A Task-Oriented Approach to Information Retrieval Evaluation

William Hersh* **and Jeffrey Pentecost**
Biomedical Information Communication Center, Oregon Health Sciences University, Portland, OR 97201.
E-mail: hersh@ohsu.edu

David Hickam
Portland VA Medical Center, Oregon Health Sciences University, Portland, OR 97201

As retrieval systems become more oriented towards end-users, there is an increasing need for improved methods to evaluate their effectiveness. We performed a task-oriented assessment of two MEDLINE searching systems, one which promotes traditional Boolean searching on human-indexed thesaurus terms and the other natural language searching on words in the title, abstract, and indexing terms. Medical students were randomized to one of the two systems and given clinical questions to answer. The students were able to use each system successfully, with no significant differences in questions correctly answered, time taken, relevant articles retrieved, or user satisfaction between the systems. This approach to evaluation was successful in measuring effectiveness of system use and demonstrates that both types of systems can be used equally well with minimal training.

Introduction

As information retrieval (IR) systems proliferate, there is increasing interest in different approaches to indexing and retrieval and how effectively those approaches benefit users. There has been a long-standing debate in the information science community over the best approach to indexing (human vs. automated; thesaurus vs. text-word) and retrieval (Boolean vs. natural language). The proponents tend to fall into two camps, those who advocate thesaurus-based human indexing with Boolean retrieval and those who favor word-based automated indexing with natural language retrieval. The former has been used almost exclusively by commercial retrieval systems, with the latter growing in popularity as end-user systems proliferate.

One of the reasons why the debate has been unresolved is the lack of user-based evaluations in operational settings. Most evaluations of natural language searching systems have been done in laboratory settings, usually with simulated queries (Salton & Buckley, 1988, 1990). A number of researchers have questioned the generalizability of these results (Harter, 1992; Hersh, 1994; Swanson, 1988). There have been a few evaluations comparing human-indexed Boolean systems with automated natural language systems carried out in interactive settings (Hersh & Hickam, 1994, 1995; Robertson & Thompson, 1990; Turtle, 1994), but these were based on the relevance-based measures of recall and precision, which do not capture the value of the complete interaction between the user and system.

The methodologic limitations of these studies has led to introspection about evaluation of IR systems generally. One of us has argued that recall and precision measured only one dimension of IR system performance, and that they cannot be used alone to assess performance (Hersh, 1994). In particular, it was argued that evaluation of systems should focus more on *outcomes* of system use, such as how well the system enabled a user to answer a question, solve a problem, or otherwise meet an information need.

In the medical domain, the need to show benefit is even more pressing, since there is increased scrutiny over health care costs and the quality of care. If health care providers and payers are going to be persuaded to adopt IR technology, it must be shown to be effective in leading to better decisions, actions, and ultimately improved cost-effectiveness and/or quality of care. The goal of this study was to use a task-oriented outcome measure to assess the effectiveness of two MEDLINE systems in helping medical students to answer clinical questions. We view this approach of measuring information attained,

* To whom all correspondence should be addressed.

 CCC 0002-8231/96/010050-07

as measured by the ability to answer questions correctly, as moving beyond recall and precision to begin understanding how IR systems are used, in this case, by clinicians.

IR systems have already made modest inroads into medical education and practice. While the number of practicing physicians using these systems on a regular basis is still small (Curley, Connelly, & Rich, 1990; Williamson, German, Weiss, Skinner, & Bowes, 1989), numerous studies have shown that they will be used when they are made conveniently available in clinical settings (Haynes et al., 1990; Hersh & Hickam, 1994). Studies have also shown that even with a modest amount of training, physicians can perform adequate searches (Haynes, Johnston, McKibbon, Walker, & Willan, 1992). While bibliographic systems still predominate, other resources such as textbooks, drug compendia, and practice guidelines are becoming increasingly available.

While the task-oriented approach used in this study has been implemented before with textbooks (Egan et al., 1989; Hersh et al., 1994), an encyclopedia (Mynatt, Leventhal, Instone, Farhat, & Rohlman, 1992), and a factual database (Wildemuth, deBliek, Friedman, & File, 1995), it has never been carried out with a large bibliographic database. We used this approach to compare performance of a Boolean/human-indexing system with a natural language/automated-indexing system. In addition to measuring success at answering questions, we also compared:

1. User certainty in answering questions;
2. Time to answer questions;
3. Ability to find relevant articles; and
4. Satisfaction with the user interface.

An additional analysis was performed to assess the type of articles cited for evidence in answering the questions. This was done due to the increasing concern that clinical decisions be based on sound scientific evidence (Evidence-Based Medicine Working Group, 1992). Despite the increased prevalence of bibliographic databases, end-users must also be able to critically appraise the articles they retrieve. In the case of articles on therapeutic measures, the strongest scientific evidence comes from randomized controlled trials, which minimize the bias that enters into other types of clinical investigations (Guyatt, Sackett, & Cook, 1993). For articles on assessing diagnostic tests, the best studies come from clinical investigations where there is an independent, blind comparison with a reference standard for the diagnosis and an appropriate spectrum of patients are included for whom the test will be applied in clinical practice (Jaeschke, Guyatt, & Sackett, 1994). Review articles often give concise summaries of treatments and diagnostic tests, but have been found to have flaws in not covering the literature of an area adequately and not providing

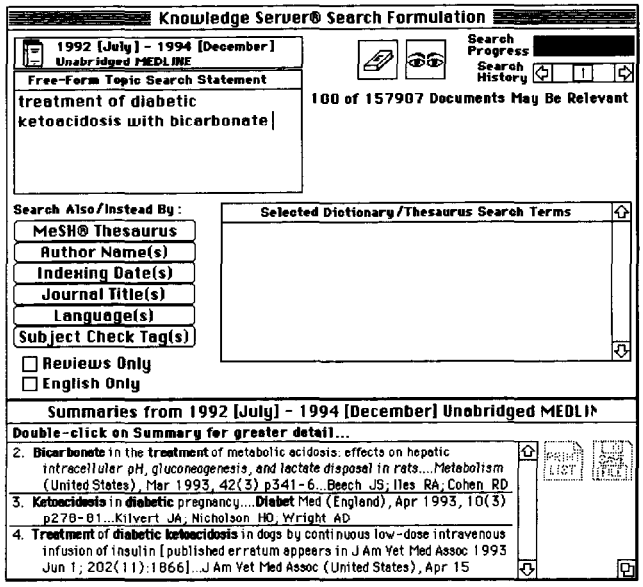

FIG. 1. Knowledge Finder user interface.

sufficient statistical analysis (i.e., meta-analysis) of the data covered (Mulrow, 1987). Commentaries and letters to the editor, which are also indexed in bibliographic databases, are considered to provide even lesser quality evidence. Thus, a final component of the study was to determine what types of articles were cited as evidence for answers to questions, looking to determine whether they had an impact on correctly answering the question.

Methods

The study was performed in the Oregon Health Sciences University (OHSU) Library, which provides access via microcomputers to two commercial MEDLINE systems which epitomize the ends of the Boolean/human-indexing vs. natural language/automated-indexing system spectrum: Knowledge Finder (KF) (Aries Systems, North Andover, MA) and CD Plus (CDP) (CDP Technologies, New York, NY).

KF is one of the earliest commercial implementations of the natural language/automated-indexing approach to bibliographic databases. While it prominently features its "free-topic" query mechanism, it also allows the use of MeSH terms with Boolean operators. KF also features relevance ranking, where retrieved documents are sorted based on the frequency of query terms they contain. Free-topic words can match against any word in the MEDLINE document, including the human-assigned MeSH terms. Originally a CD-ROM product, it now runs on networks, and the latter version used in this study provided access to the full MEDLINE database, segmented by 3–5 year intervals. KF's user interface is shown in Figure 1.

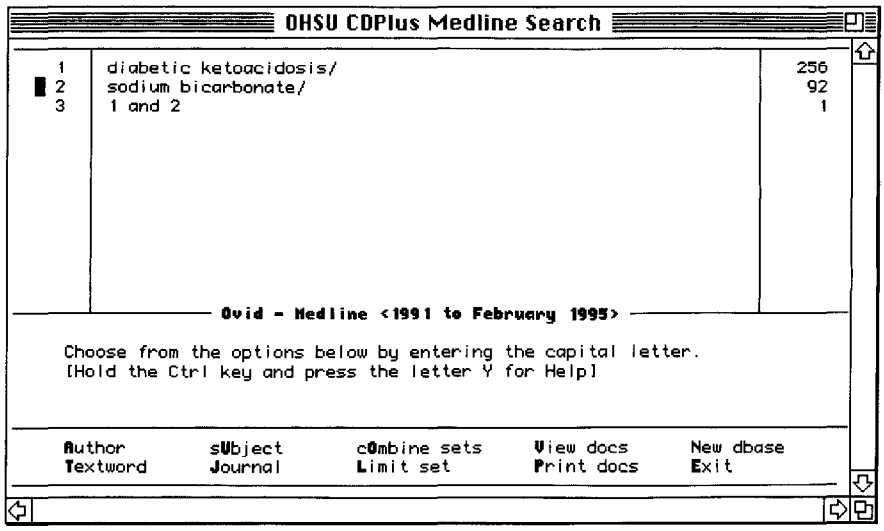

FIG. 2. CD Plus user interface.

CDP features the more traditional Boolean use of sets and, in addition, actively assists the user in selecting MeSH terms as well as applying central concept designations, subheadings, and Boolean operators. CDP also allows text word searching with Boolean operators. Like CDP, it also began as a CD-ROM product and has since migrated to the network. The version used in this study was Unix-based, accessed using a Telnet client on a Macintosh, as shown in Figure 2. It also featured access to the full MEDLINE database, segmented by 4–6 year intervals.

All 96 students in the senior class of the OHSU School of Medicine were sent a letter inviting them to participate in the study, offering a modest honorarium. About 35 students responded, with the study able to accommodate 16. Students were randomized to one of the two systems (KF or CDP) and one of four groups of three questions.

The questions for the study were obtained from an evidence-based medicine project at McMaster University, where physicians are attempting to resolve clinical controversies by searching for the best evidence (Evidence-Based Medicine Working Group, 1992). Thus the questions all had answers which were known to exist in the literature. The only modifications made to the questions were to clarify the prose in some of them and convert the few which were not in yes-no format to that format for consistency. The twelve questions and their answers are listed in Table 1. For each group of three questions, the first two were therapeutic questions, while the third was a diagnostic question.

Students did their searching in groups of two to four. The entire session took approximately three hours. After a brief orientation to the study goals and procedures, they were given a half-hour training session by a medical reference librarian with the system they were randomized to use. Following this session, they took a pre-searching test in which they were asked to answer the questions on which they would later search and rate their certainty in their answer. After the pre-searching test, they reported to their designated terminal in the library to search and attempted to answer their three questions. They were instructed to record any articles they might

TABLE 1. Questions used for evaluation of searching programs.*

1.1 Is clonidine helpful in controlling heart rate in stable atrial fibrillation with rapid ventricular response? Yes.
1.2 Does ACE inhibition improve mortality output in cardiomyopathy? Yes.
1.3 Is ferritin the best non-invasive test for diagnosing iron deficiency anemia in the elderly? Yes.

2.1. Does adding aspirin to low-intensity coumadin increase the risk of life-threatening bleeding episodes? No.
2.2 Is percutaneous endoscopic gastrostomy feeding more efficacious than nasogastric feeding in patients with prolonged neurological dysphagia? Yes.
2.3 Can the hepatojugular reflux accurately predict left ventricular filling pressures? Yes.

3.1 Do corticosteroids decrease mortality in acute alcoholic hepatitis? Yes.
3.2 Is sodium bicarbonate beneficial in correcting the acidosis of severe diabetic ketoacidosis? No.
3.3 Are arterial blood gases useful in the diagnosis of pulmonary embolus? No.

4.1. Is proprylthiouracil effective in reducing mortality from chronic alcoholic liver disease? Yes.
4.2. Are steroids useful in resolving the acute exacerbation of chronic obstructive pulmonary disease? Yes.
4.3 Is the tilt test useful in diagnosing unexplained syncope? No.

* Correct answer is in italics.

TABLE 2. The categories for user certainty of answers in the pre-searching and searching tests.

1. I am very certain about my answer.
2. I am somewhat certain about my answer.
3. I am slightly certain about my answer.
4. I am completely uncertain about my answer.

pursue in the stacks on a special form, which also asked whether they indeed attempted to find the article (it might be unobtainable or superfluous if other articles already answered the question) and whether it was helpful in answering the question. They were again asked to rate their certainty in their answer and also provide evidence in the form of citations to support their answer. (Table 2 lists the scale used for rating certainty in answers from the pre-searching and searching tests.)

Both products had limited capabilities for logging the search results. KF was able to capture the search statement and the number of references viewed by the search, but not which particular ones were actually displayed. CDP captured the search statements, set sizes, and total number of references viewed, but could not capture the references displayed. Due to network problems, some of the CDP searches were not logged. These searches were discarded, and additional searchers were recruited until two searches had been performed on each question with each system.

Upon finishing searching, each student filled out a post-study questionnaire, asking age, sex, past computer experience, past searching experience, and the clinical specialty they planned to enter. They also completed the Questionnaire for User Interface Satisfaction (QUIS) 5.0 to assess overall satisfaction with the system they used (Chin, Diehl, & Norman, 1988).

Question correctness, certainty, time, citations viewed, number of text words, number of MeSH headings, and articles recorded on the special form were assessed on a per-question basis. Correctness was compared with a chi-square test, while the remaining data was compared with repeated measures analysis of variance. Data about users, such as age, gender, past experience, and QUIS result, was compared on a per-user basis. Age and QUIS score were compared using a paired t-test, while the remaining data was compared with a chi square analysis.

Every article cited as evidence for an answer by one or more searchers was assigned to one of the following categories:

1. Randomized controlled trial;
2. Other clinical trial—an intervention but not necessarily randomized or blinded;
3. Clinical study—an observational study without a direct intervention;
4. Review article—a summary article covering other original studies;
5. Letter to the editor; or
6. Comment—an editorial or other opinion piece.

Article types were tabulated for each question by two methods. The first was the total number of each type cited as evidence by all searchers, while the second was the number of unique articles of each type. For the former, articles were counted as many times as searchers cited them, even if they were cited more than once, in order to determine the total number of each type cited by all searchers. The latter provided the total number of unique articles found by all searchers. Because this assessment was descriptive, no statistical analyses were performed.

Results

A total of 18 students participated, though the results of two had to be discarded due to search logging problems. Each question was searched four times, twice by a student using KF and twice by a student using CDP. The pre-searching test showed a trend towards better scores for CDP searchers (Table 3). However, the students' rating of certainty was very low (somewhat better than "slightly certain") and both groups' performance was worse-than-chance. (Pure guessing should have yielded a score of 50%). There were no significant differences between the two search groups in age, gender, computer ownership, previous frequency of literature searching, previous use of CDP or KF, and plans to pursue primary care or specialist careers (Table 3).

For ten of the twelve questions, all users of both systems were able to answer the questions correctly. For the two other questions, one (4.3) was answered wrong by all four searchers while the other (3.2) had one correct

TABLE 3. Characteristics of students who performed searches.*

	KF (n = 8)	CDP (n = 8)
Age (years)	26.5	32.0
Gender		
Male	3	3
Female	5	5
Number who own a computer	6	5
Reported frequency of previous searching		
Weekly	4	4
Monthly	4	4
Previously used CDP	8	8
Previously used KF	2	2
Plan to pursue primary-care specialty	4	6
Mean QUIS 5.0 score	6.94	6.69

* None of the differences were statistically significant.

TABLE 4. Results of searches performed by medical students.

	KF	CDP
Number of searches	24	24
Answers correct prior to search	29.2%	45.8%
Pre-search certainty (mean)	2.63	2.67
Answers correct following search	87.5%	83.3%
Post-search certainty (mean)	1.25	1.54
Time (mean)	27.5	31.6
Citations viewed* (mean)	9.25	55.0
MeSH headings used* (mean)	1.71	4.25
Text words used* (mean)	6.29	2.13
Articles pursued (mean)	2.63	2.67
Articles helpful (mean)	2.04	1.96

* $p < .001$.

answer (with KF) and three incorrect answers. There were no differences between users of the two systems in certainty of answer, time taken, articles possibly pursued, or articles found helpful. There were significant differences between the two systems in citations viewed, number of MeSH terms used, and number of text words used per search. CDP users looked at over six times as many citations. They also used over twice as many MeSH terms and about one-third as many text words (Table 4). There were also no difference between systems in the QUIS questionnaire (Table 3).

The questions answered incorrectly were compared with those answered correctly. There were no differences in certainty of answer, time taken, articles possibly pursued, or articles found helpful (Table 5). The study team also analyzed the wrong answers (questions 3.2 and 4.3) qualitatively to determine why searchers gave the incorrect answer. For the topic in question 3.2, three randomized controlled trials have been performed that show little value in using sodium bicarbonate for diabetic ketoacidosis. The student who gave the correct answer found a review article that described the controlled trials. Of the other three, one found one of the clinical trials but did not cite it, while the other two based their answers on a rat study, several opinion pieces, and review articles

TABLE 5. Comparison of searches yielding correct and incorrect answers.*

	Incorrect	Correct
Number of searches	7	41
Answers correct prior to search	28.6%	39.0%
Pre-search certainty (mean)	2.29	2.71
Post-search certainty (mean)	1.57	1.37
Time (mean)	27.6	29.9
Articles pursued (mean)	3.43	2.95
Articles helpful (mean)	2.85	1.85

* None of the differences were statistically significant.

TABLE 6. Article types cited as evidence for each question.*

Question	RCT	CT	CS	R	LE	C
a. Therapy questions						
1.1	3/1				2/1	
1.2	2/2	4/2		1/1		
2.1		1/1	1/1			
2.2	4/2			3/3		
3.1	1/1			6/4		
3.2	2/1			4/3		
4.1	3/1			3/2		
4.2		3/3		5/5		
Total	15/7	8/6	1/1	22/18	2/1	0/0
b. Diagnosis questions						
1.3			3/3	3/1		2/1
2.3			9/5			
3.3		3/2	2/2			
3.4		6/4	3/2	6/5		
Total	0/0	9/6	17/12	9/6	0/0	2/1

* Column abbreviations represent: Randomized Controlled Trial (RCT), Other Clinical Trial (CT), Clinical Study (CS), Review (R), Letter to the Editor (LE), and Comment (C). Each non-zero cell contains the total number of each type cited and the number of unique references cited for each question.

written before the randomized controlled trials had been performed. The main problem with question 4.3 was poor wording, in particular the adjective "useful." All of the students answered "yes" to this question, while the correct answer from the original question set was "no," based on studies showing a fairly high rate of positive tests in people who have not had syncope (fainting). Most of the searchers found randomized controlled trials that were already using the tilt test to guide treatment (perhaps inappropriately), and therefore assumed it was "useful."

The most common article type cited for therapy questions was the review article (Table 6). The next most common type cited for therapy was the randomized controlled trial, followed by other clinical trials. Thus, in this study, students were able to find correct answers with review articles as well as clinical trials. For diagnosis questions, the most common article type cited was the clinical study, which was to be expected since studies assessing diagnostic tests are not clinical trials with interventions. Indeed, the clinical trials were most likely cited because they showed the diagnostic test in question was being used to apply a therapeutic intervention. The proportions of each type of article were similar for questions answered correctly or not.

Conclusions

This study attempted to assess how two different IR systems assisted medical students in answering clinical questions. We found that they were essentially equiva-

lent, with no significant differences in proportions of correct answers, overall time taken, relevant articles retrieved, or overall user satisfaction. This study reinforced the findings of three out of four other relevance-based interactive comparisons of Boolean and natural language systems (Hersh, Buckley, Leone, & Hickam, 1994; Hersh & Hickam, 1995; Robertson & Thompson, 1990; Turtle, 1994), further showing that end-users can be easily trained and use both types of systems effectively.

While there were no differences in ability to assist in answering questions, there were other differences between the systems. CDP users on average had to look through six times as many references as KF users, although there was no statistically significant difference in overall time taken to answer the question. This finding was due to KF having a search summary that lists the title of each article for each citation which is lacking in CDP, requiring users to look at the citations in a set one at a time. The response time for paging through citations in CDP, however, is very quick, which explains why no overall greater time was required for CDP. CDP searchers also used more MeSH terms and fewer text words, which was expected due to CDP's orientation to MeSH searching and KF's emphasis on text words.

There were no differences in certainty of answer, time taken, or relevant articles retrieved between questions which were answered correctly or incorrectly. The analysis of how the students were led to wrong answers in these searches reveals some problems inherent in using bibliographic systems and information. Particularly in the question concerning the use of sodium bicarbonate for diabetic ketoacidosis, they were able to find the same quantity of information in approximately the same amount of time, but failed to retrieve or cite the pertinent clinical trials that settled the issue.

The types of articles cited as evidence for answers showed that at least in the case of therapy articles, nonprimary sources of data, especially review articles, were used most frequently. Despite the known problems with review articles, these papers were used to answer most questions correctly. Since the number of incorrectly answered questions in this study was so small, we were not able to determine whether the type of article plays a role in ability to correctly answer a question.

In summary, we have shown that medical students can be trained to use both Boolean/human-indexing and natural language/automated-indexing systems effectively to answer clinical questions. Each system required approximately the same time and effort to use. We conclude that the method of evaluating IR systems introduced in this study was effective in assessing whether these systems can be used to solve real-life problems that motivate their use. These measures also provide valuable information about user "performance" that cannot be captured by relevance-based measures.

Acknowledgments

This study was supported in part by grant LM05307 and contract LM07088 from the National Library of Medicine. The authors acknowledge Ms. Dolores Judkins, reference librarian, who provided MEDLINE training to all of the study participants.

References

Chin, J., Diehl, V., & Norman, K. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. *Proceedings of CHI '88—Human Factors in Computing Systems* (pp. 213–218). New York: ACM Press.

Curley, S., Connelly, D., & Rich, E. (1990). Physicians use of medical knowledge resources: Preliminary theoretical framework and findings. *Medical Decision Making, 10,* 231–241.

Egan, D., Remde, J., Gomez, L., Landauer, T., Eberhardt, J., & Lochbaum, C. (1989). Formative design-evaluation of Superbook. *ACM Transactions on Information Systems, 7,* 30–57.

Evidence-Based Medicine Working Group (1992). Evidence-based medicine: a new approach to teaching the practice of medicine. *Journal of the American Medical Association, 268,* 2420–2425.

Guyatt, G., Sackett, D., & Cook, D. (1993). Users' guides to the medical literature: II: How to use an article about therapy or prevention, A: Are the results of the study valid? *Journal of the American Medical Association, 270,* 2598–2601.

Harter, S. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science, 43,* 602–615.

Haynes, R., Johnston, M., McKibbon, K., Walker, C., & Willan, A. (1992). A randomized controlled trial of a program to enhance clinical use of MEDLINE. *Online Journal of Controlled Clinical Trials.*

Haynes, R., McKibbon, K., Walker, C., Ryan, N., Fitzgerald, D., & Ramsden, M. (1990). Online access to MEDLINE in clinical settings. *Annals of Internal Medicine, 112,* 78–84.

Hersh, W. (1994). Relevance and retrieval evaluation: Perspectives from medicine. *Journal of the American Society for Information Science, 45,* 201–206.

Hersh, W., Buckley, C., Leone, T., & Hickam, D. (1994). OHSUMED: An interactive retrieval evaluation and new large test collection for research. *Proceedings of the 17th Annual International ACM Special Interest Group in Information Retrieval* (pp. 192–201). Dublin: Springer-Verlag.

Hersh, W., Elliot, D., Hickam, D., Wolf, S., Molnar, A., & Leichtenstein, C. (1994). Towards new measures of information retrieval evaluation. *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care* (pp. 895–899). Washington, DC: Hanley–Belfus.

Hersh, W., & Hickam, D. (1994). The use of a multi-application computer workstation in a clinical setting. *Bulletin of the Medical Library Association, 82,* 382–389.

Hersh, W., & Hickam, D. (in press). An evaluation of interactive Boolean and natural language searching with an on-line medical textbook. *Journal of the American Society for Information Science, 46* 478–489.

Jaeschke, R., Guyatt, G., & Sackett, D. (1994). Users' guides to the medical literature: III. How to use an article about a diagnostic test, A. Are the results of the study valid? *Journal of the American Medical Association, 271,* 389–391.

Mulrow, C. (1987). The medical review article: State of the science. *Annals of Internal Medicine, 106,* 485–488.

Mynatt, B., Leventhal, L., Instone, K., Farhat, J., & Rohlman, D. (1992). Hypertext or book: Which is better for answering questions?

Proceedings of Computer-Human Interface 92 (pp. 19–25). New York: ACM Press.

Robertson, S., & Thompson, C. (1990). Weighted searching: The CIRT experiment. Informatics 10: Prospects for intelligent retrieval. (pp. 153–166). York, UK: ASLIB.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management, 24,* 513–523.

Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science, 41,* 288–297.

Swanson, D. (1988). Historical note: Information retrieval and the future of an illusion. *Journal of the American Society for Information Science, 39,* 92–98.

Turtle, H. (1994). Natural language vs. boolean query evaluation: A comparison of retrieval performance. *Proceedings of the 17th Annual International ACM Special Interest Group in Information Retrieval* (pp. 212–220). Dublin: Springer-Verlag.

Wildemuth, B., deBliek, R., Friedman, C., & File, D. (in press). Medical students' personal knowledge, searching proficiency, and database use in problem solving. *Journal of the American Society for Information Science,46,* 590–607.

Williamson, J., German, P., Weiss, R., Skinner, E., & Bowes, F. (1989). Health science information management and continuing education of physicians. *Annals of Internal Medicine, 110,* 151–160.