Evaluation of Internet Search Engines: Methodological Issues and Assumptions

Louise F. Spiteri & Nathalie Richard
School of Library & Information Studies
Dalhousie University
Halifax. Nova Scotia
lspiteri@is.dal.ca

ABSTRACT

This paper examines two questions: what methodological assumptions underlie existing evaluations of Internet search engines and how do they differ from those used to study information retrieval systems (IRSs)? An examination of 31 Internet-based studies reveals that they employ a methodology very similar to the experimental model that has been used to evaluate IRSs; in other words, the studies are driven by the evaluator, who chooses the topic, formulates the search query, and performs the relevance assessments on the documents retrieved from the search. None of these studies includes a more broadly-based user survey of the effectiveness of the search engines: focus is placed mainly upon the effectiveness of particular features of the search engine, rather than users' satisfaction with the performance of the search engine.

INTRODUCTION

The evaluation of the retrieval effectiveness of information retrieval systems (IRSs) is a well-established domain in the field of Library and Information Science that can trace its theoretical foundations and assumptions to the Aslib Cranfield tests of indexing languages and systems conducted in the late 1950s (Cleverdon 1966). The underlying assumption of these studies is that effectiveness of the IRSs can be judged according to the perceived relevance of the retrieved items. Measures used to determine retrieval effectiveness commonly include recall, precision, relevance, and searcher satisfaction.

Internet search engines differ from "traditional" IRSs in a number of ways. An IRS is normally produced by a single organisation that controls the size of the database (i.e., the number of records it contains), and the number of new records added to the database on a daily, weekly, monthly, or yearly basis. An IRS is geared to a target audience and is relatively cohesive in its content, coverage, and scope. Finally, an IRS might use a controlled vocabulary. An Internet search engine, on the other hand, tends to contain a much larger volume of records that grows at a rate that is rarely monitored on a regular basis; it is extremely difficult to know how many new records are added every day.

An identical search conducted in the same Internet search engine on two different days of the same week could produce very different results. An Internet search engine is not targeted to any one audience and lacks cohesion in its content, coverage, and scope. Finally, most Internet search engines do not use controlled vocabulary.

Given the differences between "traditional" IRSs and Internet search engines, this paper seeks to examine two questions: what methodological assumptions underlie existing evaluations of Internet search engines and how do they differ from those used to study IRSs? In order to attempt to answer these questions, 31 evaluation studies of the retrieval effectiveness of several Internet search engines were examined, with particular emphasis placed upon the following areas:

- The search engines examined
- The qualifications/backgrounds of the evaluators
- The nature of the participants used in the studies
- The nature of the queries used in the studies
- The source of the queries used in the studies
- The criteria against which is measured the effectiveness of the search engines
- The specific measures used to determine the degree to which the search engines met stated criteria

METHODOLOGICAL ASSUMPTIONS IN THE EVALUATION OF IRSs

Performance evaluation of IRSs has been based largely upon two methodologies. The experimental methodology was established in the Aslib Cranfield tests of indexing languages. These tests established the use of relevance, recall, and precision ratios as standard measures by which to evaluate the effectiveness of IRSs. The Cranfield model is an experimental one, in the sense that the queries posed to the IRS are designed by subject specialists, indexers, or the people who work with the IRS. The queries posed need not reflect real information needs, as their primary function is to test the IRS. Document relevance is based upon the degree to which the information "answers" the question (Cleverdon 1966). The MEDLARS test of indexing languages established the operational methodology, whereby real-life users submit their own queries, based upon real information needs, and determine the relevance of the items retrieved, based upon the degree to which the items addressed their information needs (Lancaster 1969).

The problem with the experimental model is that it appears to be artificial: to what extent do the subject specialists et al. reflect the average searcher who uses the IRS? The expert knowledge that these specialists bring with them vis-à-vis their knowledge of the database, as well as their searching and subject expertise, tests only how well the IRS performs for this particular group of users, and not for the larger population that might use the IRS. Since the Internet is used by a wide variety of people and because it contains information that can address a myriad of information needs, the operational methodology would seem to be a more appropriate way to evaluate the performance of Internet search engines. As shall be seen, however, many of the evaluation studies conducted to test the performance of Internet search engines appear to follow, to a large degree, the experimental methodology used to evaluate IRSs.

FINDINGS

The 31 studies (Appendix 1) examined follow the same basic procedure to evaluate the search engines, namely, (a) a series of queries is posed to each search engine, and (b) each engine is evaluated based upon the quality of its design features and, in many cases, upon its ability to retrieve documents that are relevant to the stated query.

Search engines evaluated

The average number of search engines evaluated per study is 5.9, with the highest number being 23 search engines (Falk 1997), and the lowest 2 (Tomaiuolo & Packer 1996; Stanley 1996). The eight search engines evaluated most frequently are:

Lycos:	27 studies	(87%)
AltaVista:	23 studies	(74.1%)
Excite:	20 studies	(64.5%)
Open Text Index:	17 studies	(54.8%)
HotBot:	14 studies	(45.1%)
Infoseek:	13 studies	(41.9%)
Webcrawler:	11 studies	(35.4%)
Yahoo!:	11 studies	(35.4%)

After this point, the next most frequently-evaluated search engines are Magellan, Metacrawler, and SavvySearch, at 4 studies each. The authors do not generally explain their choices of search engines; at best, they base their decisions upon what they consider to be the most popular search engines in use at the time, but provide no bases for this type of assertion.

Qualifications of the evaluators

The authors of the evaluation studies examined fall into three main categories: computer scientists/information technology professionals, librarians, and academics:

Librarians:	13 studies (41.9%)
Computer scientists/Information Technology professionals:	5 studies (16.1%)
Professors in Library & Information Science (LIS) programmes:	3 studies (9.6%)
Professors in non-LIS programmes:	3 studies (9.6%)
Computer Science (1)	
Finance and Economics (1)	
Mathematics (1)	
Unknown	7 studies (22.5%)

The inability to identify the professional qualifications of the authors could be problematic, as without this knowledge, it could be difficult to ascertain the credibility of both the evaluation studies and their results. Furthermore, one is uncertain about the "agenda" of the author; for example, does the author work for any of the search engines under review? Is the author qualified to construct a rigorous evaluation study and to interpret and analyse the data gathered?

Nature of the participants used in the evaluation studies

From a methodological point of view, this aspect of the studies is particularly troubling, as of the 31 studies examined, only 1 uses participants other than the authors (Tomaiuolo & Packer 1996). It is difficult to determine the extent to which evaluator bias affects the findings and conclusions of the other 30 studies. Errors of variability stemming from the evaluators could include the following:

- Evaluator's knowledge of, and familiarity with, the individual search engine
- Evaluator's knowledge of, and familiarity with, the search features of the individual search engine
- Evaluator's ability to construct an effective search query based upon a stated information need
- Evaluator's ability to judge the output of the search query based upon a stated information need

Another problem that arises from these author-driven evaluations is the fact that these studies reveal not much more than the interaction of one person with a particular search engine. The results from such studies can therefore be rather artificial and unhelpful, as they indicate mostly personal preferences of an individual evaluator, rather than how well a search engine functions with a variety of searchers and information needs. Furthermore, how reflective are the evaluators of the typical searcher who uses the search engines? The thirteen librarians who conducted the evaluation studies, for example, work in the reference department, therefore one can assume, with a fair degree of certainty, that these librarians are skilled searchers and quite familiar with a variety of search engines. The queries posed by a reference librarian are quite possibly more sophisticated and specific than those posed, say, by an undergraduate student which, in turn, affects the quantity and quality of the results obtained in the search. Without a broader participant base, it is difficult to reach conclusions about the efficiency of search engines to the population beyond that of librarians, academics, or computer scientists.

Nature of the queries used

Only 11 of the evaluation studies (35%) provide the specific queries that were used to test the search engines. In the remaining 20 studies, it is extremely difficult, if not impossible, to determine the underlying basis for the evaluation of the search engines. The authors do not mention the nature of the information posed to the search engines, nor the exact queries (including search syntax used) that were used. This failure to provide the search statements means that the reader is unable to determine to what extent the results derived from these studies are affected by:

- The nature of the information need/question posed to the search engine
- The nature of the search syntax used to formulate the search query in each search engine
- The consistency of the search statement. In other words, was the same statement submitted tô each search engine using the same search syntax?

Assuming, for instance, that the same question was submitted to three different search engines, to what extent does the search syntax used affect the results of the searches? Was the same search syntax used in all three cases (e.g., a Boolean phrase)? Different search syntaxes can produce vastly different results, e.g., the keyword search *bronchial asthma* is likely to produce a larger number of hits than the Boolean search *bronchial AND asthma* which, in turn, will produce a larger number of hits than the exact phrase search *"bronchial asthma."* One might assume, also, that an exact phrase search will produce results that are more precise/relevant than a keyword search, since the former type of search specifies much more clearly the nature of the relationship between the two search terms.

Source of the queries

In 16 of the studies (51.6%), the evaluators choose their own topics to submit to the search engines and determine the quality of the results obtained from the searches. Of concern is the fact that in the only study that uses participants other than the evaluators (Tomaiuolo & Packer 1996), the participants are provided topics to search by the evaluators (but not the actual search statement). Furthermore, in this latter study, the assessment of the relevance of the retrieved documents is done not by the participants, but by the evaluators). In 51.6% of the studies, therefore, queries submitted to the search engines do not, in fact, reflect the "true" information needs of the person conducting the search. Once again, one is reminded of the experimental model used in the Aslib Cranfield tests, where a set of artificial information needs was designed to test an IRS. One is uncertain, therefore, to what extent the results of these studies are affected by the following factors:

- The searchers' understanding of the topic
- The searchers' interest in the topic
- The searchers' motivation to find "answers" to the topic
- The searchers' ability to judge the quality of the results obtained from the searches

This level of artificiality is even more evident in the one "participant" study, where the person deciding the quality of the results of the search is not the same person who conducts the search. The output of the search is dependent directly upon the searcher's understanding of the topic and the search syntax used to conduct the query (i.e., "input factors"). In this case, however, the persons who determine the input factors of the search are not the ones who determine the output of the search.

Criteria by which the search engines were evaluated.

All 31 articles examined provide the criteria used to evaluate the search engines. The criteria most commonly used are as follows, with an average of 2.5 criteria used per study:

• Search features (e.g., Boolean operators)	20 studies (64.5%)
• Display/output features (e.g., relevance rankings)	15 studies (46.5%)
• Precision of the retrieved documents	13 studies (41.9%)
• Help features	8 studies (25.8%)
• Recall (i.e., number of documents retrieved)	6 studies (19.3%)
• Database size	4 studies (12.9%)
• Speed/response time	4 studies (12.9%)
• Database content	3 studies (9.6%)

With the exception of precision, most of these criteria focus upon design aspects of the search engines. These criteria do not take into account the searchers' educational background, subject expertise, and experience with using Internet search engines, all of which could impact upon the searchers' ability to search effectively and to make relevance assessments. These criteria appear to assume that a search engine works in isolation; as shall be seen, there is no room by which to judge the searchers' satisfaction with the way the search engine works. The focus of the studies, therefore, appears to be upon the search engine as a structural entity, rather than with the degree to which it satisfies the needs of its users.

Measures by which criteria are determined.

Search features: all 20 studies that use this criterion describe the search features available in the search engines evaluated

Display/output features: all 15 studies that use this criterion describe the display/output features available in the search engines evaluated

Precision

The evaluation of IRSs has been based frequently upon the measures of precision and recall. Recall is defined as the total number of relevant documents retrieved vs. the total number of relevant documents in the system. Given the enormous size and mutability of Internet databases, recall is all but impossible to measure, since there is virtually no way of telling how many relevant documents exist in a database on any given topic on any given day. It is for this reason that recall is measured very rarely in these 31 studies.

Precision is based upon the calculation of the number of relevant documents retrieved from the total number of documents retrieved. Since Internet searches often retrieve hundreds, thousands, and sometimes hundreds of thousands of documents, the practicality of calculating a true measure of precision becomes untenable. Ten of the 13 studies that measured precision thus limited their analysis to a specific number of documents retrieved:

- 4 studies: the first 10 documents retrieved
- 1 study: the first 20 documents retrieved
- 1 study: the first 25 documents retrieved
- 4 studies: no indication of how many documents were analysed

The calculation of precision is predicated also upon the presumption of the "relevance" of the document. Of the 13 studies that look at precision, 9 provide no operational definitions of relevance and how relevance is measured: they indicate only whether or not the retrieved documents were found to be relevant. Three studies use a ranking system by which relevance is measured, based upon three- or five-point scale (Clarke & Willett 1997; Ding & Marchionini 1996; Leighton 1995). These ranking systems are described in detail, as are the operational definitions used to measure relevance. One study uses benchmarking as a means of measuring relevance: the authors predetermined which documents would "answer" the question and rate the documents discovered by the participants based upon the degree to which these documents match the benchmarked documents (Tomaiuolo & Packer 1996).

Help features: all 8 studies that use this criterion describe the help features available in the search engines evaluated

Recall: the 6 studies that use this criterion are concerned not with the relevance of the documents retrieved: their definition of recall is the total number of documents retrieved in each search.

Database Size: all 4 studies that use this criterion count the total number of indexed documents contained in each database/search engine

Database Content: all 3 studies that use this criterion describe the quality of the content of selected documents, but no mention is made about the total number of documents examined

Response Time: of the 2 studies that use this criterion, only one explains how response time was measured, namely, the difference between the time at which the query is entered and the time at which the search engine displays the results of the query (Chu & Rosenthal 1996).

SUMMARY OF METHODOLOGIES USED TO EVALUATE INTERNET SEARCH ENGINES

- In most of the studies (96.7%), the subjects of the study are the evaluators themselves, which could lead to a serious concern about the reliability and validity of the results. Furthermore, these results reflect only one person's experience with particular search engines, and not that of a larger population of Internet users

- The evaluators/participants are generally assumed to be experienced Internet users and searchers, and hence might not reflect the variety of search experiences that is true of the larger population of Internet users

- When participants are used, the queries they submit do not reflect their own, real, information needs. This situation could affect the participants' (a) motivation to conduct an effective search; (b) understanding of the nature of the query; and (c) ability to assess the relevance of the documents retrieved. Point (c) might be moot in the one participant-based study, since the evaluators made the actual relevance assessments

- The majority of the criteria and measures used are concerned more with the design features of the search engines, rather than with the searchers' assessment of the utility of these features to their queries

- Relevance is not always defined in the studies; of more concern, perhaps, is the fact that relevance judgements are not always made by the same person(s) who submitted the queries.

- A number of the studies appear to be more in the line of reviews rather than evaluations; they tend to just state which search/output/help features are available, the number of hits retrieved, the size of the database, and so forth. Emphasis in these types of studies is once again the design features of the search engine rather than its utility to a real end-user

COMPARISON OF EVALUATION METHODOLOGIES: IRSs AND INTERNET SEARCH ENGINES

A detailed examination of the methodologies used to evaluate information retrieval systems (IRSs) is beyond the scope of this paper. Several interesting patterns emerge, however, in the evaluation of IRSs, and they contrast quite noticeably, in some cases, with the patterns observed in the Internet studies.

Most recent evaluations of IRS systems follow the operational methodology, i.e., they are searcher-driven: the participants formulate their own queries based upon their own information needs; furthermore, relevance assessments of the documents retrieved are made by the participants. As has been indicated, this is far from the pattern used in the Internet studies where, in most cases, the participants of the studies are the authors themselves. In the case where other participants are used, the queries submitted to the search engines do not always reflect a "real" need on their part. Finally, relevance judgements are made almost exclusively by the authors, even if the latter did not conduct the search themselves. This "Internet model" bears a strong resemblance, therefore, to the experimental model used in the Aslib-Cranfield experiments, whereby evaluation is driven almost exclusively by experts, both in the choice of queries and in the assessments of relevance.

Several criteria and many measures have been proposed and used for evaluating the performance of IRSs, but there is a lack of agreement as to what constitutes successful performance, or which are the best existing evaluation measures. Examples of criteria and measures that have been used in IRS studies include (Saracevic 1988a, 1988b, 1988c; Tague & Schultz 1989):

Criterion	**Measure**
Relevance	Precision
Efficiency	Search session time Relevance assessment time Actual search cost
Utility	Worth of search results in $$ Worth of search results vs. time expended Worth of search results vs. effort expended Value of search results as a whole
User Satisfaction	Searcher's contribution Searcher's understanding of the request Searcher's understanding of purpose of request Searchers' thoroughness in exploring request Searcher's knowledge of use of database Use of appropriate search terms Search results Searcher's satisfaction with completeness of results Searcher's confidence in completeness of results Importance of completeness of results Searcher's satisfaction with precision of results Importance of precision of results to searchers
Success	Searcher's judgement of overall system success

It is interesting to note that only two of the criteria mentioned above are to be found in the 31 studies examined, namely, relevance (using the measure of precision), and efficiency (speed of the search). Unlike their Internet counterparts, these IRS criteria and measures do not incorporate very much the interface of the search system. For example, does the use of particular search features (e.g., Boolean vs. keyword searching), or help features, etc., affect searchers' judgement of system success? On the other hand, these IRS criteria and measures are very much driven by the needs of the searchers, i.e., they attempt to equate the quality of an IRS with the degree to which the system serves the various needs of the searcher. These "user-driven" criteria and measures are noticeably lacking in the 31 Internet studies examined.

CONCLUSION AND RECOMMENDATIONS

The methodology employed by the 31 studies is very similar to the experimental model that has been used to evaluate IRSs; in other words, the studies are driven, for the most part, by the evaluator, who chooses the topic, formulates the search query, and performs the relevance assessments on the documents retrieved from the search. None of these studies includes a more broadly-based user survey of the effectiveness of the search engines: focus is placed mainly upon the effectiveness of particular features of the search engine, rather than users' satisfaction with the performance of the search engine.

Given the growing popularity of Internet search engines amongst a large variety of searchers, it would be useful for evaluation studies of these search engines to adopt the operational methodology used to assess IRSs. In other words, what is needed is to study the ability of the search engines to help searchers find information that is pertinent to their individual needs. The following criteria are therefore suggested for future user-based evaluation studies of the effectiveness of Internet search engines:

Nature of the request: this criterion looks at (a) the general purpose of the request submitted by the searcher, (b) the subject category of the request, and (c) the subject area of the request. Are questions conducted in Medicine, for example, more likely to produce more relevant hits?

Searcher characteristics: this criterion looks at the searchers' educational background and subject expertise to see whether either of these factors affects the searchers' ability to judge the relevance of their hits. This criterion looks also at the searchers' experience with using Internet search engines, as this characteristic could impact upon the success of a search.

Search interface: this criterion is concerned with looking at (a) which features are available in a search engine; (b) which features are used by the searcher; and (c) the searcher's evaluation of the effectiveness of the features that were used in the search

Relevance: this criterion looks at the precision of the hits provided by the search engine. Because of the mutable size of Internet databases, an artificial recall number will need to be established (e.g., assess the relevance of the top 20 hits)

Searcher satisfaction: this criterion looks at the searchers' overall satisfaction with the ability of the search engine to provide information useful for their needs.

The scholarly community has little knowledge about whether the users' search attempts are successful or in vain and whether there might be better ways of guiding users in their information pursuits. There does not appear to be a scientific body of knowledge available on the topic nor an appropriate model for conducting investigations in this area. A series of user-based evaluations of the effectiveness of Internet search engines is thus needed to answer questions such as: which search engines are most effective in meeting users' needs; which search features are used most frequently; which search features are perceived by the users as being most effective, and so forth. Development or improvement of Web-based search tools is proceeding without the benefit of such knowledge; consequently, users of the tools are making their choices of tools in the absence of reliable comparisons and good guidance.

APPENDIX 1: EVALUATION STUDIES OF INTERNET SEARCH ENGINES

Bradley, Phil. (1998). *Multi-search engines: A comparison.* http://www.philb.com/msengine.htm

Brunelle, Bette S. (1996). Smart systems, smart searches. *Online information 96. Proceedings of the International Online Information Meeting, London, England, December 3-5, 1996*

Bullwinkle, Davis. (1997). Critical analysis of selected Internet search engines. *Arkansas libraries*, 54(2/3): 3-10.

Chu, Heting. & Rosenthal, Marilyn. (1996). Search engines for the World Wide Web: A comparative study and evaluation methodology. *Global complexity: Information, chaos and control. Proceedings of the 59th ASIS annual meeting, Baltimore, Maryland, October 21-24, 1996.* ed., Steve Hardin. Medford, NJ: Information Today: 127-135.

Clarke, Sarah J. & Willett, Peter (1997). Estimating the recall performance of Web search engines. *Aslib proceedings*, 49(7): 184-189.

Cooke, Alison. McNab, Alison. & Anagnostelis, Betsy. (1996). The good, the bad, and the ugly: Internet review sites. *Online Information 96. Proceedings of the International Online Information Meeting, London, England, December 3-5 1996.*

Courtois, Martin P. & Baer, William M. & Stark, Marcella. (1995). Cool tools for searching the Web: A performance evaluation. *Online*, 19(6): 14-32.

Ding, Wei. & Marchionini, Gary. (1996). A comparative study of Web search service performance. *Global complexity: Information, chaos and control. Proceedings of the 59th ASIS annual meeting, Baltimore, Maryland, October 21-24, 1996.* Ed., Steve Hardin. Medford, NJ: Information Today: 136-142.

Northwestern University Library. (1997). Evaluation of selected Internet search tools. http://www.library.nwu.edu/resources/internet/search/evaluate.html

Falk, Howard. (1997). World Wide Web search and retrieval. *The electronic library*, 15(1): 49-55.

Feldman, Susan. (1997). Just the answers please: Choosing a Web search service. *Searcher*, 5(5): 44-50, 52-57.

Kimmel, Stacey. (1997). WWW search tools in reference services. *The reference librarian*, 57: 5-20.

Klingener, Alice. (1996). Web searchers: Start your engines. *Business & finance bulletin*, 102: 17-19.

Lebedev, Alexander. (1997). *Best search engines for finding scientific information in the Web.* http://www.chem.msu.su/eng/comparison.html

Leighton, H. Vernon. (1995). *Performance of four World Wide Web (WWW) index services: Infoseek, Lycos, Webcrawler and WWWWorm.* http://winona.msus.edu/is- f/library-f/webind.htm

Leighton, H. Vernon. & Srivastava, Jaideep. (1997). *Precision among World Wide Web search services (search engines): Alta Vista, Excite, Hotbot, Infoseek, Lycos.* http://www.winona.msus.edu/is-f/library-f/webind2/webind2.htm

Liberatore, Karen. (1997). Rating the search engines: A guide to the info-gathering barons. *Macworld Online*. June 5. http://macworld.zdnet.com/ns-search/netsmart_collection/features/searchin.review.html

Machovec, George S. (1996). World Wide Web search engines: Open Text, Harvest, 2ASK. *Libraries and microcomputers*, 14(6-7): 1-6.

Marks, Tracey (1997). *Search the Net: Top ten search resources.* http://www.windweaver.com/searchtools.htm

Notess, Greg R. (1997). Comparing Net directories. *Database*, 29(1): 61-64

Peterson, Richard Einer. (1997). Eight Internet search engines compared. *First monday*, 2. http://www.firstmonday.dk/issues/issue2_2/peterson

Seong, Jeong-Chang. (1997). Survey of 4 search tools. http://www.cs.uga.edu/~seong/search.html

Sherman, Chris. (1998). Search engine help documentation and resources on the Web. *Online*, 22(6): 51-56.

Stanley, Tracey. (1996). *Alta Vista vs. Lycos*. http://www.ariadne.ac.uk/issue2/engines

Tomaiuolo, Nicholas G. & Packer, Joan G. (1996). Web search engines: Key to locating information for all users or only the cognoscenti? *International Online Information Meeting*. Oxford: Learned Information: 41-48.

Venditto, Gus. (1996). Search engine showdown. *Internet world*, May: 79-86.

Westera, Gillian. (1996). Search engine comparison: Testing and retrieval accuracy. http://www.curtin.edu.au/curtin/library/staffpages/gwpersonal/senginestudy/results.htm

Westera, Gillian. (1996). Search engine comparison: Exploration of changes in results over time. http://www.curtin.edu.au/curtin/library/staffpages/gwpersonal/senginestudy/results2.htm

Winship, Ian R. (1995). World-Wide Web searching tools: An evaluation. *Vine*, 99: 49-54.

Your complete guide to searching the Net. (1997) *PC Magazine Online*. December 2. http://www.zdnet.com/pcmag/features/websearch/suit.htm

Zorn, Peggy. Emanoil, Mary. Marshall, Lucy. & Panek, Mary. (1996). Advanced Web searching: Tricks of the trade. *Online*, May/June: 14-28.

REFERENCES

Cleverdon, Cyril. (1966). *Aslib Cranfield research project. Factors determining the performance of indexing systems. Volume 1. Design.* Cranfield: College of Aeronautics.

Lancaster, F. W. (1969). MEDLARS: Report on the evaluation of its operating efficiency. *American documentation.* 29(2): 119-142.

Saracevic, Tefko. et al. (1988a). A study of information seeking and retrieving. I. Background and methodology. *Journal of the American Society for Information Science*, 39(3): 161-176.

Saracevic, Tefko. et al. (1988b). A study of information seeking and retrieving. II. Users, questions, and effectiveness. *Journal of the American Society for Information*, 39(3): 177-196.

Saracevic, Tefko. et al. (1988c). A study of information seeking and retrieving. III. Searchers, searches, and overlap. *Journal of the American Society for Information*, 39(3): 197-216.

Tague, Jean. & Schultz, Ryan. (1989). Evaluation of the user interface in an information retrieval system: a model. *Information processing & management*, 25(4): 377-389.