they *altered their behavior* in ways that could be measured, in real time, by a suitably instrumented system. Our findings suggest, quite generally, that some aspects of behavioral dynamics may provide unobtrusive indicators of system performance.

Categories and Subject Descriptors

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

General Terms

Experimentation, Human Factors.

Keywords

Experiment design, Analysis techniques, User modeling, Adaptive IR Systems

1. INTRODUCTION

Search systems learn about their users from direct inputs such as query terms and possibly from contextualizing or personalizing information. Models underlying search systems use these parameters to select a maximally relevant set for the searcher, however, the ideal set is generally not returned. Searchers have learned to overcome this difficulty, at least to the extent necessary to make the systems useful. We ask what people are *doing to maximize the performance* they get from search systems. If we can learn what people do to overcome system failure, perhaps we can build systems that monitor user behavior for indicators of failure, so that adaptation can become a two-way street.

We report a factorial experiment in which we intentionally manipulated a standard search system to produce two types of degraded results, and studied how people solve the problem of search failure. We compared searches conducted with the degraded systems to those conducted using the standard. We found that the searchers changed their search behavior, and by doing so, were able to achieve the same level of success as did those using the standard system.

2. RELATED WORK

Several recent studies have suggested that using a better search system may not always lead to improvements in search outcomes. In a precision-focused task, Turpin and Scholer [5] asked searchers to find a single relevant document as quickly as possible. Searches conducted using degraded systems were completed just as quickly as were those using better systems, with no difference in search success. Studying a recall-focused task Allan, Carterette, & Lewis [1] found that searcher productivity was different only at the extremes of system performance (*bpref* < 60% *and* *bpref* > 90%); across the center of the performance range, no significant difference was found. User error rates were not significantly affected by system performance at any point in the full range. Together, these findings suggest that searchers adapt their search behavior to compensate for variability in system performance. This is, of course, a rational response for anyone who has learned (through repeated use) that search system performance varies considerably depending on the search topic [2], and for whom no better alternative is available.

The idea that adaptive behavior can be measured in real time is supported by Turpin and Hersh [4]. Searchers using an enhanced system for a question-answering task, answered questions correctly in the same proportion as did users of a baseline system, and did so more efficiently. That is, users of the baseline system submitted 3 times as many queries and displayed more documents in achieving comparable success. The study also found, for an instance-recall task, that the difference in the number of queries entered was not statistically significant. Data from both experiments are summarized in Figure 1. The trends in both situations suggest that one tactic for adapting to lower system performance is to issue more queries over the course of a search. We examine this question further in the study reported below.

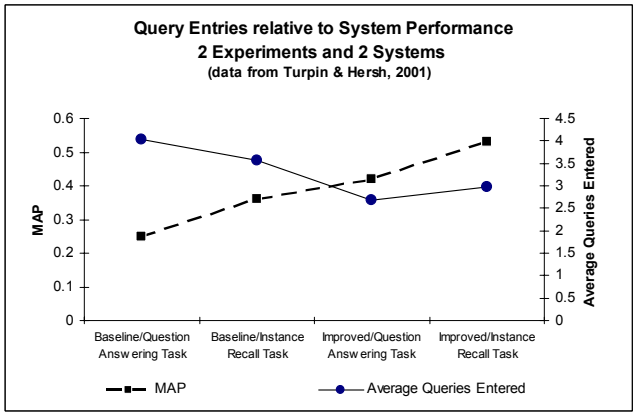

Figure 1. Query Entry and System Performance, from [4]

3. METHOD

3.1 Experimental design

3.1.1 Subjects

36 subjects were recruited on the campus of a large mid-Atlantic university. Subjects included undergraduate and graduate students, as well as non-students. They were paid $15 for their time, and to motivate search effort, were told that an additional $40 would be paid to the subject who "found the most good information sources, and the fewest bad sources." Each subject searched for information on the same set of 12 topics, searching for four topics in each of three "blocks". Subjects were randomly assigned to one of three conditions: a control condition or one of two experimental conditions. All sessions were conducted over a three week period in the same quiet, isolated room. Subjects were told that they would search using the Google® system, which was, in fact, the standard system behind the experimental interface.

3.1.2 Protocol

Introduction. All subjects signed an informed consent form and completed a pre-experiment questionnaire, which collected demographic information and information about prior search experience and attitudes.

Search task assignment. Next, subjects were informed of their experimental task in a mock job description, which provided context for the activity. The subject's job was to find as many "good information sources" as possible for an unspecified "boss" who needed information about 12 topics. A *good information source* was described as one "*you* could and *would* use to get information about the topic." Subjects were told that there was no time limit on searching, and that "there may be some times when there is little or no good information on a topic."

System training. Next, they were shown the experimental system, as the researcher demonstrated it by searching for an example topic. Subjects were then required to practice by searching on at least one practice topic, with an option to continue practicing until ready to begin the experimental task (35 did only one practice topic; one subject did 2). The example topic and practice topics were the same for all subjects. After they practiced, and any questions were answered, subjects started the first topic.

Experiment. When each of the 12 topic searches was started, the system displayed the topic statement and subjects completed a paper pre-search questionnaire. When ready, subjects clicked "start topic search" and saw the search interface. Using it, subjects entered queries in the search box, received search results, browsed the items listed in the results lists, "clicked open" any items to inspect a corresponding information source (website or other document form), and entered additional queries as needed. Information sources were opened in a second monitor, so that subjects saw the list of items and the information source at the same time. Subjects used the Firefox® browser for within-page search and to navigate through opened sources without restriction.

The display of each results list included a checkbox next to each item, which subjects used to "tag" the corresponding source if it was judged to be a "Good Information Source" (GIS) for the topic. Any item that had been previously tagged by the subject as a GIS for the topic, was displayed subsequently with the check already placed in the checkbox, indicating that the item had already been identified as *good*. Subjects could uncheck the checkbox. We scored the assessment of the *goodness* of each item based on the subject's *last* indication given. Subjects could judge (that is, tag as GIS*)* only sources displayed on a results list (i.e., if a subject found a good source by exploring a website, that source could be tagged as GIS *only* if the subject could induce the system to present it in the search results). When done searching on a specific topic, the subject clicked a series of buttons to confirm completion and could not, thereafter, return to the topic.

After a topic was completed, the system prompted the subject to complete a paper post-search questionnaire for that topic. Next, the subject clicked a "continue" button and the cycle of Topic-display→Pre-search-questionnaire→Topic-search→Post-search-questionnaire began anew. Subjects repeated this sequence until all twelve topics were complete or until they quit the experiment.

Debrief. Finally, subjects completed a post-experiment questionnaire, were debriefed regarding the deceptive aspects of the experiment, and received payment for participation.

3.1.3 Instruments

The pre-experiment questionnaire collected demographic information and information about prior search experience and attitudes. Table A details attitudinal and experience-related information gathered. Other questionnaire data is unrelated to the results reported here, and will be discussed elsewhere.

Table A. Pre-experiment Questions on Attitudes toward and Experience with Searching

Subjects were asked to indicate their level of agreement with the statement.	
Statement	Possible Value
I usually find what I am looking for on the Internet or World Wide Web.	1=strongly disagree 6=strongly agree *(6-point Likert scale)*
I am interested in online searching.	
I enjoy trying new ways to use the Internet or World Wide Web.	
I am familiar with Google searching.	
Google can find anything I need.	

3.1.4 Factorial Design

A 3x3 diagram-balanced factorial design was used. Topic order was controlled, with each subject assigned to one of 12 search orders, which balanced topic frequency across the three blocks, with the exception of two topics[1]. One subject in each group searched in each of the 12 order assignments, for a total of 432 searches. Searches were conducted in three blocks of four topics each. The blocks differed in that subject groups searched in different conditions during the middle or treatment block. Block 1 was a *pre-treatment block* (control), in which all three groups searched in the standard condition. During Block 2 (*treatment block*) each group searched in a different condition. In Block 3 all subjects again searched in the standard condition. Subjects were not informed of the blocking, and no break was given between the blocks. Data from the third block has been used in this analysis,

[1] Due to an error, topic 8 was searched one extra time in the treatment block and one less time in the post-treatment block, and topic 6 was searched one less time in the treatment block and one more time in the post-treatment block.

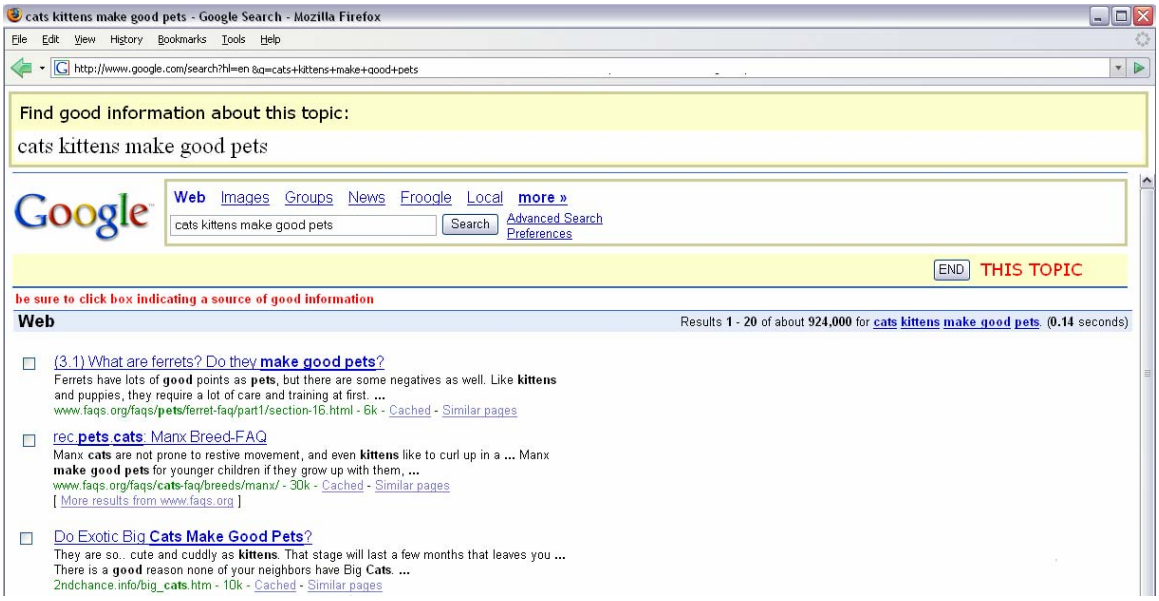

Figure 2. Experimental search interface

however results from only the first two blocks are reported here. Results from the final block will be discussed elsewhere.

In the control condition, the standard search system was used in all three blocks. In the experimental conditions, subjects searched using the standard system in the first block, and then in their assigned experimental condition during the treatment block. One experimental condition, *consistently-low-rankings* (CLR), displayed results from positions 300-320 of a standard retrieved list. The second experimental condition, *inconsistently-low-ranking* (ILR), displayed documents from various ranks in the standard retrieved list.

3.2 Experimental systems

3.2.1 Underlying system

Queries entered by subjects were passed through a proxy server, which stored the queries and other data collected. Queries were submitted to Google® in real time. All queries requested a 20 item list via *url* parameters. All results lists returned by Google® were "scraped[2]". The lists were parsed, advertising and sponsored items were stripped away, and the Google® links "Cached - Similar pages" were removed. The html for each resulting item was stored before display.

3.2.2 The standard system

This displayed items in the order returned by Google®, with the top-ranked item first, and all subsequent items returned in order. During the treatment block, subjects in the control group continued to receive standard results from Google®.

3.2.3 The experimental systems

The two experimental conditions were created by manipulating both the queries submitted by subjects and the search results returned by Google®. Starting ranks were altered according to the subject's assigned condition, as follows: For the consistently-low-rankings (CLR) condition, the query always requested a list starting at Google®'s 300th ranked item. The design was intended to mimic the failure of a system with little or no information in a topic domain. For the inconsistently-low-ranking (ILR) condition, the starting point of the displayed list varied within a topic search; some lists did not start at the 300th ranked item (see Table B). This design was intended to mimic a system with maladaptive mechanisms such as automatic query expansion, where the system fails to converge correctly on the search topic.

Table B. Starting Rankings for the ILR Condition

Queries	Rankings Displayed *(displayed as rankings 1 – 20)*
First, Second	300 – 320
Third	120 – 140
4th-5th	300 – 320
6th	1 - 20
7th	300 – 320
8th	120 – 140
9th – 10th	300 – 320
11th	1 – 20
12th to last	300 – 320

3.2.4 Equipment

One monitor displayed the experimental system. Web pages and documents opened by subjects were displayed in a second monitor. All subjects used the Firefox® browser and each used a familiar type of computer mouse.

3.2.5 Interfaces

The experimental system has two interfaces: (1) for interaction with information about and instructions for the experiment, including the introductory text, requests to complete paper

[2] "Scraping" is a process that extracts data from a webpage. The experimental system formatted the scraped data in the modified display, which was returned to the subject (see 3.2.5).

surveys, and the display of topic statements prior to search and (2) the experimental search interface, which was entered from the first. It displayed two frames, with the topic statement always visible in the upper frame, and a modified Google® interface in the larger lower frame (see Figure 2, above).

After the first search was completed, the upper frame showed a "reminder" box, reporting total elapsed search time since the start of the first search, the number of topics completed, and the number not yet finished. The box was updated at the start of each topic. The standard navigational links usually appearing on the Google® search interface were displayed but disabled. Every item in the results list was left-aligned and displayed using the text and formatting obtained from Google®. A single checkbox to the left of each item was used to "tag" good information sources. Because each results list was limited to no more than twenty items, "next page" links were visible but disabled. When a list of fewer than twenty items was returned by Google®, only the returned items were displayed. For the two degraded systems, the standard Google® results counts and timing text (e.g. "Results 1 - 20 of about *xxx* for *xxxxx* [query terms]. (0.*xxx* seconds)") was altered to indicate that the list started at rank 1. Any "did you mean…" links and "hint" messages returned by Google® were displayed. The "did you mean…" query suggestion link was "live"; subjects could use the link to submit a query suggested by Google®. The link to each item in the list was live, and subjects clicked those links to open information sources. Buttons were provided for confirming that a search was completed.

Table C: Measurements for each topic search

	Variable Names for:	
Measurement/*description*	**Items Judged to be Good**	**All Items**
total item displays	*GIDs*	*AIDs*
total number of item displays during the search; includes repeated displays of the same item		
unique items	*GUI*	*AUI*
number of unique items displayed during the search; excludes repeated displays of the same item		
tagged item displays	*GTIDs*	*ATIDs*
total number of item displays for items tagged as a GIS by the searcher; includes repeated displays of the same tagged item		
tagged unique items	*GTUI*	*ATUI*
total number of items tagged as a GIS by the searcher; items are counted as tagged only once during the search; this count excludes repeated displays of tagged items		
total good items in the "collection" for the topic	*TGI*	--
total number of items tagged as GIS for the topic by any searcher during the experiment and judged to be a good source by the researcher		

Table D. Variables reported in this analysis

Variable name	Ratio	Description
Measures of System Performance		
GPrec	$\frac{GIDs}{AIDs}$	fraction of displayed items that are good items
GRec	$\frac{GIDs}{TGI}$	fraction of all known good items for the topic that are displayed during the search
Measures of Searcher Performance		
Number of good sources found	--	number of unique good sources found during the topic search (*GTUI*)
Elapsed topic search time	--	elapsed time in minutes from initial query entry to end of topic search (*ETTime*)
Searcher selectivity	$\frac{ATUI}{AIDs}$	fraction of displayed items that a searcher tags as GIS
Searcher accuracy	$\frac{GTUI}{ATUI}$	fraction of the tagged items that are good items
Searcher detection rate	$\frac{GTIDs}{GIDs}$	fraction of the good source displays that a searcher tags as GIS
Measures of Searcher Behavior and System Response		
Query rate	$\frac{QueriesEntered}{ETTime}$	number of queries entered per minute of elapsed topic search time
Average list length	$\frac{AIDs}{QueriesEntered}$	the average length of a displayed list
Unique items per query	$\frac{AUI}{QueriesEntered}$	average number of unique items displayed during the search, per query entered
Item display repetition rate	$\frac{AIDs - AUI}{AIDs}$	fraction of item displays that repeat a previously displayed item

4. Measures

4.1.1 System data

As each search progressed, measures of search experience were logged in a database, including (a) the beginning timestamp for each search, (b) each query entered (with timestamp), (c) codes for messages and query suggestions returned by Google®, (d) each item displayed to the subject and its rank order in the display, (e) a record of each item tagged as a good information source (GIS), and (f) the ending timestamp for each search.

4.1.2 Judgment data

After the completion of all 36 experimental sessions, the researcher (CLS) judged the *goodness* of each tagged GIS source. Sources were identified by the full *url*s used to open them. Because subjects were instructed to search for *good information sources*, not "good entry pages", the following rule was used for judging websites: if the entry page was not good, but one navigational link could be used to reach the needed information,

or if one entry in the site's search mechanism could do so, the source was judged as *good*. The researcher was blind to the search conditions under which each source was tagged. Sources were judged in *url* alphabetical order within topic groups, so all sources tagged for a topic were judged at the same time. A 4-level judgment scale was used: *good*, *marginal*, *bad,* or *missing* (link no longer viable). Overall, the distribution of judgments of tagged items, including items found by more than one subject, was: 51.8% good, 19.3% marginal, 24.4% bad, and 4.5% missing.

4.1.3 Derived variables

Using the data described above, the measurements listed in Table C were computed for each topic search and then used to compute several ratios detailed in Table D. Both tables are above.

5. ANALYSIS

5.1 Subjects

Subject characteristics. Pre-experiment measures revealed no significant differences among the three subject groups with regard to prior experience with and attitudes about web searching and Google®, as well as demographic characteristics (for all measures, ANOVA $F(2,33) < 1.284$, $p>0.289$).

Subject attrition. Three subjects from the control group, two from the CLR group, and one from the ILR group quit the experiment before completing the final block, but after completing all the searches in the first two blocks. Data from their completed searches was retained and data from their incomplete searches was excluded from the analysis, leaving 416 complete topic searches for analysis.

5.2 Isolation of system effects

We know from prior research [2] that several large effects are likely to be present in our data. Search topics vary in difficulty. Searchers have different search styles, predilections, and idiosyncrasies. Searchers conducting a series of twelve searches in an experimental setting are affected by the researcher's demand that they search without a break until the task is complete; searchers grow tired or bored with the task and early searches may be performed differently than later searches. The factorial design of our experiment enables us to isolate the effect of the system from all of these confounding effects, using a general linear model.

The model used was developed to evaluate collaborative searching systems [6] and has also been used in evaluation of Question Answering Systems [7]. For this study, a simple model relates a measure, *y,* produced by a user, *u*, using a system treatment *s* while engaged in searching on a topic *t*, at position *p* in the search order. The equation is:

$$y_{ustp} = \lambda_u^{(U)} + \lambda_s^{(S)} + \lambda_t^{(T)} + \lambda_p^{(P)} + \varepsilon$$

where the main effects are represented by the λ parameters, and the term ε represents other variation unaccounted for in the model, and random error. This model allows us to estimate the size of each of the three confounding effects (User: *U*, Topic: *T*, and Position, *P*) for each measurement, for each search, and to subtract these effects. The resulting measure is composed of the effect of the system plus other variation and random error.

$$y_s = y_{ustp} - (\lambda_u^{(U)} + \lambda_t^{(T)} + \lambda_p^{(P)}) = \lambda_s^{(S)} + \varepsilon$$

Data from all 416 completed searches was used in the computation of y_s for each measurement. Subsequently we report on measurements from which *topic, subject and position effects have been subtracted.*

5.3 Analysis using planned contrasts

The design is a 2 x 2 multivariate analysis with planned contrasts. Each contrast tested a set of first order and second order differences. The first order difference (Δv), the change in the average of any specific measure from the first block of four searches to the average of that measure from the second block of four searches, is computed for each group. This becomes the new dependent variable. The second order difference ($\Delta\Delta v$), which

Table F Contrasts for system performance. *NOTE: Topic, subject and position effects have been removed from this data – see above*

n=48	Pre-treatment Block $m \pm sem$	Treatment Block $m \pm sem$	Δv	$\Delta\Delta v$
		v=GPrec		
control	-0.003 ± 0.013	0.026 ± 0.013	0.029 ± 0.018	*NA*
CLR	0.004 ± 0.008	-0.011 ± 0.008	-0.015 ± 0.011	-0.044 ± 0.022 *
ILR	0.000 ± 0.010	-0.016 ± 0.006	-0.016 ± 0.012	-0.045 ± 0.016 *
		v=GRec		
control	-0.046 ± 0.015	0.057 ± 0.015	0.103 ± 0.021	*NA*
CLR	0.019 ± 0.025	-0.016 ± 0.013	-0.035 ± 0.028	-0.138 ± 0.035 **
ILR	0.027 ± 0.031	-0.041 ± 0.011	-0.068 ± 0.033	-0.171 ± 0.043 ***

How to read the above table. Significant differences are noted in the last column: * α=.05, ** α=.01, *** α=.001. For example, for searches conducted in the pre-treatment block by subjects in the control group, average GPrec was -0.003, with a standard error of the mean (sem) of 0.013. For searches conducted in the pre-treatment block by subjects in the CLR group, average GPrec was 0.004 with sem of 0.008 . For searches completed in the treatment block, for the control group the same measure was 0.026 with sem of 0.013, and for subjects in the CLR group it was -0.011 with sem of 0.008. For the control group, in the treatment block GPrec was 0.029 greater than in the pre-treatment block ($\Delta v_{control}$), with sem of 0.018. For the CLR group, the measure was 0.015 less in the treatment block than in the pre-treatment block (Δv_{CLR}), with sem of 0.011. The change for the control group is greater than the change for the CLR group, with a difference of -0.044 ($\Delta\Delta v_{CLR}$) with sem of 0.022. The difference is significant at α=.05.

measures the effect of our treatment, is the difference between *the change in the measure for the control group* ($\Delta v_{control}$), and *the change in the measure for the treatment group* ($\Delta v_{treatment}$). This second difference was computed once for each treatment group (ILR and CLR). Thus, for variable *v*

$$\Delta\Delta v_{CLR} = (\gamma_{CLR} - \alpha_{CLR}) - (\gamma_{control} - \alpha_{control})$$

$$\Delta\Delta v_{ILR} = (\gamma_{ILR} - \alpha_{ILR}) - (\gamma_{control} - \alpha_{control})$$

where for group *i* (control, CLR and ILR), α_i is the mean of the pre-treatment block and γ_i is the mean of the treatment block.

6. RESULTS

6.1 System performance.

To verify that system performance was altered as intended, we examined the effect of our manipulations on the retrieval results, as summarized in Table F, above. We define *GPrec* as the fraction of displayed items that are good items, measured over all item displays presented during a topic search. This is analogous to system precision, except that we measure the *goodness of a source*, rather than the more traditional *relevance of a document/page. GRec* is defined as the fraction of all known good items for the topic (good items found by all subjects in the experiment) that were displayed during the search. This measure is analogous to system recall, except again, we measure *goodness of a source*, rather than *relevance of a document/page*.

Contrast analysis confirmed that system performance was degraded during the treatment block for searches conducted using either experimental system. For searches using the CLR system, retrieval results had lower GPrec than did searches by the same subjects during the pre-treatment block. The decline in GPrec was significantly different from the change in GPrec for subjects using the standard system in both blocks (the control group) (*v=GPrec*, $\Delta\Delta v_{CLR}$ = -0.044, f=4.789, *df* 1, p<.05). Similarly, the ILR system produced lower GPrec than the standard system. This decline in GPrec was also significantly different from the corresponding change for searches in the control group (*v=GPrec*, $\Delta\Delta v_{ILR}$ = -0.045, f=4.981, *df* 1, p<.05). Results are similar for GRec, with lower GRec for both the CLR group (*z=GRec*, $\Delta\Delta v_{CLR}$ =-0.138, f=11.514, *df* 1, p=.001), and the ILR group (*z=GRec*, $\Delta\Delta v_{ILR}$ =-0.171, f=17.630, *df* 1, p<.001).

6.2 Searcher *performance.*

We examined two basic measures of searcher performance, the *number of good sources found* during a topic search, and the time spent searching (*elapsed topic search time).* Contrast analysis revealed no significant differences between groups for either of these measures. We also considered three additional measures of searcher performance: *selectivity, accuracy, and detection rate*. *Searcher selectivity* is the fraction of the displayed items tagged by the searcher. *Searcher accuracy* is the fraction of the items tagged by the searcher that were judged to be good sources. Contrast analysis reveals no significant differences between groups for either of these measures. Finally, *searcher detection rate* is the fraction of the good source displays that were tagged. For subjects using the CLR system, the searcher detection rate was greater than for searches by the same subjects during the pre-treatment block. This increase is significantly different from the change for subjects in the control group (*v=searcher detection rate*, $\Delta\Delta v_{CLR}$ =0.380, f=13.380, *df* 1, p<.001). A similar result was found for the ILR system, (*v=searcher detection rate*, Δv_{ILR} =0.367, f=28.244, *df* 1, p<.001). Table G below summarizes the contrast results for searcher performance.

Table G Contrasts for searcher performance. *NOTE: Topic, subject and position effects have been removed from this data – see above*

n=48	Pre-treatment Block *m* ± *sem*	Treatment Block *m* ± *sem*	Δ*v*	ΔΔ*v*
v=Number of Good Sources Found				
control	−0.039 ± 0.231	0.280 ± 0.219	0.319 ± 0.318	*NA*
CLR	0.103 ± 0.202	−0.182 ± 0.155	−0.285 ± 0.255	−0.604 ± 0.408
ILR	−0.063 ± 0.242	−0.098 ± 0.193	−0.035 ± 0.310	−0.354 ± 0.401
v=Elapsed Topic Search Time				
control	−0.252 ± 0.398	−0.066 ± 0.195	0.186 ± 0.443	*NA*
CLR	−0.084 ± 0.269	0.122 ± 0.239	0.206 ± 0.360	0.020 ± 0.571
ILR	0.336 ± 0.318	−0.056 ± 0.235	−0.392 ± 0.395	−0.578 ± 0.535
v=Searcher Selectivity				
control	0.011 ± 0.012	0.003 ± 0.007	−0.008 ± 0.014	*NA*
CLR	−0.005 ± 0.006	−0.008 ± 0.007	−0.003 ± 0.009	0.005 ± 0.017
ILR	−0.006 ± 0.005	0.005 ± 0.004	0.011 ± 0.006	0.019 ± 0.011
v=Searcher Accuracy				
control	−0.033 ± 0.040	0.030 ± 0.057	0.077 ± 0.070	*NA*
CLR	0.044 ± 0.035	−0.041 ± 0.055	−0.045 ± 0.065	−0.122 ± 0.095
ILR	−0.010 ± 0.046	0.004 ± 0.058	−0.040 ± 0.074	−0.117 ± 0.099
v=Searcher Detection Rate				
control	0.106 ± 0.040	−0.156 ± 0.055	−0.238 ± 0.068	*NA*
CLR	−0.019 ± 0.042	0.071 ± 0.072	0.142 ± 0.083	0.380 ± 0.108 ***
ILR	−0.091 ± 0.040	0.142 ± 0.049	0.129 ± 0.063	0.367 ± 0.105 ***

Significant differences are noted in the last column: * α=.05, ** α=.01, *** α=.001

6.3 Searcher *behavior* and system response.

We are most interested in measures available to a search system without access to explicit performance feedback or document value judgments. We examined the following measures of searcher behavior and system response: *query rate, average list length, unique items displayed per query, and item display repetition rate.* Table H below summarizes the contrast results.

Query rate is the average number of queries entered per minute. Contrast analysis revealed that the query rate is greater for searches conducted using the CLR system during the treatment block. This increase in query rate was significantly different from the change in the rate for searches conducted by subjects in the control group (*v=query rate*, $\Delta\Delta v_{CLR}$ =0.268, f=7.554, *df* 1, p<.01). Interestingly, no significant difference was found in the query rate for searches conducted using the ILR system.

Average list length is the average number of items displayed in each list, including repeated displays of items. Contrast analysis revealed that the average list was shorter for searches using the CLR system than for searches by the same subjects during the pre-treatment block and that this change was significantly different from that for the control group (*v=average list length*, $\Delta\Delta v_{CLR}$ =-1.171, f=5.023, *df* 1, p<.05). No significant difference was found for searches conducted using the ILR system.

Unique items per query is the average number of unique items displayed for every query entered during a search. Contrast analysis revealed that unique items per query was greater for the ILR system than for the control group (*v=unique items per query*, $\Delta\Delta v_{ILR}$ =1.705, f=5.957, *df* 1, p<.05). No significant difference was found for the CLR system. It is important to note that the ILR system shifted the starting rank of some lists (see table B, above). While 85% of searches experienced at least one shift, only 8% experienced the maximum 4 shifts. Lists that were not shifted always started at the 300th ranked item, the same rank used for all CLR lists. While unique items per query was lowest for ILR, it was not significantly different from that of the CLR group (*v=unique items per query*, $\Delta\Delta v_{CLR\text{-}ILR}$ =0.020, f=2.409, *df* 1, p>.10).

Item display repetition rate is the fraction of item displays that repeat a previously displayed item. There was a significant effect relative to the control group for both the CLR group (*z=item display repetition rate*, $\Delta\Delta v_{CLR}$ =-0.084, f=9.063, *df* 1, p<.01) and the ILR group *z=item display repetition rate*, $\Delta\Delta v_{ILR}$ =-0.104, f=13.727, *df* 1, p<.001).

7. RESULTS

The results of Table F confirm that our manipulation does degrade system performance, as measured by 1) the lower density of good items among those presented and 2) lower coverage of all known good items. Nonetheless, searchers using either of the degraded systems found as many good sources as did searchers using the better system, and they did so in the same amount of time, findings consistent with [1, 4, & 5]. How did searchers manage?

A searcher might simply tag a lot of sources, with the expectation that some of them would be scored as *good*; doing so would lower searcher selectivity. Alternatively, a searcher might use lower quality expectations and select marginal or bad sources, which would reduce searcher accuracy. *Neither* strategy was used by our subjects. Note that our task assignment, "find the most good *and fewest bad* sources possible" was designed to make these strategies unappealing. Instead, success using the poor systems appears to flow from an ability to recognize the few good sources displayed; users of the poor systems have higher detection rates than those who used the standard system. We propose therefore that *searchers adapt their scanning behavior to compensate for poor system performance.*

Table I. Contrasts for searcher behavior and response.
NOTE: Topic, subject and position effects have been removed from this data – see above

n=48	Control Block $m \pm sem$	Treatment Block $m \pm sem$	Δv	$\Delta\Delta v$
v=Query Rate				
control	0.014 ± 0.033	–0.098 ± 0.037	–0.112 ± 0.050	*NA*
CLR	–0.042 ± 0.048	0.114 ± 0.067	0.156 ± 0.082	0.268 ± 0.096 **
ILR	0.028 ± 0.041	–0.016 ± 0.041	–0.044 ± 0.058	0.068 ± 0.101
v=Average List Length				
control	–0.143 ± 0.221	0.384 ± 0.171	0.527 ± 0.279	*NA*
CLR	0.153 ± 0.286	–0.491 ± 0.321	–0.644 ± 0.430	–1.171 ± 0.513 *
ILR	–0.009 ± 0.196	0.107 ± 0.213	0.116 ± 0.289	–0.411 ± 0.518
v=Unique Items per Query				
control	0.384 ± 0.346	–0.391 ± 0.305	–0.775 ± 0.461	*NA*
CLR	–0.085 ± 0.321	–0.240 ± 0.376	–0.155 ± 0.494	0.620 ± 0.676
ILR	–0.299 ± 0.299	0.631 ± 0.260	0.930 ± 0.396	1.705 ± 0.634 *
v=Item Display Repetition Rate				
control	–0.025 ± 0.015	0.038 ± 0.014	0.063 ± 0.021	*NA*
CLR	0.010 ± 0.015	–0.011 ± 0.010	–0.021 ± 0.018	–0.084 ± 0.027 **
ILR	0.014 ± 0.015	–0.027 ± 0.007	–0.041 ± 0.017	–0.104 ± 0.024 ***

Significant differences are noted on the last column: * α=.05, ** α=.01, *** α=.001.

Our degraded systems failed in two different ways, producing different response characteristics, and we find evidence that searchers responded differently to the two systems. Searchers using the CLR system (the consistently degraded condition) were more likely to receive short results lists than were searchers using the standard system. These users entered more queries per minute than did searchers in the other two groups. Searchers who used the ILR system (the inconsistent condition) received a greater number of unique items for each query entered, and their query rate did not increase. We propose that *searchers increase their query rate when faced with some, but not all, types of system failure*. This idea is supported by the findings of [4].

Users of either degraded system received fewer repeated displays of items than did searchers using the standard system. Since the CLR system always delivers the same results for the same query, this finding suggests that *searchers faced with a failing system are less likely to reenter previously entered queries during a topic search*.

8. CONCLUSIONS AND FUTURE WORK

Conclusions. We find that searchers using either degraded system were as *successful* as those using a standard system. In achieving this success, they adapted their behavior in ways that depend on the characteristics of the failure. Two adaptations appear, on average, to be indicative of the degraded performance: (1) an increase in the rate of query entry (a user behavior) and (2) a decrease in the rate of repeated item displays (a system response characteristic). Both of these can be observed, in real time, by a suitably instrumented system.

Limitations and future work. These results are encouraging, but further work is necessary. For example, the lower detection rate among users of the standard system may be due, at least in part, to other factors. It may reflect *satisficing* behavior on the part of subjects in the control group; subjects facing a relative abundance of good sources may simply select the best among them, without trying to optimize by seeking and selecting *all* the available sources. A similar effect, termed *saturation*, has been reported [4]. It is also possible that our pre- and post-search instruments (not shown here for reasons of space) which asked subjects to predict the number of sources they would find, caused subjects to "anchor" on the expected number of sources. "Anchored" subjects may stop searching when their expectations are met, even though additional sources are available. Finally, subjects using the standard system may have been able to attend less to searching and more to *differentiating* the goodness of sources. This phenomenon would probably lower the rate of agreement in tagging, which can be explored by examining within-group agreement levels in future analysis.

On the other hand, the design of the CLR system may have produced an exaggerated effect, as each list it returned started at the 300th ranked item, and CLR users were more likely to receive empty or shorter lists. This may have alerted them to the degraded performance, increasing their attentiveness. As reported, when faced with the degraded system CLR searchers increased their pace of query entry. The increased speed may have caused more typos and misspellings in query terms, which would in turn increase the likelihood of empty or short lists, reinforcing rushed behavior in a cycle of failure. The possible effect of the length of returned lists can be examined in future experiments.

In sum, however, this study finds significant and observable differences in user behavior when the system is not serving them well. Most crucially, from the point of view of system design, we must determine whether the differences in user behavior (amplified here, by our experimental design) are large enough that a system can effectively estimate whether it is serving or failing its user.

9. REFERENCES

[1] Allan J., Carterette B. and Lewis J. When will information retrieval be 'Good Enough'? In *28th annual international ACM SIGIR conference on Research and development in information retrieval.* (Salvador, Brazil). ACM, New York, NY, USA, 2005, 433-440.

[2] Lagergren E. and Over P. Comparing interactive information retrieval systems across sites: the TREC-6 interactive track matrix experiment. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval.* (Melbourne, Australia). ACM, New York, NY, USA, 1998, 164-172.

[3] Sun Y. and Kantor P. B. Cross-Evaluation: A new model for information system evaluation. *J. Am. Soc. Inf. Sci. Technol., 57*, 5 (2006), 614-628.

[4] Turpin A. H. and Hersh W. Why batch and user evaluations do not give the same results. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval.* (New Orleans, Louisiana, United States). ACM, New York, NY, USA, 2001, 225-231.

[5] Turpin A. and Scholer F. User performance versus precision measures for simple search tasks. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval.* (Seattle, Washington, USA). ACM, New York, NY, USA, 2006, 11-18.

[6] Wacholder N., Kelly D., Kantor P., Rittman R., Sun Y., Bai B., Small S., Yamrom B. and Strzalkowski T. A model for quantitative evaluation of an end-to-end question-answering system. *J. Am. Soc. Inf. Sci. Technol., 58*, 8 (2007), 1082-1099.

[7] Morse, E.L., Scholtz, J. Kantor, P., Kelly, D. and Sun, Y. (2005). An Investigation of evaluation metrics for analytic question answering. ARDA Metrics Challenge Report. [Available from E. Morse emile.morse@nist.gov]

.