
While it is true that many factors can influence information seeking success, the amount of effort exerted by the user in completing a task can be a key factor. This study examined the log files of high performing and low performing users of an experimental system to examine differences in their behaviors performing retrieval tasks. The study examined aspects of elaboration, redundancy, depth, and effort to determine which might be used to predict user success in a retrieval task. Results did not show significant differences between the groups according to elaboration, redundancy, and depth, but did reveal significant differences when comparing the following effort measures: number of documents opened, number of piles used, and number of documents placed into piles. Results indicate that the amount of effort is a good indicator of success in information seeking tasks.

Headings:

    Searching

    Information Storage and Retrieval Systems

    Evaluation/Information Storage and Retrieval Systems
A COMPARISON OF RETRIEVAL STRATEGIES BETWEEN HIGH PERFORMERS AND LOW PERFORMERS OF AN EXPERIMENTAL IR SYSTEM

by
Earl W. Bailey Jr.

A Master’s paper submitted to the faculty of the School of Information and Library Science of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Science in Information Science.

Chapel Hill, North Carolina
April 2008

Approved by

______________________________
Diane Kelly
1. INTRODUCTION

Research in retrieval has for many years focused on measuring, describing, and even quantifying success in information seeking, with the laudable goal of making information easier to locate. Initial research on systems and channels has given way to a focus on the user and how that user encounters and creates meaning from the process (Case, 2002, p226). Current research focuses on studying the three Rs – Representation, Retrieval, and Reasoning.

Numerous models have been developed to represent the process of information seeking or retrieval, including anomalous states of knowledge (Belkin, 1980), psychological relevance (Sperber & Wilson, 1986), information search process (Kuhlthau, 1993), Leckie’s general model (Leckie, 1996), sense making (Dervin, 1999), and berry picking (Bates, 1989). Individual search behaviors have been categorized into models as well, including successive fractions (Meadow & Cochrane, 1981), pearl growing (Markey & Atherton, 1978), interactive scanning (Hawkins & Wager, 1982), and building blocks (Harter, 1986). Likewise, many measures of performance have been proposed, including recall, precision, relevance, average search length, fallout, asymmetric uniqueness, and expected search length (Losee, 1998).

Each of these models and measures have gone through the standard cycle of new ideas, with initial enthusiasm over the “New Kid in Town” (Eagles, 1976) giving way to more realistic evaluations or the next new idea. But research has shown that few behaviors are found in isolation, rather combined in actual practice (Marchionini, 1995).
This normative combining of behaviors is useful when examining and quantifying particular users, but is less useful when attempting to predict how a user might perform upon a given task. In fact, much of the finding of research in the field of information retrieval has been inconclusive or difficult to directly relate to actual users in a predictive manner. This is a measure of how difficult the field is to actually study rather than a deficit of researchers’ intents or efforts.

Determining the competence of a user by their behavior is a task that has been studied as a part of research into interactive retrieval systems and artificial intelligence systems, with the goal of using these determinations to tailor system output more towards the actual user. Studies comparing expert behaviors and novice behaviors also touch upon the differences in users interacting with a particular system. Unfortunately, few studies have focused on differences in how these two groups perform searches and whether these differences can be used to predict performance on retrieval-based tasks.

This study examines and compares two groups of subjects using an experimental retrieval system, a high performing group and a low performing group. Performance was measured based upon the subjects’ calculated Recall, or the ratio of the number of documents they selected as relevant to the task and the total number relevant in the collection. The study specifically reports on the retrieval methods of these groups as potential indicators of group membership. Our original research question was: Is amount of effort a better predictor of search success than use of specific tactics? The broadness of the topic led us to split this question into several other research questions, some of which we expected to show no significant difference between groups. These questions are:
Q1: Are there significant differences between the groups in the amount of effort?  
Q2: Are there significant differences between the groups in depth of results viewed?  
Q3: Are there significant differences between the groups in their use of search tactics?  

We hypothesized that the differences between the high performing and low performing groups would be seen in effort (Q1) and depth (Q2) rather than in use of search tactics (Q3). Since this study makes use of multiple comparisons, we set our significance threshold to 0.01 to reduce the risk of false positives. Formally, these hypotheses are shown below:

H1: There are significant differences between the groups in measures of effort.  
H2: There are significant differences between the groups in measures of depth.  
H3: There are no significant differences between the groups in measures of search tactics.  

Effort is defined by Merriam-Webster as the total work done to achieve a particular goal (http://www.merriam-webster.com). For retrieval, we chose the variables number of documents opened, number of documents placed into piles, number of piles used, number of search iterations, number of search terms, number of unique search terms, and search time to represent effort. For H2, Depth is measured by number of results viewed by the user during the course of the search.

To measure use of search tactics for Q3, we consulted the literature and chose the model used by Hembrooke, Granke, Gay, and Liddy (2005) to compare the search behaviors of domain experts and novices. This model organized search tactics into two
broad groups, Elaboration and Redundancy. Elaboration is represented by the specific tactics of broadening and refining. Redundancy is represented by specific tactics of backtracking, plural making and taking, topic term usage, poke-n-hope, and kitchen sink. These specific tactics are discussed in more detail in the literature review. Tactics involving Boolean structures were specifically omitted because the subjects were instructed that they were not supported by the search interface. We expected that we would find no significant differences between groups in the use of any of these tactics.

In short, we predicted that elbow grease would be the key factor in determining successful search interactions rather than use of specific search tactics. Naturally, it is not surprising that effort plays a role in retrieval success since it clearly does so in so many other activities. But measurable differences in variables representing effort could serve to help quantify this facet of user behavior. If observed, these results could have significant impact upon the future design of search interfaces as well as upon the design of future retrieval research.

2. LITERATURE REVIEW

Much of the research analyzing how users might construct queries or navigate through retrieval results has been in the area of Interactive Information Retrieval (IIR) and information-seeking behavior. IIR systems study user search behaviors, both implicit and explicit, to create relevance feedback (Ruthven, Lalmas & Rijsbergen, 2003). This feedback is often used by the system under study to model the user and to use this model to personalize the interface or results. The focus of this research is on the creation of interactive retrieval systems, but the findings are also useful for the examination of user
behaviors. Here, however, generalities are difficult to come by. In fact, even such quantifiable measures as display time are shown to have substantial different meanings across users (Kelly, 2004). Additionally many researchers agree information seeking behaviors can be altered by such variables as task or time (Vakkari, 2000), explaining differences in user behaviors as examples of their influence. While helpful in research pertaining to system design, these explanations do not suffice for predictive user modeling.

A great deal of research has been conducted on the acquisition and effect of expertise, both within a particular discipline, or domain, and related to a particular task or system. Expertise has been called “the ability, acquired by practice, to perform qualitatively well in a particular task domain” (Frensch and Sternberg, 1989). Experts differ from novices both qualitatively and quantitatively, not only having stored more information but also having “organized that information into more structurally and hierarchically meaningful patterns” (LaFrance, 1989). Researchers have examined this difference as it relates to information retrieval, studying its effects upon search term selection (Hembrooke, Granka, Gay, & Liddy, 2005), search tactics (Hsieh-Yee, 1993), search tactic formulation (Wildemuth, 2004), and knowledge acquisition (LaFrance, 1989). Hsieh-Yee (1993) theorized that expert searchers know how to cope with their reduced knowledge while novice searchers are poor enough at searching that domain knowledge did not assist them in formulating tactics. Baloglu (2003) studied both high and low achievers in Education, concluding that high achievers “evaluate the information that they study more critically, organize it conceptually, and make comparisons and contrasts” more effectively than do the low achievers.
Domain expertise as it relates to retrieval can be defined as “knowledge of a subject area (i.e., domain) that is the focus or topic of the search,” and is “conceptually distinct from knowledge of searching techniques” (Wildemuth, 2004). Additionally, domain experts use well-rehearsed strategies to enhance recall and process knowledge, giving them an “enhanced ability to recall the appropriate scripts or schema of their field of expertise” (Solomon, 1992). Vakkari (2002) notes that experts in a subject tend to generate better results than those with less experience, but that novices tend to show more gains when assisted by query expansion or term suggestion. Wildemuth (2004) examined the tactics used by searchers in an effort to understand how they formulate and reformulate search strategies. She also examined differences in those tactics based on domain knowledge. The results were structured into models designed to illustrate the cognitive processes represented by the specific search strategies. She found that her participants moved through the models at greater frequency when their domain knowledge was low, a result she attributed to the number of changes needed by students to retrieve appropriate records. She also noted that the sequence of moves changed as domain knowledge changed, although increased familiarity with the interface might have impacted those results.

Unfortunately, most research in expertise and retrieval also has had a limited focus, primarily examining novice/expert differences with an eye towards assisting novices to perform more like experts or the creation of expert systems to accomplish this goal. Examination of how search behaviors evolve as expertise increases can shed light upon the search behaviors of all users such that their behavior might be used to predict their success or even to afford assistance to the user, but much of the research is
inconclusive or contradictory when generalizing the findings beyond the scope of the experiments.

Examination of search strategies or behaviors has produced limited findings, but the examination of search tactics may prove more promising. While search strategies represent the general roadmap or plan for information seeking, search tactics describe the individual actions taken or move made when conducting a search (Bates, 1979). Tactical actions used in conducting searches have been studied and categorized in many different ways. Bates (1979) created a taxonomy of search tactics with four broad categories with 17 idea tactics. She studied the literature on information searching and reference processes looking for tactics that would improve the effectiveness of a search. Tactics were grouped into the general categories of monitoring, file structure, search formulation, and term. Bates also classified models into 4 groups, those idealizing search, those representing searching, those teaching how to search, and those facilitating searching. She placed this work on tactics in the facilitating group, but it might well be seen as belonging to the idealizing group since it provided no real method by which to incorporate particular tactics.

Much of the early research on search tactics is based upon the use of bibliographic databases. The current use of full-text databases renders this research less useful (Marchionini, 1995). For example, Harter and Rogers-Peters (1985) studied the literature and gathered together the heuristics mentioned in current research. They then classified 101 individual tactics into six broad classes of philosophical attitudes and overall approach; language of problem description; record and file structure; concept formulation and reformulation; recall and precision; and cost efficiency. Many of the heuristics and
tactics were firmly based upon the database systems then in use, and so are of less use when studying full-text searching.

Hembrooke, Granka, Gay, and Liddy (2005) explored differences between experts and novices in generating search terms both with and without feedback. The study measured the presence or absence of search tactics, indicators of complexity, and time per trial. The results indicated that novices engaged in fewer effective strategic search behaviors and that experts used elaboration more frequently and employed more complex and unique terms. The study also developed a typology of search term query strategies and tactics, shown in Table 1.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Tactic</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elaboration</td>
<td>Broadening</td>
<td>The extent to which a user begins with a specific query and expands the scope of the search phrase over successive trials.</td>
</tr>
<tr>
<td>Elaboration</td>
<td>Refining</td>
<td>The extent to which a subject begins broadly and narrows the search with increasing specificity.</td>
</tr>
<tr>
<td>Redundancy</td>
<td>Backtracking</td>
<td>The frequency with which a searcher reuses prior search terms over successive trials.</td>
</tr>
<tr>
<td>Redundancy</td>
<td>Plural Making/Taking</td>
<td>Reflects instances when a user repeatedly incorporates similar nouns into their search attempt, with the slight modification of making the word plural or singular.</td>
</tr>
<tr>
<td>Redundancy</td>
<td>Topic Terms</td>
<td>The extent to which the user incorporates the given query terms as their search terms.</td>
</tr>
<tr>
<td>Redundancy</td>
<td>Poke-n-Hope</td>
<td>The extent to which a searcher retains the same basic structure throughout all search queries, changing only a single word within each trial.</td>
</tr>
<tr>
<td>Redundancy</td>
<td>Kitchen Sink</td>
<td>The extent to which a searcher incorporates search terms related to the subject, but not specific to the query task.</td>
</tr>
</tbody>
</table>

Table 1: Search Strategies (Hembrooke, Granka, Gay, & Liddy, 2005)

The concept of Elaboration represents the user’s capacity to construct a query that reflects understanding of the search topic. This concept comes in part from Nordlie’s
(1999) study of user interactions with OPAC and reference intermediaries. Nordlie noted that the beginning queries were often not sufficient to complete a specified task. These queries might be insufficient for a number of reasons ranging from cognitive to linguistic, and may be opaque to the observer. What could be observed, however, was the steps taken after the initial query to elaborate upon it. (Nordlie, 1999) In Hembrooke’s typology, elaboration is represented by the tactics of broadening and narrowing. The concept of Redundancy, represented by backtracking, plural making and taking, topic terms, poke-n-hope, and kitchen sink, reflects limited understanding of the topic. Subjects employing these tactics might simply reuse terms in multiple queries in an effort to discover new results or might use terms that could be relevant but have no real connection to the task before them. What links these tactics together is the seemingly random nature of the terms selected from query to query. Elaboration tactics, on the other hand, follow observable broadening and refining from one query to the next.

Comparing the search tactics of different groups of users is not a new concept, nor is it one in which the results are likely to be conclusive. In their examinations of the search behaviors of medical students, Wildemuth and Moore (1995) concluded that no generalizations can be made about the “relationship between search performance and the number of statements executed, terms used, citations retrieved, or types of moves used.” Further, they found the strong possibility that the searches might contain syntactical or typographical errors and fail to use any terms from a controlled vocabulary.

Fidel (1984) noted two kinds of searchers, operationalists who use optimal strategies to enhance Precision and conceptualists who use facets to enhance Recall. Conceptual moves were defined by Fidel as moves that change the actual meaning of the
query, such as broadening, narrowing, and related descriptors. Operational moves were defined as moves that use system features to modify a retrieval set, such as synonyms, variant spellings, controlled vocabulary, and date. Fenichel (1980) in contrast surveyed retrieval literature and concluded that there existed considerable variations in search styles and approaches to searching. Further, searchers also tend to use intuition when applying any rules or structures (Fidel, 1991). Research has shown, in fact, that a particular user’s background, experience, and current mental state all have a bearing on their information seeking behaviors (Harter, 1992). Other researchers, e.g. Vakkari (2000) and Kuhlthau (1993), have shown that users are also greatly influenced by where they are in the search cycle.

It is clear that incorporating all the factors that influence search success or failure is a monumental task, perhaps even an unrealistic task. For this reason, many researchers have narrowed their focus to certain aspects, certain tactics, or certain environments. The results of this narrowing focus have yielded interesting results, even if those results cannot then be generalized to other situations. However, the review of literature revealed a lack of research about the relationship between effort and search success.

Zipf (1989) proposed that individual actions are all influenced by a “Principle of Least Effort.” Individuals, according to Zipf, would take actions designed to incur the least effort on their part. Wilson (2006) asserts that “information providers must assume that information users will adopt very simple search strategies in seeking information.” Studies examining end user searches seem to support this principle, showing that end users search differently than expert searchers, less systematically with shorter queries (Markey, 2007). If our users desire the least effort, have mixed results from search
tactics, and see no effect of knowledge upon their searches, it is not at all surprising that
differences between high achievers and low achievers might be better explained by
amount of effort rather than any of those prior concepts.

3. METHOD

This study is a secondary analysis of data collected in another study designed to
test how retrieval system ratings might be influenced by subjects’ perceptions of their
own performances within that system (Kelly, et al, 2008). The original study included 60
subjects and gathered data in log files that were not affected by the manipulations in that
study. The log files and actual performance information from that study formed the basis
of the data for this study. To identify high and low performers for the current study,
subjects from the previous study were ranked by their average recall and placed into three
groups:

   Group 1: High Performers (9 subjects, Recall from .31 to .52)
   Group 2: Average Performers (41 subjects, Recall from .17 to .29)
   Group 3: Low Performers (10 subjects, Recall from .13 to .16)

   The demarcations between the groups were selected based upon clustering of
Recall results. For the High Performers, there was a noticeable gap between Recall
averages of .29 and .31, with a cluster of three users to each side of the gap. Low
Performers were also selected based upon clustering, with the added constraint of
adhering to a similar fraction of the total population. The demarcation between .16 and
.17 was therefore selected to ensure similar sample sizes. Subjects in the high and low
performing groups were included in this study.
3.1 Subjects

The subjects consist of nineteen undergraduates from UNC ranging in age from 18 to 21. The subjects reported their majors as Biology (4), Information Science (4), Business (2), Education (2), Journalism (1), Communications (1), History (1), Psychology/English (1), Linguistics (1), Economics (1) and Biomedical Engineering (1). The subjects are split by gender, 12 male and 7 female. Fourteen subjects reported that they were Fairly Experienced with searching, with three subjects reporting Very Experienced. One subject reported a search familiarity as Fairly Inexperienced. Eighteen subjects reported that they searched the Web at least daily, with one subject reporting usage at least weekly.

Subjects were recruited using a solicitation email sent to all university undergraduates and used an online scheduling form to specify their availability. Subjects were offered $10 compensation to participate in the study. Subjects were contacted via email to finalize the study date and time.

3.2 Collection

The TREC-8 Interactive Track collection (Hersh & Over, 1999) was used in the study. This collection consisted of a corpus of 210,158 articles from the Financial Times of London 1991-1994, a set of aspectual search topics, and a set of relevance judgments. Topics were selected from those used by Harper and Kelly (2006), who modified topics within the TREC-8 system in order to allow use of the relevance judgments already in the corpus. Topics were altered to create situations where users would find exhaustive and
comprehensive information, but retained the original topicality to ensure that the original
relevance judgments would remain valid. Text was also added to the topics to describe
particular information-seeking scenarios. A sample topic is shown in Figure 1.

<table>
<thead>
<tr>
<th>Tropical Storms</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
</tr>
<tr>
<td>You have up to 15 minutes to find and save documents related to this topic.</td>
</tr>
</tbody>
</table>

Fig. 1. Sample Topic.

The relevance assessments come directly from the TREC-8 Interactive Track. The recall measure used is the traditional definition as the proportion of the relevant documents retrieved by users. In this case, placing a document into a pile was used to denote retrieval of that document.

3.3 Protocol

The study used a between-subjects experimental design. Subjects used an experimental information retrieval system to search for newspaper articles relevant to three provided search topics. The subjects were randomly placed in one of the four experimental groups. Topic order and group membership were counterbalanced.
according to a Latin square design. Subjects were tested independently by multiple researchers. Each session lasted for up to one hour.

We adhered to the following procedure during the study. Upon arrival, subjects were greeted by the researcher and escorted to the test facility. Subjects were asked to read a Waiver of Written Consent and to verbally consent to taking part in the study. Documentation of the Waiver was given to subjects for their records. Subjects were asked to complete a demographic questionnaire that gathered background information about them such as age, major, and search experience. Subjects were asked to watch a ten minute system tutorial on the computer using a sample topic. Subjects were then given the first topic and given control of the system. Subjects were instructed to search on the topic until they were satisfied or until time ran out. Subjects were notified that they had approximately 15 minutes to complete the task. Upon completion of the topic or the end of time allocated, the subject was asked to complete a post-task questionnaire evaluating the system while the researcher uploaded their data. Steps 5 and 6 were repeated for two more tasks. Subjects were asked to complete an exit questionnaire evaluating the system. Subjects were given $10 compensation and asked to complete a receipt. The post-task questionnaire was not analyzed for this study because the focus was on log data. The exit questionnaire was also not used in this study because it was implemented after the experimental manipulation in the first study.
3.4 Instruments and Measures

3.4.1 Demographic Questionnaire

The Demographic Questionnaire consisted of six questions. Questions 1-4 were used to detail data for the study sample. Questions 5 and 6 were designed to differentiate inexperienced searchers from experienced searchers. Since the overwhelming majority of subjects reported values of Fairly Experienced or better, these questions were not useful as grouping variables for this study. The questions from this questionnaire are reproduced in Figure 2.

![Demographic Questionnaire](image)

Fig. 2. Demographic Questionnaire.

3.4.2 XRF User Interface and System

The study used the XRF interface developed by Harper and Kelly (2006), a specialized search tool designed to allow users to establish context for their searches. The XRF interface allows users to create customizable piles to sort retrieved documents. The
piles are color coded to assist identification, and documents can belong to multiple piles. Documents can be added or removed from piles as the user desires. The user can see at any time how many documents are in a pile and the name of the pile. Clicking on a pile retrieves the documents placed within the pile. Users can also click on a Similar button to retrieve documents similar to those already in a pile. The system uses a retrieval engine developed with the Lemur toolkit (http://www.lemurproject.org/) with ad-hoc capabilities provided using the Okapi BM25 model and default settings from Lemur. The relevance feedback (RF) searching uses 25 feedback terms and the Lemur KL-divergence language modeling approach. Response times for the system vary from 2.9 seconds for ad-hoc searches to 6.2 seconds for RF searches. All user interactions with the system were logged and archived. The XRF interface is shown in Figure 3.

![Fig. 3. XRF Interface.](image-url)
3.4.3 XRF System Log Files

The XRF System log files recorded the actions taken by the subject during each task. The actions logged included the start time and end time of each search, time stamps for all other activities, an ending time stamp, query strings, use of review pile function, use of similar to pile function, use of more results function, and use of add to pile function.

The log files were analyzed manually to count occurrences of queries, results viewed, documents viewed, documents placed into piles, and piles used. The queries were then examined for examples of Elaboration or Redundancy tactics. For Q1, we noted number of documents opened, number of documents placed in piles, number of search iterations, number of search terms, and number of unique search terms. For unique search terms, we used stemming, since the XRF system did stem words. For Q2, we noted values for results viewed. Finally, for Q3, we noted the use of broadening and refining Elaboration tactics as well as Redundancy tactics of backtracking, plural making and taking, topic term usage, poke-n-hope, and kitchen sink.

4. RESULTS

To answer Q1, we used the grouping variable Group, representing whether the user was a member of the high Recall group or the low Recall group as an independent variable and the dependent measures Docs Opened, Docs in Piles, Piles Used, Search Iterations, Search Terms, Unique Search Terms, and Search Time. We first examined the means and standard deviations of the variables, seen in Table 2.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Docs Opened</td>
<td>Low</td>
<td>30</td>
<td>32.000</td>
<td>15.541</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>27</td>
<td>59.444</td>
<td>15.771</td>
</tr>
<tr>
<td>Docs in Piles</td>
<td>Low</td>
<td>30</td>
<td>11.867</td>
<td>4.091</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>27</td>
<td>33.296</td>
<td>15.762</td>
</tr>
<tr>
<td>Piles Used</td>
<td>Low</td>
<td>30</td>
<td>6.800</td>
<td>2.497</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>27</td>
<td>14.370</td>
<td>4.280</td>
</tr>
<tr>
<td>Search Iterations</td>
<td>Low</td>
<td>30</td>
<td>8.733</td>
<td>5.552</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>27</td>
<td>8.333</td>
<td>4.731</td>
</tr>
<tr>
<td>Search Terms</td>
<td>Low</td>
<td>30</td>
<td>19.733</td>
<td>16.646</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>27</td>
<td>10.852</td>
<td>11.162</td>
</tr>
<tr>
<td>Unique Terms</td>
<td>Low</td>
<td>30</td>
<td>7.333</td>
<td>4.722</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>27</td>
<td>5.259</td>
<td>3.046</td>
</tr>
<tr>
<td>Search Time</td>
<td>Low</td>
<td>30</td>
<td>11.705</td>
<td>2.877</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>27</td>
<td>13.256</td>
<td>1.179</td>
</tr>
</tbody>
</table>

Table 2. Q1 Variable Means.

There are clear differences in the means for Docs Opened, Docs in Piles, Piles Used, and Search Terms. A closer examination also shows a large discrepancy in standard deviation scores for Docs in Piles, indicating a need to test for equality of variances. We examined the significance of the comparisons of the means using a standard t-test in SPSS, also using Levene’s Test to determine whether equal variance could be assumed. As noted above, we set our $p$ value to 0.01 to guard against false positives since we were conducting multiple comparisons. Results are shown in Table 3.

The Levene’s Tests on the variables showed several cases where equal variance cannot be assumed, most notably Docs in Piles, Piles Used, and Search Time. For this reason we will use the Levene Test on all analyzed variables and compare t-test results that do not assume equal variance where indicated. For our comparison of the variables representing effort, the variables Docs Opened, Docs in Piles, Piles Used and Search Time show significance at 0.01 or better when comparing members of the Low scoring and High scoring groups, partially supporting H1.
In order to examine Q2 we used the independent variable Group with the dependent variable Docs Viewed. We first examined the means and standard deviations of the variables, seen in Table 4.

The High performing group shows higher values in both mean and standard deviation. We examined the significance of these differences using a standard t-test in SPSS. As noted above, we set our $p$ value at .01 to guard against false positives since we were conducting multiple comparisons. Results are shown in Table 5.
The variable Docs Viewed, here representing Depth, does not show significance at .01 or better when comparing members of the Low scoring and High scoring groups, refuting our hypothesis H2.

In order to examine Q3 we used the independent variable Group with the dependent variables Refine and Broaden representing Elaboration and the dependent variables Backtrack, Plural M/T, TopicTerms, PokeNhope, and Ksink representing Redundancy. We first examined the means and standard deviations of the variables, seen in Table 6.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Refine</td>
<td>Low</td>
<td>30</td>
<td>2.567</td>
<td>2.885</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>27</td>
<td>1.111</td>
<td>2.118</td>
</tr>
<tr>
<td>Broaden</td>
<td>Low</td>
<td>30</td>
<td>0.933</td>
<td>1.081</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>27</td>
<td>0.667</td>
<td>0.920</td>
</tr>
<tr>
<td>Backtrack</td>
<td>Low</td>
<td>30</td>
<td>1.200</td>
<td>1.297</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>27</td>
<td>0.630</td>
<td>0.967</td>
</tr>
<tr>
<td>Plural M/T</td>
<td>Low</td>
<td>30</td>
<td>0.900</td>
<td>1.517</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>27</td>
<td>0.481</td>
<td>0.753</td>
</tr>
<tr>
<td>TopicTerms</td>
<td>Low</td>
<td>30</td>
<td>14.200</td>
<td>11.496</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>27</td>
<td>8.148</td>
<td>6.809</td>
</tr>
<tr>
<td>PokeNhope</td>
<td>Low</td>
<td>30</td>
<td>3.300</td>
<td>3.323</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>27</td>
<td>1.593</td>
<td>2.422</td>
</tr>
<tr>
<td>Ksink</td>
<td>Low</td>
<td>30</td>
<td>1.800</td>
<td>2.987</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>27</td>
<td>0.778</td>
<td>1.601</td>
</tr>
</tbody>
</table>

Table 6. Q2 Variable Means.

There are clear differences in the means for TopicTerms, and smaller differences in several of the other mean comparisons. Closer examination also shows a difference in standard deviation scores for TopicTerms. We examined the significance of these differences using a standard t-test in SPSS. As noted above, we set our $p$ value at .01. Results are shown in Table 7.
Levene’s tests show that equal variances cannot be assumed for many of the variables in this group, including Refine, Backtrack, Plural M/T, TopicTerms, and Ksink. Thus, for these analyses we use t-test results that do not assume equal variance. None of the variables representing Elaboration or Redundancy show significance at .01 or better when comparing members of the Low scoring and High scoring groups, confirming our hypothesis H3. The significance for TopicTerms is low enough, however, to indicate a potential for further study.
5. DISCUSSION

Our original research questions and hypotheses are summarized below in Table 8.

<table>
<thead>
<tr>
<th>Questions</th>
<th>Hypotheses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1: Are there significant differences between the groups in the amount of effort?</td>
<td>H1: Yes</td>
</tr>
<tr>
<td>Q2: Are there significant differences between the groups in depth of results viewed?</td>
<td>H2: Yes</td>
</tr>
<tr>
<td>Q3: Are there significant differences between the groups in their use of search tactics?</td>
<td>H3: No</td>
</tr>
</tbody>
</table>

Table 8. Research Questions.

Examination of the findings for the Elaboration and Redundancy tactics shows no clear relationship between use of these tactics and membership in either the High or the Low results group. It is clear that none of these variables can be used predictably to determine user performance, answering Q3 negatively as predicted. The results from Hembrooke, Granka, Gay, and Liddy (2005) are not in contradiction to these results as they compared domain experts to novices rather than groups with high and low performance in recall.

The findings for the remaining questions are more mixed. The variable Results Viewed showed no significant difference between groups, in contrast to our prediction for Q2. For this study, the hypothesis H2 that there would be differences in the groups in quantity of results viewed must be rejected. Examination of the remaining variables shows support for a positive answer for Q1. Four of the variables, Docs Opened, Docs in Piles, Piles Used, and Search Time show significant differences between the High and Low groups. This supports our hypothesis H1 that there would be differences in effort between the groups, although not all measures show the same significance. Even with this limitation, it is important to note that the most significant differences noted between
the high and low performing groups is how many documents they opened, how many
documents they put into piles and how many piles they used. Since placing documents
into piles is the stated goal of the exercise (Kelly, 2006), this would indicate that high
performing users simply performed this action more often than the low performing group,
confirming that effort plays a key role in search success. It could also mean that the
simple act of using piles and placing documents into them somehow structures the
process more successfully for the user.

It must also be noted that the experimental interface used for this study was not
designed to support all user behaviors in information seeking. The most notable tactic
unavailable to users is the use of Boolean operators in their searches. While important to
include for completeness, other studies have shown that users do not use these tactics
even when they are available (Borgman, 1996), therefore their lack does not negate study
results. The original study also limited users in the amount of time to search on each
topic, which could have impacted some users more than others. We also note that we set
our significance threshold to 0.01 to reduce the risk of false positives since this study
makes use of multiple comparisons. This is in line with current research indicating the
need for lower significance thresholds when investigating multiple comparisons.

The principle of parsimony asks scientists to consider the least complex
explanation for an observation. Perhaps this is also applicable in measuring overall
retrieval success – that success primarily springs from the effort of the searcher rather
than the specific actions taken by the searcher.
6. CONCLUSION

The study of information seeking covers many topics, including systems benchmarking, interaction modeling, domain expertise, search strategies, and search tactics, but little work has been done on the effects of effort upon search success. While it is true that system and interface design play important roles in success, the amount of effort exerted by the user in completing a task can be a key factor. This study examined the log files of high performing and low performing users to test to see if the high performers expended more effort during retrieval tasks. We also examined tactical user behaviors representing Elaboration and Redundancy to see their effect upon search success. Discovering the effects of effort on information seeking success can help to redirect research into areas of more direct actual benefit to the users of these systems.

Future research into the effects of effort will broaden the applicability by using a more standard search interface to allow for the capture of more types of user behavior. For instance, subjects can be split into groups and given different instruction sets to determine if greater effort, and therefore greater success, can be influenced through instruction. Subjects can also be selected from a more diverse population to test the hypotheses outside of the original sample of university students. It is expected that broadening these variables may cause the results to be less definitive, but using these variables predictably requires a broad data set. The results also indicate strong significance in the use of piles when comparing the two groups. Future research could explore whether it is the use of piles that helps with search success or whether it is a by-product of some other factor in the user.
Can the amount of effort be used to predict results for other users? This study did not specifically address this question, but the results suggest a strong possibility a correlation exists. Clearly there is potential to examine these variables as a possible method for predicating user success in retrieval based tasks.
References


Hsieh-Yee, I.. (1993). Effects of Search Experience and Subject Knowledge on the
Search Tactics of Novice and Experienced Searchers. Journal of the American

Kelly, D. and Belkin, N. J. (2004). Display time as implicit feedback: Understanding task
effects. In Proceedings of the 27th Annual ACM International Conference on
Research and Development in Information Retrieval (pp. 377–384). Sheffield.

Johnson, N. A., Ke, W., Oh, S., Poljakova, A., Rodriguez, M. A., van Noord, M.
users' evaluations of an interactive IR system. UNC SILS Technical Report #TR-


Differences for Knowledge Acquisition. SIGART Newsletter, 108. 6-14.

professionals: A general model derived from research on engineers, health care

Analytic Models of Performance."


