User Adaptation: Good Results from Poor Systems

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

csmith, kantor @scils.rutgers.edu

ABSTRACT

Several recent studies have found only a weak relationship between the performance of a retrieval system and the "success" achievable by human searchers. We hypothesize that searchers are successful precisely because they alter their behavior. To explore the possible causal relation between system performance and search behavior, we control system performance, hoping to elicit adaptive search behaviors. 36 subjects each completed 12 searches using either a standard system or one of two *degraded* systems. Using a general linear model, we isolate 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 are *as successful* as those using the standard system, but that, in achieving this success, they *alter 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 seek a maximally relevant set of "returns", however, the set returned is generally not ideal. Searchers have learned to overcome this difficulty, at least to the extent necessary to make systems useful. We ask what people are doing to *maximize the performance* of 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.

SIGIR'08, July 20–24, 2008, Singapore.

We report a factorial experiment in which we intentionally manipulated a standard search system to produce degraded results, and studied how people solve the problem of search failure. We compared searches that used the degraded systems to those that used the standard. We found that searchers changed their search behavior, and by doing so, were able to achieve the same level of success as did those using a 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. Turpin and Scholer [8] examined how quickly searchers could find a single relevant document. Searches conducted using degraded systems were completed just as quickly as were those using better systems, with no difference in search success. Allan, Carterette, & Lewis [1] found that searcher productivity was different only at the extremes of performance (*bpref* < 60% *and bpref* > 90%); no significant difference was found across the center of the range, and error rates were not significantly affected by system performance. Together, these findings suggest that searchers adapt their search behavior to compensate for variability in system performance. This is a rational response for anyone who has learned, through repeated use, that search system performance varies considerably depending on the search topic [2].

The idea that adaptive behavior can be measured in real time is supported by Turpin and Hersh [7]. Searchers using a baseline system answered questions correctly in the same proportion as did users of an enhanced system, but they did so less efficiently. Users of the baseline system submitted 3 times as many queries in achieving comparable success. In the same study, for an instance-recall task, the difference in the number of queries entered was not statistically significant. Data from both experiments are summarized in Figure 1. The trends suggest that one tactic for adapting to lower system performance is to issue more queries during search. We examine this question in the study reported.

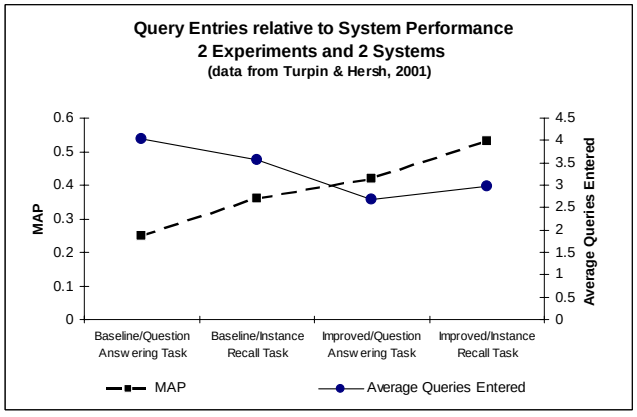

Figure 1. Query Entry and System Performance, from [7]

3. METHOD

3.1 Experimental design

36 subjects were recruited on the campus of a large mid-Atlantic university. Each was paid $15 to search on the same set of 12 topics, 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." Subjects were randomly assigned to experience one of three conditions during the middle 4 searches of their runs: a control condition or one of two experimental conditions. Subjects were told that they would search using the Google® system, which did underlie the experimental system. Subjects provided demographic information and information about prior search experience and attitudes in a pre-experiment questionnaire.

As their experimental task, subjects were asked to find and identify "good information sources" for an unspecified "boss." "Good information source" was defined as one "*you* could and *would* use to get information about the topic." There was no time limit on searching. Subjects observed a demonstration of the experimental system and completed a practice topic before beginning the experiment. Each topic search proceeded as follows: (1) the topic statement was displayed, (2) the subject completed a pre-search questionnaire, (3) the subject searched, and when done searching, (4) the subject completed a post-search questionnaire. Subjects searched as they normally would with the Google® search interface, except that subjects used a checkbox to "tag" each "Good Information Source" (*GIS*) found for the topic. If a tagged source was displayed again during the topic search, the box was displayed as checked; subjects could uncheck the box. We scored the subject's assessment of the *goodness* of each item using the *last* indication given. When done searching on a topic the subject could not return to the topic. Subjects were debriefed after completing the experiment.

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 (see Figure 2). Block 1 was a *pre-treatment* control block, in which all three groups searched in the same standard condition. During Block 2, the *treatment block*, each group searched in a different condition. In Block 3 all subjects were returned to the standard condition. Subjects were not informed of the blocking, and no break was given between blocks. Subjects in the control group searched using the standard system in all three blocks. Subjects in the experimental groups searched using the standard system in the first block, and then in their assigned experimental condition during the treatment block. One experimental treatment, *consistently-low-rankings* (CLR), displayed results from positions 300-320 of lists retrieved by Google®. The other experimental treatment, *inconsistently-low-ranking* (ILR), displayed documents

Block	1 - pre-treatment				2 - treatment				3 - post-treatment				
Position #	1	2	3	4	5	6	7	8	9	10	11	12	Group
System	standard				standard				standard				Control
	standard				Inconsistently low rankings				standard				ILR
	standard				consistently low rankings				standard				CLR

Figure 2. Block structure

from various ranks in the standard retrieved list. Data from the third block has been studied, but only results from the first two blocks are reported here. Results regarding the final block will be discussed elsewhere. Further discussion of the design can be found in [5].

3.2 Experimental systems

3.2.1 Underlying system

Queries entered by subjects were passed through a proxy server, which submitted the queries to Google® in real time, and stored the queries and other data collected. Standard Google® *url* parameters were used to request 20-item lists. All results lists returned by Google® were "scraped[2]" using screen-scraper®[4]. All 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

The system displayed items in the order returned by Google®, with the number-1-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

For the consistently-low-rankings (CLR) condition, the query transmitted by the proxy server requested a list starting at Google®'s 300th ranked item. This design was intended to mimic the failure of a system having little or no information in a topic

Table 1. 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

[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.

[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).

domain. For the inconsistently-low-ranking (ILR) condition, the starting point of the displayed list varied within a topic search (see Table 1, above). The design was intended to mimic a maladaptive mechanism, such as an automatic query expansion process that fails to converge correctly on a search topic.

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.

3.2.5 Interfaces

The experimental system had two interfaces: (1) controlled the experiment, including introduction, requests to complete questionnaires, and initial display of topic statements and (2) controlled searching with the topic statement displayed in an upper frame, and a modified Google® interface in a larger lower frame (see Figure 3).

After the first search was completed, the upper frame displayed a "reminder" box, reporting total elapsed 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. Standard navigational links 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 enabled the searcher to tag as GIS each good source found. "Next page" links were visible but disabled, so that subjects could not view results on subsequent pages. When a list with 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)") were altered to indicate that the list started at rank 1. Any "did you mean…" and "hint" messages returned by Google® were displayed. The "did you mean…" query suggestion link was "live"; subjects could use it to submit the revised query. The link to each item in the list was live, and subjects clicked those links to open information sources.

4. Measures

4.1 System data

As each search progressed, measures characterizing the search experience were logged, 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.

4.2 Judgment data

After all 36 experimental sessions were completed, 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 displayed entry page was not good, but one navigational link from it could be used to reach the needed information, or if one (obvious) entry in the site's search mechanism could do so, the source was judged *good*. The researcher was blind to the search conditions under which each source was tagged. All sources tagged for a topic were judged at the same time, in alphabetical order by *url*. A 4-level judgment scale was used: *good*, *marginal*, *bad,* or *missing* (link no longer viable).Sources were judged as *marginal* if they were about the topic, but did not cover all aspects of the topic statement. Sources that were not about the topic were judged to be "bad." The distribution of judgments, including items found by more than one subject, was: 51.8% good, 19.3% marginal, 24.4% bad, and 4.5% missing.

4.3 Derived variables

Using the data described above, the measurements listed in Table 2 were computed for each topic search and then used to compute ratios detailed in Table 3. Both tables are below.

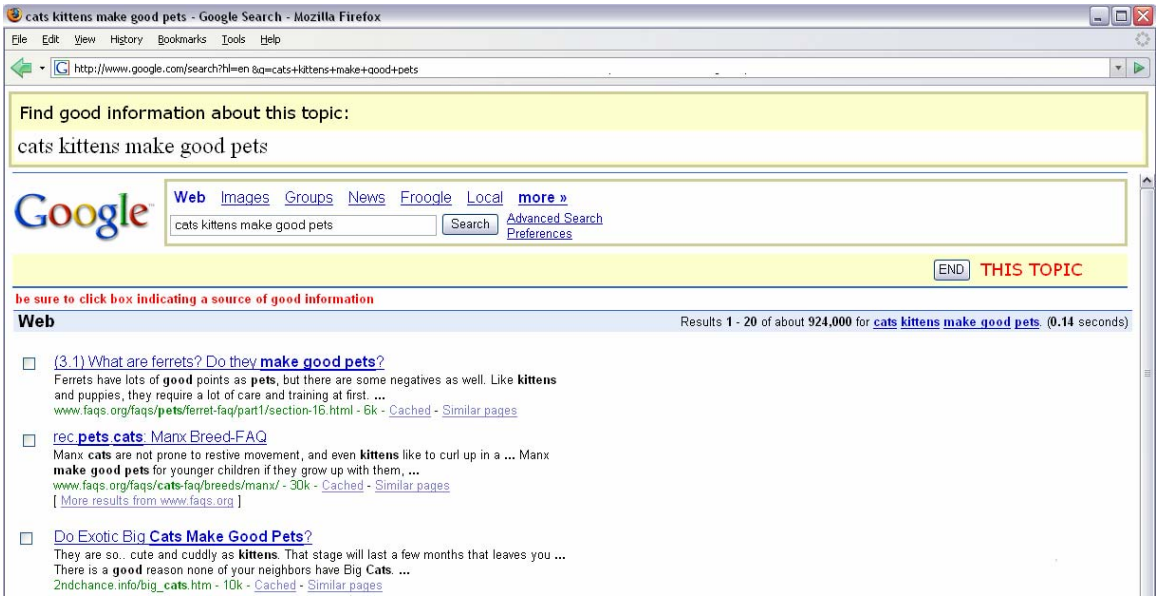

Figure 3. Experimental search interface

Table 2. Measurements for each topic search

Measurem'nt / *description*	Variables: Items Judged Bad	Items Judged Marginal	Items Judged Good	All Items
total item displays	--	--	*GIDs*	*AIDs*
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*
number of item displays for items tagged as GIS by the searcher; includes repeated displays of the same tagged item				
tagged unique items	*BTUI*	*MTUI*	*GTUI*	*ATUI*
number of items tagged as 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				

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®. No significant differences were found in the demographic characteristics of each group (for all measures, ANOVA $F(2,33) < 1.284$, $p>0.289$).

Subject attrition. Six subjects (3 control, 2 CLR, 1 ILR) quit the experiment before completing the third block, but after completing all the searches in the first two blocks. Data from their completed searches was retained and data from incomplete searches excluded. 416 topic searches are used for analysis.

5.2 Topic searches

The general characteristics of the 416 completed topic searches are presented in Table 4.

Table 4. General characteristics of 416 topic searches

Variable	Mean	S.E.M.	Min.	Max.
ETTime	6.53	.158	1.53	22.73
Queries Entered	5.52	.218	1	30
ATUI	3.32	.127	0	13
GTUI	1.72	.103	0	12
MTUI	.64	.046	0	6
BTUI	.81	.063	0	11

Table 3. Variables reported in this analysis

Variable name	Ratio / Variables	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	*GTUI*	number of unique good sources found during the topic search
Number of bad sources found	*BTUI*	number of unique bad sources found during the topic search
Number of marginal sources found	*MTUI*	number of unique marginal sources found during the topic search
Elapsed topic search time	*ETTime*	elapsed time in minutes from initial query entry to end of topic search
Good item ratio	$\frac{GTUI}{ATUI}$	fraction of the tagged items that are good items
Marginal item ratio	$\frac{MTUI}{ATUI}$	fraction of the tagged items that are marginal items
Bad item ratio	$\frac{BTUI}{ATUI}$	fraction of the tagged items that are bad items
Searcher selectivity	$1-\left(\frac{ATIDs}{AIDs}\right)$	fraction of item displays that a searcher does not tag as GIS
Searcher sensitivity	$\frac{GTIDs}{GIDs}$	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 repetitions	$\frac{AIDs - AUI}{AIDs}$	fraction of item displays that repeat a previously displayed item

5.3 Isolation of system effects

We know from prior research [2] that three large effects are likely to be present in our data: (1) *search topics* vary in difficulty, (2) *searchers* have different search skills, styles, cognitive abilities and domain knowledge, and (3) searchers conducting a series of twelve searches in an experimental setting are likely to experience *learning effects*. Searchers 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 these three confounding effects, using a general linear model. The model used was developed to evaluate collaborative searching systems [9] and has also been used in evaluation of Question Answering Systems [3, 6].

For our study, this model relates 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 random error. This model allows us to estimate the size of each confounding effect (User: *U*, Topic: *T*, and Position, *P*) from the ensemble of measurements, for each search, and to subtract these effects. The resulting measure comprises the effect of the system plus random error.

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

For example, the calculation for *elapsed topic search time* (ETTime) for user 2 (subject 2), working on topic 2, in position 1 (the first search in the experiment) is: $\text{ETTime}_{u2,t2,p1} = 9.283$ minutes. We model the main effects of *U*, *T*, and *P* and compute $(\lambda_2^{(U)} + \lambda_2^{(T)} + \lambda_1^{(P)})$. In this example case the parameters are $\lambda_2^U = 3.549$, $\lambda_2^T = 1.993$, $\lambda_1^P = 3.017$, thus the model estimates $ETTime'_{u2,t2,p1} = 8.559$. We compute and save the difference between the measured value and the estimated value (9.283 - 8.559 = 0.721). This value is the effect of the system used during the search, plus random error. In cases where the model "overshoots" the measured value, the saved value will be negative. The values are visualized in Figure 4, below.

Data from all 416 completed searches were used in the computation of $y_s + \varepsilon$ for each measurement. In what follows, we discuss measurements from which *topic, subject and position effects have been subtracted* and report on the 288 searches completed during the first two blocks (36 subjects x 8 searches each).

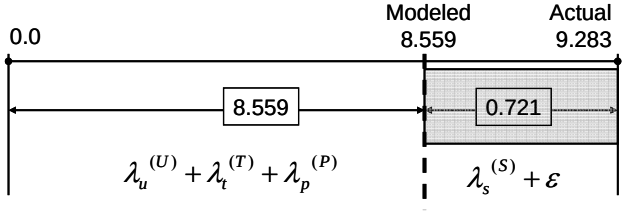

Figure 4. Calculation of saved value for $\text{ETTime}_{u2,t2,p2}$

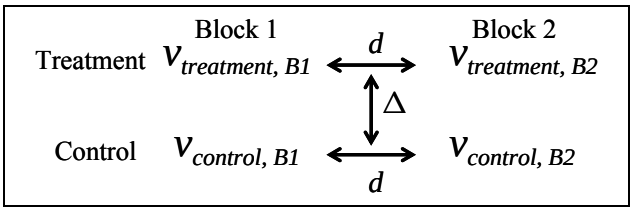

Figure 5. Contrast components

5.4 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 (d_*v*), the change in the average of any specific measure from the first block of four searches to the average of that measure in the second block of four searches, is computed for each of the three subject groups. This becomes the new dependent variable. The second order difference (Δ_*v*), measures the effect of our treatment. Figure 5 diagrams the components. Δ_*v* is the difference between *the change in the measure for the control group* $(\text{d\_}v)_{control}$, and *the change in the measure for the treatment group* $(\text{d\_}v)_{treatment}$. thus, for variable *v*,

$$\Delta\_v_{treatment} = (d\_v)_{control} - (d\_v)_{treatment}$$

For example, the contrast Δ_v_{CLR} for elapsed topic search time measures the difference and tests the hypothesis that *d_v* for the control group is equal to *d_v* for the CLR group. As Figure 6 depicts, these two *d_v* values are not significantly different.

Δ_*v* contrasts were computed for each treatment group (ILR and CLR) for each measure. The results for system performance are summarized in Table 5, results for searcher performance in Table 6, and the results for searcher behavior and system response are summarized in Table 7.

6. RESULTS

6.1 Confirming system performance.

Contrast analysis confirmed that during the treatment block system performance was degraded in both treatment systems. For the CLR group, relative to the pre-treatment block in which the standard system was used, GPrec was lower in the treatment block. The decline in GPrec is significantly different from the change in GPrec (a statistically significant increase) for the control group, which used the standard system in both blocks

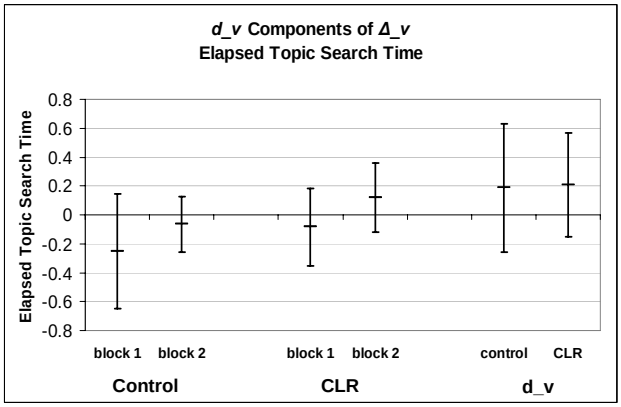

Figure 6. *d_v* components of Δ_v_{CLR} contrast for ETTime

Table 5. Contrasts: system performance. *NOTE: Topic, subject and position effects have been removed from this data*

n=48	$d\_v \pm S.E.M.$	$\Delta\_v \pm S.E.M.$
v=GPrec		
Control	0.029 ± 0.018	
CLR	-0.015 ± 0.011	-0.044 ± 0.022 *
ILR	-0.016 ± 0.012	-0.045 ± 0.016 *
v=GRec		
Control	0.103 ± 0.021	
CLR	-0.035 ± 0.028	-0.138 ± 0.035 **
ILR	-0.068 ± 0.033	-0.171 ± 0.043 ***

Significant Δ_v noted in last column:
* α=.05, ** α=.01, *** α=.001

(*v=GPrec*, Δ_v_{CLR} = -0.044, f=4.789, *df* 1, p<.05). Similarly, the ILR system produced lower GPrec in the treatment block and the decline is significantly different from the corresponding change for the control group (*v=GPrec*, Δ_v_{ILR} = -0.045, f=4.981, *df* 1, p<.05). Results are similar for GRec, with lower GRec in the treatment block for both the CLR group (*v=GRec*, Δ_v_{CLR} =-0.138, f=11.514, *df* 1, p=.001), and the ILR group (*v=GRec*, Δ_v_{ILR} =-0.171, f=17.630, *df* 1, p<.001).

6.2 Searcher *performance.*

We examined four basic measures of searcher performance: the *number of good sources found* (GTUI) during a topic search, the number of bad source found (BTUI), the number of marginal sources found (MTUI) and the time spent searching (ETTime). For both treatment groups, relative to the control group, the block-to-block changes are not significantly different for any of these measures.

We are interested how accurately our subjects identified good information sources. *Good item ratio* is the fraction of the items tagged GIS that were subsequently judged to be good sources. For both the CLR and ILR groups, relative to the control group, the block-to-block change in this ratio is not significantly different. We also examined the characteristics of the tagged items that were not good, using the *marginal item ratio* (the fraction of tagged items that were subsequently judged to be "marginal") and the bad item ratio (the fraction of tagged items that were judged "bad"). For the ILR group, the block-to-block changes for both ratios are not significantly different relative to the control group. For subjects in the CLR group, a block-to-block decline in the marginal item ratio is not significantly different relative to the corresponding change for the control group, but the change in the bad item ratio is. Relative to the pre-treatment block, this ratio was higher in the treatment block, a significantly different change relative to the control group (*v=bad item ratio*, Δ_v_{CLR} = 0.221, f=11.984, *df* 1, p=.001).

We also considered two other measures of searcher performance: *selectivity* and *sensitivity*. Searcher selectivity is the fraction of the displayed items *not* tagged GIS by the searcher. For both treatment groups, relative to control, the block-to-block changes in selectivity are not significantly different. Searcher sensitivity is the fraction of the good source displays that the searcher tagged. For subjects in the CLR group, relative to the pre-treatment block, searcher sensitivity increased in the treatment block. The increase is significantly different

Table 6. Contrasts: searcher performance. *NOTE: Topic, subject and position effects have been removed from this data*

n=48	$d\_v \pm S.E.M.$	$\Delta\_v \pm S.E.M.$
v=Number of Good Sources Found		
control	0.319 ± 0.318	
CLR	-0.285 ± 0.255	-0.604 ± 0.408
ILR	-0.035 ± 0.310	-0.354 ± 0.401
v=Number of Bad Sources Found		
control	0.028 ± 0.213	
CLR	-0.160 ± 0.229	-0.188 ± 0.313
ILR	$0.132 \pm .208$	0.104 ± 0.309
v=Number of Marginal Sources Found		
control	$0.063 \pm .132$	
CLR	$-0.208 \pm .163$	$-0.271 \pm .209$
ILR	$0.146 \pm .149$	$0.083 \pm .221$
v=Elapsed Topic Search Time		
control	0.186 ± 0.443	
CLR	0.206 ± 0.360	0.020 ± 0.571
ILR	-0.392 ± 0.395	-0.578 ± 0.535
v= Good Item Ratio		
control	0.077 ± 0.070	
CLR	-0.045 ± 0.065	-0.122 ± 0.095
ILR	-0.040 ± 0.074	-0.117 ± 0.099
v= Marginal Item Ratio		
control	0.004 ± 0.040	
CLR	-0.039 ± 0.044	-0.043 ± 0.059
ILR	0.029 ± 0.038	0.024 ± 0.058
v=Bad Item Ratio		
control	-0.106 ± 0.045	
CLR	0.115 ± 0.053	0.221 ± 0.070 **
ILR	0.012 ± 0.049	0.117 ± 0.072
v=Searcher Selectivity		
control	-0.003 ± 0.013	
CLR	0.009 ± 0.011	0.012 ± 0.017
ILR	-0.004 ± 0.008	-0.001 ± 0.014
v=Searcher Sensitivity		
control	-0.238 ± 0.068	
CLR	0.142 ± 0.083	0.380 ± 0.108 ***
ILR	0.129 ± 0.063	0.367 ± 0.105 ***

Significant Δ_v noted in last column:
* α=.05, ** α=.01, *** α=.001

relative to control (*v=searcher sensitivity*, Δ_v_{CLR} =0.380, f=13.380, *df* 1, p<.001). A similar result was found for subjects in the ILR group (*v=searcher sensitivity*, Δ_v_{ILR} =0.367, f=28.244, *df* 1, p<.001).

6.3 Searcher *behavior* and system response.

We are specifically interested in measures "available to a search system" that has no access to explicit performance feedback or document value judgments. We examined the following measures of searcher behavior and system response: *query rate,*

Table 7. Contrasts: searcher behavior and system response. *NOTE: Topic, subject and position effects have been removed from this data*

n=48	$d\_v \pm S.E.M.$	$\Delta\_v \pm S.E.M.$
v=Query Rate		
control	-0.112 ± 0.050	
CLR	0.156 ± 0.082	0.268 ± 0.096 **
ILR	-0.044 ± 0.058	0.068 ± 0.101
v=Average List Length		
control	0.527 ± 0.279	
CLR	-0.644 ± 0.430	-1.171 ± 0.513 *
ILR	0.116 ± 0.289	-0.411 ± 0.518
v=Unique Items per Query		
control	-0.775 ± 0.461	
CLR	-0.155 ± 0.494	0.620 ± 0.676
ILR	0.930 ± 0.396	1.705 ± 0.634 *
v=Item Display Repetitions		
control	0.063 ± 0.021	
CLR	-0.021 ± 0.018	-0.084 ± 0.027 **
ILR	-0.041 ± 0.017	-0.104 ± 0.024 ***

Significant Δ_v noted in last column:
* α=.05, ** α=.01, *** α=.001

average list length, unique items displayed per query, and item display repetitions.

Query rate is the average number of queries entered per minute. For the CLR group, query rate increased in the treatment block, a change that is significantly different from the block-to-block change for the control group (for which there was no statistically significant change) (*v=query rate*, Δ_v_{CLR} =0.268, f=7.554, *df* 1, p<.01). Interestingly, for the ILR group, the change in query rate is not significantly different from the corresponding change for the control group.

Average list length is the average number of items displayed in each list returned, including repeated displays of items. For the CLR group, the average list was shorter in the treatment block than in the pre-treatment block, a change that is significantly different from the corresponding change for the control group (*v=average list length*, Δ_v_{CLR} =-1.171, f=5.023, *df* 1, p<.05). For the ILR group, the change in average list length is not significantly different from the change for control.

Unique items per query is the average number of unique items displayed for every query entered during a search. For the ILR group, unique items per query increased in the treatment block, a change that is significantly different from the corresponding change for the control group (*v=unique items per query*, Δ_v_{ILR} =1.705, f=5.957, *df* 1, p<.05). For the CLR group, the block-to-block change in unique items per query is not significantly different from the change for control. It is important to note that the ILR system shifted the starting rank of some lists (see Table 1, above). 85% of searches experienced at least one shift, while 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 highest for the ILR group, the block-to-block change is not significantly different from that of the CLR group (*v=unique items per query*, $\Delta\_v_{CLR\text{-}ILR}$ =0.020, f=2.409, *df* 1, p>.10).

Item display repetitions is the fraction of item displays that repeat a previously displayed item. For subjects in either treatment group, relative to the pre-treatment block, item display repetitions were lower in the treatment block, a change that is significantly different from the corresponding change for the control group (*v=item display repetitions*, CLR: Δ_v_{CLR} =-0.084, f=9.063, *df* 1, p<.01; ILR: Δ_v_{ILR} =-0.104, f=13.727, *df* 1, p<.001).

7. DISCUSSION

Our results (Section 6.1) confirm that our manipulations did degrade system performance, as measured by 1) the lower density of good items 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, 7, & 8]. How did they do it?

A searcher might simply tag a lot of sources, with the hope that some of them would be scored as *good*, a strategy that would be revealed as reduced searcher selectivity; our subjects did not use this strategy. Alternatively, a searcher might lower his or her quality expectations and select marginal or bad sources. Users of the CLR system used to this strategy, at least in part, as revealed by the larger fraction of their tagged items that were judged "bad." Users of the ILR system did not use this strategy. Note that our task assignment, "find the most good *and fewest bad* sources possible" was designed to make strategies such as reduced selectivity or lower quality expectations unappealing, but that they might appear in a different task context.

Relative to those who used the standard system, users of the degraded systems had higher detection rates. This finding suggests that when using the degraded systems, success was achieved, at least in part, because users were able to recognize the few good sources that were displayed. We propose therefore that *searchers adapt their scanning behavior to compensate for poor system performance.*

Our degraded systems failed in two different ways, which should elicit different response characteristics. We find evidence that searchers did respond 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 also 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 the items they find are neither "good" nor "sufficiently various" over several iterations of retrieval results.* This idea is consistent with the findings of [7].

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 resubmit previously entered queries during a topic search.* We plan further analysis of the logged data to investigate this issue.

8. CONCLUSIONS AND FUTURE WORK

Conclusions. We find that searchers using either of the degraded systems were as *successful* as those using a standard system. In achieving this success, they modified 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 occurrence of repeated item displays (a system response characteristic). Both of these can be observed, in real time, by a suitably instrumented system, at the server side.

Limitations and future work. These results are encouraging, but further work is necessary to eliminate other possible effects. For example, the lower detection rate shown by users of the standard system may be due, at least in part, to other factors. It may reflect a kind of *satisficing* behavior on the part of subjects in the control group. Subjects facing a relative abundance of good sources may simply select a few that seem best among them, without trying to optimize by seeking and selecting *all* the available good sources. A similar effect, termed *saturation*, has been reported elsewhere [7]. 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, cause 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, presented with relatively more "good" results, 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 of the data we have collected.

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 (the reasons for this are currently being investigated). These characteristics of the lists may have alerted subjects to the degraded performance, increasing their attentiveness. As reported, when faced with the CLR degraded system, searchers increased their pace of query entry. The increased pace may have caused more misspellings in query terms, exacerbating a cycle of shorter lists and even more rushed behavior, causing further misspellings. Since we recorded the frequency of Google's spelling suggestion messages, we are able to use them as a rough measure of misspellings (a spelling suggestion is unlikely for gross misspellings and typos). We examined misspelling-messages-per-query-entered (MMQE) and misspelling-messages-per-short-list (MMSL). For both measures we found no significant differences between CLR and the control group during the treatment block. (ANOVA: MMQE, f=2.219, *df* 3, p=.087; MMSL, f=.492, *df* 3, p=.689). The possible effect of the length of returned lists will be examined elsewhere.

Taking into account these limitations, however, this study finds significant and observable differences in averaged user behavior when the system is not serving them well. We note, finally, that we observe these differences in the mean, where accuracy is improved by combing results for several persons, and several searches. For these differences to be useful in system design, we must find that differences in user behavior (amplified here, by our experimental design) are large enough that a system can estimate them during the course of a search, and thus estimate effectively whether it is serving or failing its user.

9. ACKNOWLEDGEMENTS

The authors wish to thank the anonymous reviewers for comments and helpful suggestions. We also thank Todd Wilson, at screen-scraper®, for his system and his help, and our 36 hard working subjects. P.K. thanks Donald Kraft and David Evans for helpful conversations on modeling users rigorously (although we are not there yet).

10. 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] 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]

[4] screen-scraper®. www.screen-scraper.com

[5] Smith, C.L. and Kantor, P.B., User Adaptation: Good Results from Poor Systems. Laboratory for Advanced Information Research (LAIR) Technical Report, LAIR/TR-08/01. (2008)

[6] 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.

[7] 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.

[8] 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.

[9] 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.