A THEORY FOR THE MEASUREMENT OF INTERNET INFORMATION RETRIEVAL

Steven Leonard MacCall, B.A., M.S.

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

May 1999

APPROVED:

Ana D. Cleveland, Major Professor
Donald B. Cleveland, Committee Member
Amanda Spink, Committee Member
Jon I. Young, Committee Member
Roberta J. Beals, Committee Member
Philip M. Turner, Dean of the School of Library and Information Sciences
C. Neal Tate, Dean of the Robert B. Toulouse School of Graduate Studies

MacCall, Steven Leonard, A Theory for the Measurement of Internet Information Retrieval. Doctor of Philosophy (Information Science), May 1999, 131 pp., 53 tables, references, 53 titles.

The purpose of this study was to develop and evaluate a measurement model for Internet information retrieval strategy performance evaluation whose theoretical basis is a modification of the classical measurement model embodied in the Cranfield studies and their progeny. Though not the first, the Cranfield studies were the most influential of the early evaluation experiments. The general problem with this model was and continues to be the subjectivity of the concept of relevance. In cyberspace, information scientists are using quantitative measurement models for evaluating information retrieval performance that are based on the Cranfield model. This research modified this model by incorporating enduser relevance judgment rather than using objective relevance judgments, and by adopting a fundamental unit of measure developed for the cyberspace of Internet information retrieval rather than using recall and precision-type measures. The proposed measure, the Content-bearing Click (CBC) Ratio, was developed as a quantitative measure reflecting the performance of an Internet IR strategy. Since the hypertext "click" is common to many Internet IR strategies, it was chosen as the fundamental unit of measure rather than the "document." The CBC Ratio is a ratio of hypertext click counts that can be viewed as a false drop measure that determines the average number of irrelevant content-bearing clicks that an enduser check before retrieving relevant information. After measurement data were collected, they were used to evaluate the reliability of several methods for aggregating relevance judgments. After reliability coefficients were calculated, measurement model was used to compare web catalog and web database performance in an experimental setting.   Conclusions were the reached concerning the reliability of the proposed measurement model and its ability to measure Internet IR performance, as well as implications for clinical use of the Internet and for future research in Information Science.

ACKNOWLEDGMENTS

The successful completion of this dissertation was due to the time, talents, and patience of a wide range of individuals. First and foremost, my major advisor, Ana D. Cleveland, gave me the benefit of her expertise, generosity, and friendship through the highs and lows of the process. Mere words do not do justice in praising Dr. Cleveland, so I simply say “gracious.”

Special thanks also go to Donald B. Cleveland, whose contribution at key points during the development of this work was indispensable. Thanks also to my other committee members, Amanda Spink, Jon Young and Roberta Beals, for their time and talents.

I would also like to acknowledge the efforts of people who were crucial in the recruitment of subjects: Bud Prather, Suzanne Davis, Thomas McHattie, Marc Armstrong, and Jerry McKnight.

Finally, I would like to say a large and hearty thank you to my parents, whose understanding and support helped see me through to the a end of this project. It is my great honor to dedicate this work to them.

TABLE OF CONTENTS

Page

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Purpose of the Study

The purpose of this study is to develop and evaluate a measurement model for Internet information retrieval strategy performance evaluation the theoretical basis of which is a modification of the classical measurement model embodied in the Cranfield studies and their progeny.

Background

The goal of information retrieval (IR) systems is to retrieve *relevant* information. Since the first days of computerized systems, this goal has guided the development of measurement models that evaluate the performance of IR strategies. However, despite a large and varied body of critical information science research, the relevance-based approach to evaluation remains the dominant model as evidenced by the essential similarities between the first large scale evaluations, the 1960s Cranfield studies, and the most recent large scale evaluations, the 1990s Text Retrieval Conferences (TREC) studies. Though not the first, the Cranfield studies are the most influential of the early evaluation experiments and in this dissertation, the Cranfield studies and their progeny are referred to as the classical IR measurement model. The general problem with classical measurement model is the subjectivity of the concept of relevance.

The classical IR measurement model has been captured by Rees and Saracevic's (1966) four requirements for the quantification of the effectiveness of IR systems:

1. A criterion (or criteria) such as a phenomenon, value, quantity, measure, or dimension which can serve as the basis for a measuring unit. This criterion has to adequately represent one or more purposes of the system: relevance.
2. A measuring unit (measure) in terms of which the performance of the system is quantified: e.g., recall ratio and precision.
3. A measuring instrument (yardstick): relevance judgment independent of the system including the scales and instructions involved--in this connection it is to be noted that humans are acting as the analogs of measuring devices and are subject to variation.
4. A methodology for measuring: Cranfield approach.

This study proposes and evaluates a measurement model that is based on a modification of the general quantitative approach, embodied in the classical measurement model, for the information space of the Internet, what is commonly referred to as *cyberspace*.

In cyberspace, information retrieval tasks use the Internet information infrastructure to access resource collections. This study looks at a subset of Internet IR, namely the World Wide Web (WWW), which involves the use of *catalogs* and *databases*. Web catalogs are collections of WWW sites and WWW pages that are classified by human indexers. These web catalogs can be domain-independent collections, such as Yahoo! or domain-specific collections, such as Medical Matrix for medicine. Web databases are collections of WWW pages that are automatically indexed by algorithms and are generally searched by keyword to produce rank-ordered lists of documents. As is the case with web catalogs, web databases can be domain-independent collections, such as Alta Vista or can be domain-specific, such as Medical World Search for medicine. What is needed is a way to quantitatively compare the performance of each of these Internet IR approaches so that endusers and professional searchers are aided in their selection of the proper tool at the proper time. The proposed measurement model is designed to meet this need.

This dissertation is divided into two parts. After performance data were collected, they were first analyzed in a measurement study in which the proposed measurement model was evaluated for reliability. After reliability coefficients were calculated, the data were then used in a demonstration study in order to demonstrate the use of the proposed measurement model. The measurement study was a necessary precursor because the generated reliability coefficients enabled better interpretation of the results of the demonstration study.

Two Methodological Questions in Information Science Measurement Theory

Any developmental work in IR measurement theory in the late 1990's must consider the research findings of the past twenty years concerning relevance, specifically, those studies which conclude that the objective relevance of the classical approach must be modified in order to reflect the subjective/cognitive theoretical point of view. Additionally, current theoretical work in IR evaluation must also account for the new IR environment, namely, the cyberspace of Internet IR. Endusers now sit at personal computers that are connected to the world wide information infrastructure and are using retrieval tools in a distributed environment that are only now being evaluated in academic research. In cyberspace, Information Science currently relies on a measurement model for evaluating information retrieval performance that is based on the classical measurement model developed for a prior information retrieval environment. Research is needed to modify the classical measurement model in order to incorporate the dynamics of the Internet information retrieval environment and to incorporate relevance research findings.

This dissertation is concerned with analyzing two problem areas in Information Science measurement theory. The first is the traditional problem concerning the use of objective human relevance judgments as measuring instruments and the second is a new problem concerning the implication of the dynamic documents of cyberspace on classical IR measurement units, such as recall and precision.

Are Human Relevance Judgments Objective Measuring Instruments?

Ideally, a reliable scientific measurement model is based on a valid criterion that is used to construct an objective measuring instrument to facilitate quantitative comparison using a stable measuring unit. A classic example is the measurement of length using a meter stick as the measuring instrument. The unit of measure for a meter stick is the *meter*, whose stability is defined based on the valid criterion that the velocity of light is a universal constant and can be used to objectively establish the "meter" as the distance that light travels in 1/299,792,458 second (Kranz, Luce, Suppes, & Tversky, 1971). By defining the meter as a universal standard, the meter stick becomes an objective measuring instrument for calculating the length of spatially extended objects using *meter* as the measuring unit. As a result, scientists have a reliable methodology for comparing spatially extended objects: "A" (measuring five meters) is longer than "B" (measuring 4.5 meters). The meter stick is widely accepted, and there is little controversy over its objectivity as an instrument for measuring length. More importantly, the meter stick is considered legitimate in advance of its use, which enables scientists to use the meter stick in their research without having to be concerned with its legitimacy for measurement.

It has long been the goal in Information Science to discover a similarly reliable measurement model to facilitate the quantitative comparison of IR strategy performance. In the classical IR measurement model, the criterion of relevance is operationalized as the "human judgment of topical relevance assigned to a document." The "objective" human relevance judgment then has the role of the objective measuring instrument, and recall and precision are the two primary units of measure. This model uses human relevance judgments as the measuring instrument in IR evaluation by requiring that an information scientist determine in advance which documents are judged as objectively relevant to given topics (usually by a panel of subject experts) in order to evaluate IR system performance in retrieving relevant documents. Use of the classical measurement model allows quantitative comparisons of IR strategies:

- "Strategy A" scoring 50% recall (retrieving one half of all documents relevant to a given query) is better than "Strategy B" scoring 25% recall (retrieving one quarter of all documents relevant to the same given query).
- "Strategy C" scoring 75% precision (three quarters of all retrieved documents are relevant to a given query) is better than "Strategy D" scoring 50% precision (one half of all retrieved documents are relevant to the same given query).

In terms of measurement theory, using relevance as a criterion in this way implies that each determination of the relevance of a particular document to a given topic in advance of its retrieval by the IR strategy under study is objective and thus universally valid. This invariance assumption is made, as in the case of the meter stick, by the assignment of an unvarying attribute (relevance) to a specific object (a document).

The classical IR measurement model has been heavily criticized on methodological grounds. The particular complaint is with step three in Rees and Saracevic's four requirements for the quantification of IR effectiveness listed above, in which human relevance judgments are stipulated as objective when applied to topic-document relevance relationships. This complaint has led to the conclusion that *any* attempt to use objective human relevance judgments as a basis for an IR measuring instrument creates an unreliable measurement model. In a recent summation of this line of criticism, Stephen Harter concludes that "[there is no] valid interpretation of the meaning of the results of retrieval testing that are based on fixed, unchanging relevance judgments" (Harter 1996, p. 37) leading to the methodological conclusion that

> the nature and extent of the measurement errors introduced by using Cranfield instruments is essentially unknown. For all intents and purposes, this situation is unchanged from the early days of IR experimentation (Harter & Hert, p. 20).

A problem concerning IR measurement models thus arises when human relevance judgments are used as measuring instruments, especially when researchers make invariance assumptions about relevance relationships, i.e., when they stipulate human relevance judgments as fixed, objective relationships between documents and topics when creating a test collection in advance of experiments that measure IR performance. The primary criticism is this: The assumption of objective human relevance judgments, the criterion for binary or n-ary relevance decisions made in advance of IR evaluation experiments, interferes with the measurement of IR performance because in reality, actual human relevance judgments are subjective and can vary significantly. In terms of measurement theory, the “relevant document” cannot be part of a stable unit of measure, even when experimentally or statistically controlled, because using human relevance judgments as measuring instruments introduces a source of systematic error into the classical IR measurement model. To address this problem, research is needed to investigate the proposition that this measurement error can be reduced if variation in human relevance judgments is accounted for within IR measurement theory.

In summary, this study develops the Content-bearing Click (CBC) Ratio, a quantitative measure that addresses the relevance criticisms aimed at the classical measurement model by incorporating enduser, rather than objective, human relevance judgments. Before describing the CBC Ratio, the next section discusses a second problem with the classical measurement model.

What is the Fundamental Unit of Measurement in Information Science?

As the 20th century closes, Information Science must confront a second methodological problem about the classical IR measurement model, namely, what is the unit of measure to be? The advent of the cyberspace of Internet IR has led to the application and/or adaptation of the classical measurement model to IR evaluation of strategies that retrieve Internet-based documents from the WWW or from digital libraries. It is not clear that the classical measurement model, based on measurement units such as recall and precision that were designed for retrieving documents as traditionally construed, will remain valid in an era of potentially dynamic documents. Arguably, one can now distinguish between a static document, in which its content remains the same over time, and a dynamic document, in which its content may change over time. Examples of static documents would include print books, print journal articles, bibliographic records of books and journal articles, and the results of individual database searches. Examples of dynamic documents include electronic books, electronic journals, WWW pages and digital library documents.

From the perspective of IR measurement theory, immediate concerns can be raised about how to count dynamic documents. Before the advent of the cyberspace of Internet IR, this question was straightforward. Documents that were counted were those having their informational content on physical media: paper (books, journals, and articles) or other media (e.g., microfilm). With digitized information, dynamic documents, such as WWW pages or the content pages from digital libraries, can no longer be assumed as remaining the same document over time.

To further analyze this problem, a distinction between derived and fundamental measuring units is helpful. In measurement theory, a derived unit is defined in terms of fundamental units, which, in turn, are defined by independent arbitrary standards. Referring back to the example above, the meter would be a fundamental unit of measure because its length is arbitrarily defined as the distance in which light travels in 1/299,792,458 second. An example of a derived measure is *density* because its measurement requires the prior measurement of mass and volume. In classical IR evaluation, precision and recall are derived measurement units because they are defined in terms of a ratio of *document counts*, which, in turn, are arbitrarily defined as fixed, unchanging entities. From this perspective, static documents are the fundamental measuring unit of the classical IR measurement model. However in cyberspace, documents are potentially dynamic because they are in digital formats. Even more critical is the fact that Internet-based documents are distributed. In a library setting, scholarly papers and books will not change content after their arrival on the premises. However, in a distributed digital environment, document content might be updated at any time. Clearly, rethinking the unit of measure used in IR evaluation is warranted because we can no longer rely on documents remaining static over time as is the assumption of the units of measure associated with the classical measurement model. The CBC Ratio substitutes a fundamental unit of measure (the CBC), designed to account for the reality of dynamic Internet documents. The CBC Ratio is a modification of the classical IR measurement model because of its use of human relevance judgments made on the CBC, rather than on static documents.

In summary, the implication of the dynamic nature of cyberspace implies that IR strategies cannot be evaluated effectively with recall and precision as measuring units as traditionally conceived because of concerns about the reliability of human relevance judgments and because they are derived from static documents as the fundamental unit of measure. This study proposes a modification of the classical IR measurement model that incorporates a view of relevance and a fundamental unit of measure developed for the cyberspace of Internet IR.

Problem Statement

In cyberspace, information scientists are using a measurement model for evaluating information retrieval performance that is based on the classical model developed for a prior information retrieval environment. Research is needed to modify the classical model in order to incorporate the dynamics of the Internet IR environment and to account for relevance research findings from the subjective/cognitive point of view.

Significance of Study

In primary care medicine, the domain of study for this dissertation, research is needed to help determine the most effective way to maximize the effective transfer of information from knowledge sources to clinical decision-makers in the health community. The flow of information in medicine has increased dramatically due to recent technological advances in local, national, and international communication networks. The rapidly evolving information infrastructure, primarily the Internet, offers numerous possibilities for delivering medical information electronically in ways all members of the community find accessible. With the ever growing cost of health care, it becomes important for medical informatics and information science researchers to determine ways to maximize the effectiveness of information networks because there is general acknowledgment that the effective transfer of medical information is vital for ensuring high medical standards, minimizing costs, and increasing productivity. There is currently a need to establish reliable measurement models in order to determine which information retrieval strategies are most effective in delivering relevant information to endusers in general and family medicine physicians in particular.

However, the Internet has evolved so quickly that there is little research concerning retrieval evaluation so that most Internet users do not know if they are using the most efficient and effective retrieval strategies. More fundamentally, there is little basic research attempting to describe the nature of Internet IR and its possible theoretical relationship to established IR strategies. It is not known whether basic research applicable in the traditional IR domain will also hold for Internet IR. For example, much of current Internet IR research is focused on a database retrieval model in which "search engines" with retrieval algorithms attempt to retrieve relevant information from enormous databases of WWW pages or e-mail archives. Internet IR research concerning search engine retrieval performance borrows from evaluation methodologies used for bibliographic databases. However, it is not known whether this extension of IR evaluation to Internet IR is valid.

Additionally, there are monothetic/hierarchical classification schemes currently in use to manually organize Internet-based catalogs of information resources. Many such approaches rely on classification and indexing theories that were developed for information resources on physical media, such as the Dewey Decimal Classification (DDC) for books or the Medical Subject Heading (MeSH) for bibliographic document surrogates. Internet IR research concerning the evaluation of how well these knowledge structures facilitate the retrieval of relevant information is rare, and it is therefore unknown whether this extension of classification and indexing theories to Internet-based information resources is valid.

Finally, it is not known whether relevance-based models of information behavior developed in research from the user perspective and based on observations made in traditional IR settings can be validly extended to the Internet IR environment.

Definitions

Monothetic/hierarchical Classification: A general strategy for organizing information that uses a single, fixed structure to control the subject-oriented description of resources, generally organizing to the most specific subject heading.

Web Catalog: Collections of WWW resources, such as Yahoo (http://www.yahoo.com) and Medical Matrix (http://www.medmatrix.org), that use a monothetic/hierarchical classificatory approach to manually organize their content.

Automatic Indexing: A general strategy for organizing information that uses computational algorithms called search engines to describe the intellectual content of documents, usually based on the number of term occurrences and position of terms in a document.

Web Database: Collections of WWW resources, such as Medical World Search (http://www.mwsearch.com) and Alta Vista (http://www.altavista.digital.com), that use a computational indexing strategy to automatically organize their content.

Information Retrieval (IR) Strategy: For this study, an IR strategy is one of four search tools assigned to judges, specifically, two manually classified web catalogs (Medical Matrix and Yahoo) and automatically indexed web databases (Medical World Search and Alta Vista).

Content-bearing Click (CBC): Any hypertext click that is used to retrieve possibly relevant information as opposed to a hypertext click that is used for other reasons, such as the "search" click that begins a database search or a "navigation" click that is used to traverse a WWW-based information resource.

Research Questions

This study addresses the following five research questions and hypotheses. Research questions 1 and 2 deal with the measurement study and research question 3, 4, and 5 deal with the demonstration study.

Measurement Study

R1. Does a modified classical IR measurement model translate into a reliable methodology for measuring the performance of IR strategies in the Internet environment?

H1. A measurement model is reliable if its reliability coefficient exceeds the standard for a discipline:

$\alpha(H_1) > .70$

R2. What is the most reliable procedure for aggregating human relevance judgments in the Internet IR environment?

H2: Human relevance judgments are more reliable measuring instruments when aggregated based on the subjective perception of cognitive difficulty of questions rather than when aggregated based on the objective cognitive difficulty of questions rather than based on the cognitive class of individual judges and rather than over all judges regardless of cognitive class:

α(aggregated across question classified by subjective cognitive difficulty) >

α(aggregated across question classified by objective cognitive difficulty) >

α(aggregated across judges by cognitive class) >

α(aggregated across judges regardless of cognitive class)

Demonstration Study

R3. In terms of Internet IR performance, which web catalog employs the more effective strategy for the manual monothetic/hierarchical classification of WWW resource collections?

H3: In Internet IR tasks, a domain-specific web catalog performs significantly better than a domain-independent web catalog:

CBC Ratio(Medical Matrix) > CBC Ratio(Yahoo!)

R4. In terms of Internet IR performance, which web database employs the more effective strategy for automatically indexing WWW resource collections?

H4: In Internet IR tasks, a domain-specific web database performs significantly better than a domain-independent web database:

CBC Ratio(Medical World Search) > CBC Ratio(Alta Vista)

R5. In terms of Internet IR performance, is web catalog more effective than a web database?

H5: In Internet IR tasks, a web catalog performs significantly better than a web database:

CBC Ratio(Medical Matrix) > CBC Ratio(Medical World Search)

Limitations of Study

There are several limitations that arise from the nature of cyberspace and Internet IR as viewed in this dissertation:

1. The current study treats the Internet as a large information retrieval system for answering questions that arise in the practice of primary care medicine and does not consider the Internet as a browsing tool. Therefore, no generalizations from this work can be made concerning the application of the proposed measurement model of Internet IR strategies other than from the perspective of problem-driven information retrieval.

2. Primary care physicians are one population from which the human relevance judges are taken for this study. Before any generalizations can be made outside of this group, future work will be required to verify results obtained.
3. Primary care physicians-in-residency-training are one population from which the human relevance judges are taken for this study. Before any generalizations can be made outside of this group, future work will be required to verify results obtained.
4. Medical students who are post Family Practice clerkship are one population from which the human relevance judges are taken for this study. Before any generalizations can be made outside of this group, future work will be required to verify results obtained.
5. There are aspects of Internet IR, including excessive network traffic and the availability of specific web sites, that may impact any attempt to study information retrieval in a controlled experimental research environment.

CHAPTER 2

LITERATURE REVIEW

Introduction

This chapter reviews selected Information Science evaluation research. The chapter is divided into two major sections. After the introduction, the next section discusses evaluation research that uses objective human relevance judgments in the classical measurement model. The final section discusses research from the subjective/cognitive perspective in Information Science.

One way to categorize relevance-related research in Information Science is to examine methodological approaches to the question of human relevance judgments. An important aspect of the question of relevance-based IR measurement concerns the aggregation of individual relevance judgments during an evaluation study, which is considered by Tague-Suttcliffe to be a major problem in IR evaluation (Tague-Suttcliffe, 1996). Stated differently, Harter and Hert (1997) commented that individual differences were the most general feature of IR evaluation study data and therefore important information concerning individual differences in relevance assessments remains hidden as a result of the pooling of individual searches that is required when computing general performance statistics such as recall and precision. From the subjective/cognitive perspective, this aggregation of human relevance judgments results in a loss of critical information concerning the relevance behavior of judges and conceals individual differences in relevance judgments that may very well indicate why a particular strategy performs well or poorly. In short, the objective perspective is based on the assumption that there is little or no variation in human relevance judgments, while the subjective/cognitive perspective explores these judgments in research about information behavior.

The methodological distinction between the objective and subjective/cognitive relevance perspectives may be sharpened in a measurement theory framework by using Reliability Theory (RT). In RT, any set of measurements has a total variance, i.e., a set of scores that attempt to reflect an underlying phenomenon, which can be broken down into two components: "true" score variance and error variance. The latter is commonly referred to as "measurement error." The two sources of variation are related according to the following:

$$V_t = V_\infty + V_e$$

where V_t is the total obtained variance, V_∞ is the true score variance, and V_e is the error variance (Kerlinger, 1992). Applying this framework to human relevance judgments, the central controversy of how to use relevance in IR measurement models becomes whether to consider variation in human relevance judgments as part of measurement error variance or true score variance during IR measurement.

Viewing the situation from this perspective is useful because it illustrates the different approaches that each research perspective must take to deal with variations in human relevance judgments. The objective relevance perspective, in which variation in human relevance judgments is not explicitly considered, treats this variation as random (V_e) and subsequently uses quantitative research designs that minimize the impact of random influences. The subjective/cognitive relevance perspective, in which variation in human relevance judgments is explicitly considered, treats this variation as a part of the true score variation (V_{∞}) which must be directly addressed in research design. From a reliability perspective, both perspectives have the same goal: to maximize the true score variance component (V_{∞}) and minimize the component reflecting variance due to measurement error (V_e) of the total obtained variance (V_t) during measurement.

However, neither perspective on human relevance judgments considers the effect of the potentially dynamic documents of the cyberspace environment of Internet IR. To do this, the concept of “relevant document” must be broken down into its component parts: “relevance” and “document” so that their symmetrical nature can be explored. A case can be made that the classical IR measurement model views the relationship between relevance and documents as symmetrical in that *static* human relevance judgments are mapped to *static* documents. A second case can then be made that the subjective/cognitive perspective views this relationship as asymmetrical in that *dynamic* human relevance judgments are mapped to *static* documents (thus creating problems for the classical IR measurement model). However, the third case must be examined: whether symmetry can be restored by mapping *dynamic* human relevance judgments to the *dynamic* documents by introducing a new fundamental unit of measure for the cyberspace of Internet IR.

The relationship between human relevance judgments and documents is summarized in Table 2.1.

After presenting a selection of research representing the classical IR measurement model in the next section, the literature review will turn to research from the subjective/cognitive point-of-view in the final section where criticism of the classical IR

Table 2.1

Symmetry and Asymmetry in the Relevance-Document Relationship

Relevance Perspective	View of relevance	What is counted
Classical IR Measurement Model	Static	Static documents
Subjective/cognitive Point-of-View	Dynamic	Static documents
Cyberspace of Internet IR	Dynamic	Dynamic documents

measurement model's use of static human relevance judgments will be explored. The specific critique is that static relevance judgments are an artifact of systems-oriented thinking that ignores the dynamic nature of relevance. However, the subjective/cognitive point-of-view can itself be subjected to criticism that it ignores the dynamic nature of cyberspace. This comment is based on the observation that the subjective/cognitive perspective position is based on a relevance-document asymmetry that is least *untested* in the cyberspace of the Internet IR environment and may very well itself an *unnecessary artifact* of the era of the batch retrieval of static documents.

These issues reduce to two basic questions that are addressed in the literature review:

1. How does research using the classical IR measurement model account for variation in human relevance judgments (if at all)?
2. How does research critical of the classical IR measurement model account for the potentially dynamic documents of the Internet IR environment (if at all)?

Objective Relevance: The Classical Information Retrieval Measurement Model

Information retrieval evaluation has been a concern of information scientists since the beginning of the field. As reported by Gull (1987), Mortimer Taube's concern that library catalogs in card and book form and printed indexing were not able to meet the demands of research-oriented users led to the development of coordinate indexing for computer-based texts. The growth in the amount of computer-based information in the 1950's also lead to a proliferation of means for indexing and abstracting this new type of information. It quickly became clear that evaluation methodologies were needed to determine which of many indexing methods were better. For example, it was known whether a single word approach, such as Taube's Uniterm approach, was better that multiple words or phrases.

Recall and Precision Studies

It is generally acknowledged that the first large-scale IR evaluation studies were Cranfield I (Cleverdon, 1962) and Cranfield II (Cleverdon, Mills, and Keen, 1967), which were run at the Cranfield Institute in England in the early 1960's. The Cranfield II study was especially noteworthy. The purpose of the Cranfield II study was to develop an experimental model in which IR performance could be measured in a laboratory-type setting free of the contamination of extraneous variables. For example, variation in human relevance judgments was experimentally controlled based on the uniform indexing used for the text collection of 1,400 topically related documents using an expert panel of judges who predetermined relevant documents in advance of the evaluation experiment. With this control achieved, the Cranfield II study was able to determine the best retrieval performance based on recall and precision measuring units. The experimenters were able to directly manipulate the independent variable. Cranfield II was a groundbreaking example of experimental research design in Information Science because a highly controlled setting was reached, thus paving the way for future studies of this type. In brief, the Cranfield II study used static human relevance judgments made in advance of the experiment and aggregated them across all judges. Its test collection was restricted to static, paper-based documents of aeronautical research.

Another significant recall and precision study was Lancaster's 1968 paper that evaluated the National Library of Medicine's MEDLARS system. According to Harter (1996), this study was a prime illustration of the potential variation in human relevance judgments and the resulting loss of information when these judgments were aggregated. Relevance judgments were treated as objective; however, as operationalized by Lancaster, they were shown not to be sufficiently stable for aggregation. Mean recall and precision values were 58% and 50% respectively, but as Harter pointed out, this aggregation of relevance judgments actually led to a loss of information about how the system performed at the level of the individual user. Other methodological concerns of the Lancaster study were that the human relevance judgments remained static throughout the study and the documents were static bibliographic records of static medical research documents.

Beginning in 1992, a major new IR evaluation initiative, the Text Retrieval Conferences (TREC), began as an annual meeting of IR researchers who wished to compare their retrieval systems and strategies (Harmon, 1992). In order to do this, an evaluation methodology was established as follows: First, a set of static documents was agreed to in advance as being relevant to a given set of queries so that each TREC researcher could use search the same pre-defined set of relevant documents in batch mode. Because TREC attempted to capture the interactivity inherent in the nature of digital libraries, it is considered by many to be an improved evaluation methodology over the original Cranfield II (Harter and Hert, 1997). However, there was no allowance made for dynamic documents, since the text collections were drawn from paper-based documents.

As the research reviewed in this section shows, recall and precision studies are for the most part not directly addressing the question of whether human relevance judgments are stable enough to serve as reliable measuring instruments. The next section discusses additional studies about the classical IR measurement models that address this problem more directly.

Studies Examining the Human Relevance Judgments of Subject Experts

This section examines the role that human relevance judgments play in the classical IR measurement model. In general, research in this area attempts to better understand the basis on which relevance judgments are made so that their variability as measuring instruments can be reduced. The purpose of examining human relevance judgments in this context was to determine the situations under which they are stable enough to provide a basis to be used as reliable measuring instruments. This is important because if this reliability could be demonstrated, then a major source of error variation was controllable in the experimental environment. The ultimate goal of the classical measurement IR measurement model was to be able to validly test the performance of IR strategies in an experimental setting prior to their use in the operational environment.

The specific criticism of human relevance judges during IR evaluation is that they cannot reliably predict the relevance of documents for actual endusers (Schamber, 1994). Salton (1992) counters the criticism with research indicating that the relevance judgments of the *critical documents* of a retrieved set, i.e., the first few, are indeed a stable basis on which recall and precision can be calculated. When concentrating on the first few documents, human relevance judges produce stable judgments. In Lesk and Salton (1968), expert human relevance judgments were shown to be stable for those documents that were retrieved early in the IR process. In other words, though variation across human relevance judges did exist in terms of the total number of relevant documents retrieved, the stability of recall and precision measures was not impacted when restricted to the first few documents. Their conclusion was

> that although there may be a considerable difference in the document sets termed relevant by different judges, there is in fact a considerable amount of agreement for those documents which appear most similar to the queries and which are retrieved early in the search process (Lesk & Salton, 1968, p. 355).

From this work, the general position is maintained that variation in human relevance judgments could be treated as a source of measurement error (V_e) and can therefore be controlled statistically within an experimental methodology.

In his review of research concerning variations in human relevance judgments, Harter (1996), while lauding Lesk and Salton’s work for probing the foundations of the classical IR measurement model, nevertheless suggested that they did not consider key elements of relevance as possible independent variables. For example, Harter pointed out that the Lesk and Salton study could have benefited from the knowledge that the difficulty level of an information request can predictably affect a relevance judgment. Also, that the order of presentation of the documents and the choice of relevance scale need to be accounted for. However, Harter’s goal was not to derail the classical IR measurement model, but rather to emphasize that concerns about variations in human relevance judgments must be at the center of any research effort in this area.

In the same paper, Harter gave an example of how this might be done by sketching a modification of the classical model that would account for variation in human relevance judgments during IR evaluation and thus render quantitative evaluation possible. In order to maintain both positions, that relevance judgments vary *and* that quantitative IR evaluation is possible, Harter attempted to show that *some* aspect of relevance can be specified as invariant. Harter conjectured that relevance judgments can be made to vary predictably according to a combination of three factors: "the user's experience as a researcher" (novice, moderately experienced and highly experienced), "the user's stage of research" (just beginning research, working on design methodology, and working on conceptualization of research problem), and "the user's previous knowledge of the specific problem literature" (no previous knowledge, some knowledge, and excellent knowledge). Harter combined these factors and divides human relevance judgments into categories, implying that two factors, "differentiated relevance judgments according to cognitive state" and "topical relevance assigned in advance to a document" could be combined to produce a stable, relevance-based measurement model.

The value of Harter's approach is that he provides an opening for viewing the possibility that valid IR measurement using human relevance judgments is possible by accounting for the cognitive class of enduser as a stabilizing factor. This approach leads to the possibility of unifying the objective and subjective/cognitive relevance positions. Harter's conjecture is that human relevance judgments are stable if properly categorized and only if aggregated over that category, which requires continued study into the nature of relevance judgments. Once an understanding is achieved in this context, then valid quantified evaluation is possible because the aggregation of human relevance judgments was over a logical subset of judges rather than across all judges regardless of cognitive characteristic. This conjecture, taken in conjunction with a suitable fundamental measuring unit for the cyberspace of Internet IR, provides the basis for the measurement model for evaluating Internet IR strategy performance proposed by this study

The Classical IR Measurement Model in Cyberspace

Research in Information Science concerning the evaluation of IR strategies in cyberspace is a growing area (Schwartz, 1998). This section looks at a sample of such research to determine how the classical IR measurement model is used, specifically, in terms of relevance measures and in terms of the fundamental unit of measure. Because this area is growing so quickly, a review of research efforts must necessarily extend to significant work that is published only on the WWW, and therefore not subject to formal peer review. This research is noted below.

One of the first attempts to apply the classical model to a distributed networked retrieval environment was Marchionini, Barlow, & Hill (1994). The rationale for this study was that there existed little quantitative or qualitative evidence comparing the performance of IR strategies in an interactive environment and therefore no theoretical guidance for determining the most effective strategies for use by endusers or intermediaries. Practice, rather than theory, was driving development. In their study, Marchionini, Barlow & Hill compared two types of IR systems, WAIS (Wide Area Information Server) and a Boolean-based retrieval system. Three years of the NASA Scientific and Technical Information (STI) database were loaded at NASA using a commercial WAIS server. The dependent variables were recall and precision, and relevance judgments were made by endusers on a five-point scale. However, these relevance judgments were static since there was no evidence of subjective/cognitive dynamics in the research design. Also, for purposes of counting, it was unclear as to what constituted a "document" in this study. Presumably, the entities that were counted were bibliographic records of the STI database.

More important than the specific results of this study are the implications they suggest for future measurement efforts in cyberspace. For example, the authors pointed out that a WAIS-type system retrieved ranked sets of documents rather than batched sets. This was a problem because precision ratios no longer had a definite number of documents, was the case for the denominator used in batch retrieval evaluation. Also, the WAIS system was interactive, which means that any metric developed for batch retrieval was suspect. Though no new metric was proposed, this exploratory study does provide a useful foundation on which to focus needed research. The authors concluded that variables for evaluating IR in the new environment were needed that were not based on individual document relevance.

In later research, Ding and Marchionini (1996) moved from the limited WAIS environment of Marchionini, Barlow, & Hill (1994) to the WWW, specifically, to the comparison of WWW search services or, better, web page databases (WBDBs). Three WPDBs were compared using the classical IR measurement model. The dependent variables were relevance-based, three of which were a variation on precision as classically conceived and the remaining two were new measures ("salience," an aggregate measure for each WPDB, and "relevance concentration," a measure designed to reflect the ratio of relevant hits in the first ten over relevant hits in the entire set of twenty). To address the problem of how to "batch" ranked retrieval output, all denominators were restricted to the first twenty documents retrieved.

Though not explicitly, this research tackled the problem of what to count in the cyberspace of Internet IR. For example, the term “relevant document” did not appear in the paper. However, there was ambiguity concerning what exactly is being counted. In the “Methods and Procedure” section of the paper, the term “hit” was used to represent the countable entity, but there was no definition of precisely what a “hit” is. Another problem concerned their treatment of hits that were orthogonal to the retrieved batch, i.e., the hits that are hypertext links within the documents of the retrieved set. Ideally, the countable entity of Internet IR should be based on the interactivity of the hypertext environment in that the counting of orthogonal hits is facilitated. Not to do so would keep intact a vestige of the batch retrieval environment, in that only those entities in the *retrieved set* are those that are counted.

Non-peer reviewed research posted on the WWW (Leighton & Srivastava, 1997) suggested new quantitative measures for the cyberspace of Internet IR. Leighton and Srivastava compared five WPDBs (Alta Vista, Excite, Hotbot, Infoseek, and Lycos) using several variations of the standard precision measure based on what they term “first twenty precision.” This metric was based on relevance judgments that are made by the researchers based on a scale of 0 to 3. This scale attempted to capture “relevance” (objective) and “usefulness” (subjective) on a linear scale. No attempt was made to provide for dynamic or changing relevance judgments; however, the authors acknowledged that there may be a problem with their treatment of relevance judgments. Their stated goal, though, was to treat the relevance question consistently.

Like Ding and Marchionini above, this study makes a contribution to the question of what to count in cyberspace. For example, they did not count “relevant documents” *per se*;

but concentrated on counting hypertext links to web pages. This is somewhat of an improvement, because the idea of a *page* in cyberspace is more dynamic than a *document* in print. Problems with this study included one that is shared with Ding and Marchionini's research above concerning their treatment of ranked retrieval output as a "batch" which does not allow orthogonal hits to be counted. A second problem concerned their unsupported postulation that twenty links is the limit to which the average user will pursue retrieved links as the result of a WPDB search. The WWW is a highly interactive environment, which raises questions of whether such a "futility" point is valid in the cyberspace of Internet IR.

In quantitative evaluation research, Bruce (1998) used a psychophysical magnitude estimation technique to measure user satisfaction with Internet IR sessions. The psychophysical approach, which has been used in Information Science to capture n-ary relevance judgments (Eisenberg, 1988), was adapted by Bruce to measure Internet search satisfaction. His results showed that this approach produces a reliable method to measuring enduser satisfaction in the Internet IR environment. Bruce's study is important, as it begins to show that quantitative measures that account for enduser judgments are possible when evaluating Internet IR. In addition, Bruce showed that Information Science theory developed in pre-Internet studies is an important source of guidance when examining the measurement problems of Internet IR. As with Bruce's research, the current study attempts to provide a quantitative measure for Internet IR that is based on theoretical work developed in pre-Internet Information Science evaluation research.

In summary, this sample of research indicates that those researchers who adapt the classical IR measurement model to cyberspace continue to ignore variations in relevance judgments. Additionally, the suggested units of measurement are based on the number of web pages or hypertext links retrieved rather than at a more fundamental level in which orthogonal links are counted.

Subjective Relevance: The Subjective/Cognitive Perspective

Criticism of IR evaluation concerning the use of objective relevance as a criterion for measuring how well IR systems perform originates in research that is user-centered. The basic finding of this very large body of research is that there is empirical evidence showing variation in human relevance judgments across individuals or groups of individuals that must be taken into account by IR measurement models. Researchers in this area argue that the stipulation of fixed relevance judgments in advance of IR experiments ignores the individual differences inherent in the phenomenon of relevance. It is their position that a disregard of variation in human relevance judgments is a crucial reliability issue that plagues the classical IR measurement model.

The subjective/cognitive perspective opens the “black box” of relevance (Cuadra & Katter, 1967) in order to determine what factors have to be accounted for when using relevance as a criterion for IR evaluation. Due to the fact that they examine posited relationships and attempt to model entities “in the mind,” the subjective/cognitive perspective has yet to produce a single approach to evaluating information retrieval that is acceptable to all researchers. The subjective/cognitive perspective therefore cannot be evaluated in the same manner as the classical IR measurement model was in the last section. The reason for this is the classical model has an archetypal methodology for measuring call the Cranfield approach (Ellis, 1994), which makes it both a well-established methodology to apply and makes it a convenient target for critics. Partially in response to the lack of an archetypal methodology, research from the subjective/cognitive perspective reviewed in this section uses qualitative designs rather than quantitative ones.

Problems with Objective Human Relevance Assessments

Schamber, Eisenberg, and Nilan (1990) noted that without an understanding of what relevance means *to* users, it seemed difficult to imagine how a system could retrieve relevant information *for* users. This was the rallying cry for those who view the objective relevance approach as misguided. However, there is not yet a stable vocabulary to describe the subjective/cognitive perspective. Researchers have used a variety of terms to describe their work, including situational relevance, subjective relevance, pertinence, and utility. Each specific approach shared a concern with the validity of objective human relevance judgments, and as an alternative, they sought to study the criteria that users employ when judging relevance. In their view, a major problem with objective relevance was that it tends to take a “snapshot” of the IR situation, rather than considering IR as a dynamic interaction between user and information store. In short, information seeking was better viewed as a process. Endusers’ problems emerged when they were forced to structure their information needs to fit the constructs imposed by an IR system. There was a tendency for objective relevance researchers to think that if a system were built rationally, then users would be effective. From the subjective/cognitive perspective, this philosophy inevitably lead to problems because much of what was considered relevant was situation-dependent according to the educational levels, culture, and attitudes of users.

Dervin and Nilan (1986), in an influential chapter in the Annual Review of Information Science and Technology, reviewed information needs and uses literature by analyzing issues relating to the conceptualizations that had been driving this research. What they discovered was a tension between theory and practice: What was produced in terms of information systems design and theory was not meeting the practical needs of both the people who used these systems and the professionals whose responsibility it was to assist those who use these systems. Dervin and Nilan concluded that there was a dearth of research and theory available to guide practitioners from any perspective other than from a systems-oriented one, which was ineffective because it concentrated on dependent variables that were easily counted, such as book circulation and other system use statistics. This focus on system-oriented measures and theories lead to what Dervin and Nilan considered to be a vicious circle, in which systems-oriented research generated results that reified the systems approach. In order to have information systems that provided utility to users, designers needed to be aware of the needs of those users, which meant that user needs should become a central focus of system design.

Typical systems-oriented research questions sought to determine the extent to which users used given system functions, the extent that users perceived barriers to the use of the system, and the extent that users reported satisfaction with the function of systems. Based on these questions, researchers would attempt to explain differences among individuals according to behavioral dimensions with predictors such as demographics, sociological, life style, and task description. Dervin and Nilan pointed out that these were laudable goals, but they are premised on an assumption that there was a problem with the user/system interface, rather than a problem with the system's conception of the user's information need. In fact, "information need" and "information use" often remained undefined in this research.

Examples of systems-oriented research include the use of circulation statistics and citation counts. If a library only has a count of how many times a book was checked out over a given time period, then researchers were limited in measuring the effective contact that a library patron has had with the book. Concerning citation counts, if a researcher only has a count of citations over a period of time, then analysis concerning the nature of the citations, such as a negative cite, was missing and not available for inclusion into a conclusion about citing behavior. In short, one cannot know just from citation counts what types of citing practices are taking place (MacRoberts & MacRoberts 1989).

The next subsections discuss two theoretical approaches to the subjective/cognitive perspective.

Subjective/Cognitive Theoretical Approach: Sense-Making

Based on research stretching over twenty years (Dervin, 1997; Dervin, 1994; Dervin, 1992; Dervin & Dewdney, 1986; Dervin, 1977), Brenda Dervin concluded that many of the questions concerning the effectiveness of information technology, design, and practice involved human actors. Therefore, questions that should be investigated included how to design databases so they would be maximally used, how satisfied are users, and why some potential users refrain from using information technology. How can the flexibility that new technologies allow be capitalized rather than merely using them do what is currently done only in greater quantities, from further distances and at faster speeds (Dervin, 1992)?

Dervin's fundamental unit of analysis was the discontinuity or gap, which is a person's perception that information is needed to solve a problem. Dervin's sense-making approach dictated that an information system, be it technological or human, consider the situational and cognitive context of the individual as an individual person. This was in contrast to the system's view, in which enduser information needs were considered in some aggregate form. For example, Dervin and Dewdney (1986) suggested that this could be done at the library reference desk through the use of neutral questions to elicit the concerns and capabilities of library patrons rather than by classifying patron information needs only according to demographic or socioeconomic considerations.

Discontinuities appeared in the relationship between reality and human sensors, between human sensors and the mind, between tongue and message created, between human and culture, between human at time one and human at time two, between human and institution, and between institution and institution. The research methodology that Dervin suggested for data collection is the micro-moment time line interview in which discontinuities were reconstructed for the researcher in the context of an information seeking timeline.

The inherent complexity of this approach has been a cause for concern. However, as stated above, the main reaction by the subjective/cognitive perspective to the perceived limits of the objective relevance position has been to produce dynamic models of information seeking and use.

Subjective/Cognitive Theoretical Approach: Cognitive Modeling

A second approach to opening the "black box" of relevance has been those who seek to apply cognitive science approaches to modeling user behavior. Neill (1992) adopted the World 3 of philosopher Karl Popper as a characterization of the "work space" in which the cognitive approach models user behavior. Popper's World 3 was based on a three world ontology. World 1 was the physical "objective" world in which things such as “books” or “documents” could be objectively counted because there was no disagreement as to their “thingness”. World 2 was the “subjective” world of private thoughts that are only accessible to individuals. World 3 was the space in which the subjective thoughts of World 2 are turned into the physical objects of World 1. According to Neill, it was the task of Information Science to investigate the properties of World 3.

Using the cognitive approach, researchers attempted to model information seeking and used mental constructs inferred from patterns of behavior. The goal was to determine how individuals and groups of individual structured information in their minds so that information systems could be built to reflect that structure rather than a generic rational structure.

Cognitive modeling could be viewed in four steps:

1. Attempt to conceptualize and define the terms used starting from the user and working out. For instance a definition of “information”.
2. Transfer of the center of interest from the information system to the user.
3. Transfer of interest from the observable behavior of the user (i.e., how much he/she used the information tools available) to the unobservable cognitive behavior of the user.
4. Shift to the premise that information is a subjective phenomenon, constructed at least to some extent by the user, and not objective.

B.C. Brookes was an early exponent of the cognitive approach (Brookes, 1980a). Brookes’ conceptual contribution was the fundamental equation of Information Science:

$$K[S] + \; I = K[S + \; S]$$

where *K[S]* was a knowledge structure, *I* represented information, and *K[S + S]* represented a modification of a knowledge structure *K[S]* after the introduction of information *I*. A knowledge structure was the cognitive scheme that a person had for possible actions. For example, when one entered a restaurant, the order of activities was clear in one's mind because there was a "logic" to it:

1. One was seated
2. One got a menu
3. One ordered
4. One ate.

In Brookes' equation, the *K[S]* represented the menu script. If one were to enter a restaurant and successfully "run" the script, then no information about the concept of "restaurant" is learned. However, should the rules of the script be violated, then the knowledge structure changes. For example, if one were to go to a parent's home, one would quickly find out that the standard restaurant script did not hold when after being informed (*I*) that no there is no menu from which to order. This information causes a modification of the restaurant script (*K[S + S]*) to produce a new knowledge structure that a parent's home is not a restaurant.

The definition of *information* in this context focused on the effect that information had on a cognitive structure rather than focusing on information in a abstract way. Various researchers in the cognitive area have defined information in this manner, including information was that which was capable of transforming an image structure

(Belkin, 1978) and information was any stimulus that altered the cognitive structure of the receiver (Paisley, 1980).

Building on these ideas, Ingwersen (1996) attempted to evolve information retrieval from a text basis to retrieval based on context or scripts. This meant that the following must be represented in information systems:

1. the topical information need
2. the underlying problem space
3. the actual work task or interest
4. the dominant work domain(s)

Ingwersen's observed that the context surrounding an information need could not be captured entirely at the textual level, which is the level at which most information retrieval systems operate ("text" retrieval). Therefore, there was a need to develop dynamic and highly interactive information systems.

The primary research methodology of the cognitive approach is protocol analysis in which information system users "pour out" what's inside their heads as they proceed during a search session. The goal of the research is to seek and model those mental processes that remain stable during the dynamic process of information seeking. This type of analysis looks for the common structures of information needs and their resolution in order to make systems smarter by modeling the dynamic nature of enduser interactions with information systems.

The Subjective/Cognitive Perspective in Cyberspace

There is research from the subjective/cognitive perspective concerning the evaluation of information retrieval in cyberspace. Some of this research, for example

Poulter 1997, is premised on the realization that in principle, techniques originating from the bibliographic IR environment are unsuitable for cyberspace. Other research, for example the 1995 Allerton Institute “How We Do User-Centered Design and Evaluation of Digital Libraries: A Methodological Forum,” is centered more firmly in the subjective/cognitive perspective.

Poulter 1997 pointed out that because the WWW is a stateless environment, compactions arise for traditional evaluation of systems. Because a connection is only made to a web search service when information was being transferred, a searcher can not refer back to earlier retrieved sets, as is the case traditional batch retrieval sessions. Another concern is the lack of a consistent in document structure on the WWW, especially when compared to bibliographic databases.

More formal efforts exist to evaluate information retrieval in cyberspace from the subjective/cognitive perspective. As noted above concerning the Allerton Institute, much of the reported research evaluated the digital library environment and was concerned with the social aspects of digital library use. Other areas of evaluation ranged from the general perspective of computer-mediated communication systems (Hilz & Johnson, 1989) to efforts in specific aspects such as electronic journal delivery (Rowland, McKnight, & Meadows, 1995). Representative research from this perspective was reviewed in Lamb 1995. In this work, Lamb classified usability research according to whether it was concerned with human computer interaction (HCI), content usability, organizational usability, or interorganizational usability. Across these categories were various perspectives for interpretation: rational, human relations, institutional, and postmodern, which placed the quantitative (rational) approach within a continuum of research that also included qualitative approaches (human relations, institutional, and postmodern).

Representative empirical work from the subjective/cognitive perspective was Covi and Kling (1996) who used a naturalistic study of the use of digital libraries by academics in molecular biology and literary theory. Their goal was to examine conditions that facilitated effective use. They defended their naturalistic approach by criticizing quantitative analyses in general for not being able account for hidden aspects of utilization lost in the attempt to compile statistical evidence of use. For example, they found that research-active faculty used digital libraries in support of publication (as an open natural system model), while librarians and computer specialists focused on digital libraries as part of their information infrastructure (as a closed rational system model). This research is valuable in that it reveals that general models of digital library models can interfere with each other. This contrast was helpful when studying the effectiveness of digital library use because, according to Covi and Kling, each point of view had external influences and embodied values that could be at cross purposes if they remained beneath the surface.

Van House, Butler, Ogle & Schiff (1996) pointed out that digital library research centered on usability assessment and interface design was too narrow for evaluating something as complex as a digital library. They concluded that because digital libraries support higher-order cognitive work, their evaluation must be in terms of how it impacted users' work. Following a similar line of thinking, Karamuftuoglu (1998) pointed out that system evaluation in the era of networked information systems must account for their collaborative nature and the necessary social informatics factors that could impact successful use of systems.

As the research reviewed in the section shows, theoretical and evaluation studies from the cognitive/subjective point of view continue to impact Information Science. The current level of research activity is indicative of optimism that evaluation models can be developed that aid system designers, system trainers, and endusers when using the Internet to retrieve information.

CHAPTER 3

METHODOLOGY

Introduction

Before comparative evaluation can take place in information retrieval (IR) research, reliable measurement models must be available. The question is whether a model for measuring IR performance that is based on a modification of the classical model can be developed that accounts for variations in human relevance judgments and meets the need for a new fundamental unit of measure. The result would be a modified classical model that enables IR researchers to determine whether one IR strategy is performing better than another in retrieving relevant information in the Internet environment.

In their recent medical informatics evaluation textbook, Friedman and Wyatt (1997) detail a two-step methodology for developing and evaluating measurement models. In brief, performance data were first used to evaluate reliability in a *measurement study*. After this was accomplished, the same data were used to demonstrate how the measurement model is applied in a *demonstration study*.

It is important to note that both the classical IR measurement model and the proposed modification produces a model that measures the performance of information retrieval *strategies* rather than the performance of *individual users*.

The purpose of this chapter is to describe the methodology for the study. The general goal of the research was to replicate the clinical decision-making environment as closely as possible. As noted below, standardized multiple-choice board review questions were used to simulate clinical cases. This meant that subjects were not dealing with actual patients nor were they able to access to diagnostic tests and lab results, which limited the amount of information available for clinical decision-making. Given the constraints imposed by the use of standardized test questions, this study sought only to determine whether *relevant information* could be retrieved by endusers supporting either the ruling in or ruling out of a suspected diagnoses. In short, the simulated clinical decision-making environment of this study represented a brief snapshot of a potentially much larger physician dialog with decision support tools of all types. It was only tested whether Internet IR strategies could be added to the existing clinical armamentarium of useful diagnostic tools.

This study addresses the following five research questions and hypotheses. Research questions 1 and 2 deal with the measurement study and research question 3, 4, and 5 deal with the demonstration study.

R1. Does a modified classical IR measurement model translate into a reliable methodology for measuring the performance of IR strategies in the Internet environment?

H1. A measuring model is reliable if its reliability coefficient exceeds the standard for a discipline:

$\alpha(H_1) > .70$

R2. What is the most reliable procedure for aggregating human relevance judgments in the Internet IR environment?

H2: Human relevance judgments are more reliable measuring instruments when aggregated based on the subjective perception of cognitive difficulty of questions rather than when aggregated based on the objective cognitive difficulty of questions rather than based on the cognitive class of individual judges and rather than over all judges regardless of cognitive class:

α(aggregated across question classified by subjective cognitive difficulty) >

α(aggregated across question classified by objective cognitive difficulty) >

α(aggregated across judges by cognitive class) >

α(aggregated across judges regardless of cognitive class)

R3. In terms of Internet IR performance, which web catalog employs the more effective strategy for the manual monothetic/hierarchical classification of WWW resource collections?

H3: In Internet IR tasks, a domain-specific web catalog performs significantly better than a domain-independent web catalog:

CBC Ratio(Medical Matrix) > CBC Ratio(Yahoo!)

R4. In terms of Internet IR performance, which web database employs the more effective strategy for automatically indexing WWW resource collections?

H4: In Internet IR tasks, a domain-specific web database performs significantly better than a domain-independent web database:

CBC Ratio(Medical World Search) > CBC Ratio(Alta Vista)

R5. In terms of Internet IR performance, is web catalog more effective than a web database?

H5: In Internet IR tasks, a web catalog performs significantly better than a web database:

CBC Ratio(Medical Matrix) > CBC Ratio(Medical World Search)

This chapter is divided into three sections. The first section describes data collection. The second section describes the procedures for the measurement study (research questions 1 and 2), and the third section describes the procedures for the demonstration study (research question 3, 4, and 5).

Data Collection

This section describes the general procedures for interacting with subjects who are serving as human relevance judges for the proposed measurement model. These procedures produced the data for both the measurement study and the subsequent demonstration study of the dissertation.

Method for Selecting and Classifying Human Relevance Judges

The human relevance judges used for this study were selected from the primary care medical community**,** specifically family practice. In order to evaluate the reliability of aggregation procedures, there had to be cognitive variation in the sample of human relevance judges. Potential human relevance judges were identified for each of the three cognitive classes, low, medium and high. A total of 36 judges were selected as presented in Table 3.1.

Table 3.1

Cognitive Classes of Human Relevance Judges

Low cognitive capability	Medium cognitive capability	High cognitive capability
12 3rd or 4th year medical students	12 Family Practice Residents	12 Family Practice clinicians

Training of Human Relevance Judges

For this research, judges were not stratified according to level of computer literacy. Rather, those judges with extensive computer and/or Internet experience were not used. To control for computer literacy and Internet experience effects, each potential human relevance judge were given a short pretest that identified approximate competency level. Judgment of computer/Internet literacy was based on a short questionnaire (Appendix A).

Instrumentation

In clinical medicine, there are recognized publishers of board review questions, which are collections of clinically relevant, multiple choice questions designed to test the knowledge of physicians. For this study, questions were taken from published board review questions in Family Medicine (Core Content Review of Family Medicine Executive Committee 1998). These questions were used to test the performance of Internet IR strategies. A Family Medicine physician competent in Internet IR was included on the dissertation committee in order to advise the investigator concerning the validity of the selected questions for use in an Internet IR session. Six questions were randomly selected as follows. Under the guidance of the physician, thirty questions were selected from the Core Content Review of Family Medicine collection and evaluated for their usability as a topic for an Internet IR session. Each of thirty identified potential questions was assigned a unique identifier and six of these identifiers were drawn from a hat to serve as the question set.

Data Collection Procedures

The following data collection procedures were used:

1. Each identified judge was randomly assigned to one of four Internet IR strategies until the total number of judges was reached. This was accomplished by assigning each judge a unique identifier and by placing these identifiers into a hat so that they could be drawn and assigned to a category. The assignment to an IR strategy group took place in numerical sequence according to the number associated with each IR strategy: IR strategy 1 – use of Medical Matrix; IR strategy 2 – use of Yahoo; IR strategy 3 – use of Medical World Search; IR strategy 4 – use of Alta Vista. For example, after three medical student judges were assigned to Medical Matrix, the next medical student was assigned to Yahoo, and so forth.
2. Prior to beginning the Internet IR session, each judge was given an "Instructions for Participants" sheet (Appendix B).
3. Each judge read the same set of six randomly selected multiple choice questions (Appendix C). To control for bias possibly present in the ordering of the question set, a Latin Rectangle design was used to assign specific question sequence for each judge. Question sequences were assigned as shown in Table 3.2 by working across each cognitive category (medical students, family practice residents, and family

practice clinicians), with each successive judge starting with the question subsequent to the prior judge's starting question.

Table 3.2

Table of Human Relevance Judge Internet IR (IIR) Strategy Assignment

	IIR Strategy 1	IIR Strategy 2	IIR Strategy 3	IIR Strategy 4
	Medical Matrix	Yahoo	MedWorld Search	Alta Vista
Medical Students	3 judges	3 judges	3 judges	3 judges
FP Residents	3 judges	3 judges	3 judges	3 judges
FP Clinicans	3 judges	3 judges	3 judges	3 judges

NOTE: IIR = Internet information retrieval; FP = Family Practice

4. After reading each multiple-choice question, each judge answered the question in the space provided on the corresponding answer sheet.
5. Before proceeding to the search phase, each judge used a psychometric scaling procedure to place a hashmark at the point on a line rating the subjective cognitive difficulty of each question:

|________________________________|

I am confident of my answer   I do not know the answer

6. Subjective level of cognitive difficulty for each question was determined by measuring the distance between the left most point of the psychometric scale and the hashmark placed by the human relevance judge. The responses were divided so that *i* = 72 for each level of difficulty. Table 3.3, shows the Question Difficulty Rating, with A_{ji} representing the difficulty rating of the i^{th} question by the j^{th} judge.

Table 3.3

Judges' Responses by Question Difficulty Rating

Question Difficulty Rating Category	**Judges' Responses**
1 (low difficulty)	A_{ij} , A_{ij} , A_{ij} , A_{ij} , A_{ij} , A_{ij} , ... A_{72}
2 (medium difficulty)	A_{ij} , A_{ij} , A_{ij} , A_{ij} , A_{ij} , A_{ij} , ... A_{72}
3 (high difficulty)	A_{ij} , A_{ij} , A_{ij} , A_{ij} , A_{ij} , A_{ij} , ... A_{72}

NOTE: A = Answer; i = question number; j = judge number.

7. Each judge was then given up to five minutes to rule in or rule out their given answer by using his or her assigned Internet IR strategy. Judges were permitted to use any specific keyword strategy when searching for information.
8. Judges were instructed to follow all clicks that were thought to be relevant and were told not to base their relevance judgments solely on the abstracted information returned by their assigned Internet IR strategy.
9. All computers used for this study had Pentium 166 Megaherz microprocessors with 16 megabytes of RAM and fast ethernet connections to the Internet. All judges used Netscape 4.5 browsing software. The "home" button of the browsing software was programmed to automatically return to the assigned Internet IR strategy, and each judge was instructed to click the "home" button should he or she want to return quickly to a blank search screen.
10. After the judge had read each question, the investigator tracked the time on task for each Internet search with a stopwatch starting with the first keystroke of the search and ending when the judge announced that information has been located that ruled in

the answer of the given question. If answers were not found within five minutes, the judges were directed to stop searching and proceed to the next question.

11. Judges were told to alert the investigator if information were found that ruled out the original marked answer. The time taken to retrieve this information was recorded as time on task.
12. The investigator counted all "false positives," i.e., those content-bearing clicks (CBC) that did not retrieve relevant information. (The method for calculating CBC ratios is described in more detail in the next subsection.)
13. The questions in each set were read, answered and searched for in sequential order. Each question was completed prior to reading the following question.
14. To control for Internet congestion, each search was run weekdays before 10:00 a.m. or after 7:00 p.m. central standard time or on a Saturday or a Sunday.
15. Three teaching physicians with five or more years practice experience who were not part of the study sample were asked to reach a consensus judgment of the objective cognitive difficulty of each of the six selected questions: low (1), medium (2), and high (3) as shown in Table 3.4.

Table 3.4

Consensus Expert Opinion Concerning Objective Difficulty Rating for Each Question

Question Difficulty Rating 1 (Low Difficulty)	Question Difficulty Rating 2 (Medium Difficulty)	Question Difficulty Rating (High Difficulty)
Questions # and #	Questions # and #	Questions # and #

Method for Calculating CBC Ratio

This section describes the procedures for calculating the Content-bearing Click (CBC) Ratio. The purpose of the CBC Ratio was to serve as a quantitative measure reflecting the performance of an Internet IR strategy. Since the hypertext "click" is common to many Internet IR strategies, it was chosen as the fundamental unit of measure rather than the "document." The CBC Ratio is a ratio of hypertext click counts. A "content-bearing" click was defined as any hypertext click that was used to retrieve possibly relevant information, as opposed to a hypertext click that was used for other reasons, such as the "search" click that began a database search or a "navigation" click that was used to traverse a WWW-based information resource. The CBC Ratio evaluated the utility of those clicks that an individual judge believed would result in the retrieval of relevant information. Other clicks, such as the aforementioned "search" and "navigation" clicks and those potential clicks that were obviously not going to produce relevant information (as judged by each subject based on contextual information, such as an abstract, that accompanies the click) were ignored. This meant that if, for example, a Medical World Search session retrieved 30,000 WWW documents, the denominator would not be "30,000," rather the denominator would be the sum of all content bearing clicks. This approach took advantage of the high level of search interactivity that the Internet affords and allowed obviously irrelevant potential clicks (as judged by the accompanying contextual information) to be ignored.

Because the CBC Ratio is not based on batch retrieval, as is the case with classical IR evaluation, there needed to be a logical way to dictate when counting should end so that the CBC Ratio can be calculated. Because in primary care medicine, a physician has only a short window of opportunity to find relevant information in order for it to be used during patient care, the search sessions for each question were limited to five minutes (Roland, Bartholomew, Courtenay, Morris & Morrell, 1992). Therefore, the CBC Ratio denominator was the count of all clicks that were stated by the subject as potentially leading to relevant information within the five-minute search window. The CBC numerator was either one or zero, reflecting whether the judge was able to retrieve relevant information in the allotted time (1) or whether the judge was not able to retrieve relevant information in the allotted time (0):

$$\text{CBC Ratio} = \frac{\text{\# of relevant Content-bearing Clicks}}{\text{\# of all relevant and not relevant Content-bearing Clicks}}$$

The CBC Ratio can be viewed as a false drop measure that determines the average number of irrelevant content-bearing clicks that an enduser must check before retrieving relevant information. As such, it is an task-oriented approach to IR evaluation (Hersh, Pentecost & Hickam 1996) that measures how effectively an IR strategy retrieves information relevant to an individual's problem.

Measurement Study

This section describes the procedures by which research questions 1 and 2 were addressed in this study.

R1. Does a modified classical IR measurement model translate into a reliable methodology for measuring the performance of IR strategies in the Internet environment?

R2. What is the most reliable procedure for aggregating human relevance judgments in the Internet IR environment?

Overview of Procedures

The evaluation of the reliability of a measurement model is accomplished with a measurement study in which the statistical quantification of measurement error is calculated (Friedman and Wyatt, 1997). The measure that reflects the quantified error estimate is *internal consistency reliability*, which is determined by Cronbach's alpha (α), a score ranging from 0 to 1 (Carmines and Zeller, 1979). A measurement model is considered reliable if it produces the same or very similar scores for identical phenomena at all time. In these terms, any relevance-based quantitative IR measurement model is unreliable unless subjective relevance is somehow taken into account.

A common method for testing the reliability of a measurement model follows a test-retest research design in which a series of measurements is taken at one point in time and are compared to another set of measurements at a different point in time. The key for test-retest reliability is that the setting in which actual measurements take place is simulated as closely as possible for the entire series of evaluation tests. If conditions are not sufficiently similar, then confounding variables may be introduced. However, Friedman and Wyatt point out that test-retest reliability is best suited for physical measurements and does not lend itself to a complicated measurement situation as exists in information systems. The main reason is that because humans are part of the measurement process, it is difficult to recreate experimental conditions that rule out confounding variables. The alternative to test-retest is the method of multiple simultaneous observations (Friedman & Wyatt, 1997).

As an alternative to test-rest, Friedman and Wyatt suggest the Method of Multiple Simultaneous Observations. The method evaluates reliability by structuring collected data during multiple applications of the measurement model. This is accomplished by collecting data about a measurement procedure using multiple applications of the measurement model and by calculating a reliability coefficient (Cronbach's alpha) that reflects the level of consistency inherent in the measurement model.

The assumption underlying this methodology is objectivist in nature: All independent observations of the same phenomenon should yield the same result. The closer the observations approach agreement for each object, the more reliable and therefore "objective" the measurement procedure can be considered to be. Disagreement reflects "subjectivity" on the part of the procedure in that it does not capture the empirical essence of the phenomena under study. An effective measurement procedure is one that restricts variability of scores to the "true score" and minimizes variability relative to other sources, which contribute to errors that erode reliability.

The proposed IR measurement procedure is investigated as an alternative to those that rely on relevance-based IR measures positing "invariant" relevance relationships between topics and documents. The proposed measurement model is designed to evaluated IR strategies based on how well an IR strategy performs in retrieving information that is cognitively accessible to judges. The retrieved information is not judged as universally relevant, but only as relevant to the immediate situation for the specific judge as reflected by a given level of difficulty. By gathering data over many judges, this performance measure can be evaluated for reliability prior to its use in comparing how well one IR strategy performs compared to another.

Testable Hypotheses

To evaluate the proposed measurement model, four procedures for aggregating human relevance judgments were compared against a standard for information technology performance (H1) and then compared to each other (H2). The results of a Cronbach's alpha (α) calculation indicated which of the three aggregation procedures is most reliable for the set of data collected in this study. Following are the hypotheses from research questions 1 and 2:

H1: A measurement model is reliable if its reliability coefficient exceeds the standard for a discipline:

$\alpha(H_1) > .70$

H2: Human relevance judgments are more reliable measuring instruments when aggregated based on the subjective perception of cognitive difficulty of questions rather than when aggregated based on the objective cognitive difficulty of questions rather than based on the cognitive class of individual judges and rather than over all judges regardless of cognitive class:

α(aggregated based on the subjective perception of cognitive difficulty of question) >

α(aggregated based on the objective cognitive difficulty of question) >

α(aggregated over cognitive class of judge) >

α(aggregated over all judges regardless of cognitive class)

Calculation of Cronbach's Alpha

The formula for Cronbach's alpha (α) below is from Friedman and Wyatt (1997):

$$\alpha = 1 - \frac{SS_{error} / (n_i - 1)(n_j - 1)}{SS_{objects} / (n_i - 1)}$$

where n_i is the number of objects and n_j the number of observations. Full details on the calculation of Cronbach's alpha are in Appendix D.

Demonstration Study

This section describes the procedures by which research questions 3, 4, and 5 were addressed in this study.

R3. In terms of Internet IR performance, which web catalog employs the more effective strategy for the manual monothetic/hierarchical classification of WWW resource collections?

R4. In terms of Internet IR performance, which web database employs the more effective strategy for automatically indexing WWW resource collections?

R5. In terms of Internet IR performance, is web catalog more effective than a web database?

According to Friedman and Wyatt (1997), the goal of a demonstration study is to inform and enhance decisions about information resources. Additionally, in this study, the purpose was to show how the proposed measurement model is applied to evaluate the performance of Internet IR strategies.

Dependent and Independent Variables

To demonstrate the proposed measurement model in an applied setting, the independent variable was “IR strategy” and the dependent variable was the CBC Ratio. The following four IR strategies were compared:

1. IR Strategy: Use Medical Matrix (Domain-specific web catalog (Medicine) available online at <http://www.medmatrix.org/>).
2. IR Strategy: Use Yahoo! (Domain-independent web catalog available online at <http://www.yahoo.com>).
3. IR Strategy: Use “Major Sites” interface of Medical World Search (Domain-specific web database (Medicine) available online at <http://www.mwsearch.com/>).
4. IR Strategy: Use “Simple Search” interface of Alta Vista (Domain-independent web database available online at <http://www.altavista.digital.com/>).

Testable Hypothesis

The purpose of the proposed measurement model was to facilitate the comparison of IR strategy performance. The following are hypotheses based on research question 3, 4, and 5:

H3: In Internet IR tasks, a domain-specific web catalog performs significantly better than a domain-independent web catalog:

CBC Ratio(Medical Matrix) > CBC Ratio(Yahoo!)

H4: In Internet IR tasks, a domain-specific web database performs significantly better than a domain-independent web database:

CBC Ratio(Medical World Search) > CBC Ratio(Alta Vista)

H5: In Internet IR tasks, a web catalog performs significantly better than a web database:

CBC Ratio(Medical Matrix) > CBC Ratio(Medical World Search)

Data Analysis: Calculation of F-Ratio

Use of analysis of variance (ANOVA) one way classification was used to indicate whether the mean CBC Ratio score for each IR strategy is significantly different.

CHAPTER 4

RESULTS

Introduction

This chapter presents the results of the evaluation and demonstration of the Content-bearing Click (CBC) Ratio, the measurement approach developed in this dissertation. The first section presents results of the measurement study, in which various procedures for aggregating human relevance judgments were tested. The second section presents results of the demonstration study, in which the performance of Internet IR strategies were measured and compared.

Measurement Study Results

The purpose of the measurement study was to test which of several aggregation procedures for human relevance judgments is the most reliable within the CBC Ratio measurement approach. Cronbach’s alpha (α) was used to evaluate reliability. (See Appendix D for Cronbach's alpha calculations.) Following are the results for research questions 1 and 2 (Table 4.1).

Research question 1: "Does a modified classical IR measurement model translate into a reliable methodology for measuring the performance of IR strategies in the Internet environment?" The data show that a modified classical IR measurement model is a reliable methodology. Based on the fact that three of the four aggregation procedures achieved an $\alpha > .70$, hypothesis 1, in which a measurement model is reliable if its reliability coefficient exceeds the standard for a discipline, is accepted.

Table 4.1

Data for the Evaluation of Research Questions 1 and 2

Aggregation Procedure for CBC Ratio Scores	Identifier	α	α > .70
Across Relevance Judges Regardless of Cognitive Class	A	.813	Yes
Across Cognitive Class of Relevance Judge- Med Student	B	.076	No
Across Cognitive Class of Relevance Judge- FP Resident	C	.697	No
Across Cognitive Class of Relevance Judge- FP Clinician	D	.690	No
Across Questions by Objective Cognitive Difficulty	E	.902	Yes
Across Questions by Subjective Cognitive Difficulty	F	.883	Yes

Note. Med = Medical; FP = Family Practice.

Research question 2: "What is the most reliable procedure for aggregating human relevance judgments in the Internet IR environment?" The data shows that the most reliable procedure for aggregating human relevance judgments was across questions by objective cognitive difficulty. The two procedures that aggregated human relevance judgments across questions produced higher reliability than those which aggregated across judges by cognitive class or across judges regardless of cognitive class. Hypothesis 2, in which human relevance judgments were predicted as more reliable measuring instruments when aggregated based on the subjective perception of cognitive difficulty of questions rather than when aggregated based on the objective cognitive difficulty of questions rather than based on the cognitive class of individual judges and rather than over all judges regardless of cognitive class, was rejected.

Demonstration Study Results

The demonstration study addressed how the proposed measurement model, the CBC Ratio, was applied in a real-life setting to measure Internet IR strategy performance.

Comparisons were made between two web catalogs, Medical Matrix and Yahoo and two web databases, Medical World Search and Alta Vista. Medical Matrix and Medical World Search were selected for comparison because the domain of the study was primary care medicine. Table 4.2 shows the results for research questions 3, 4, and 5.

Table 4.2

Results for Research Questions 3, 4, and 5

Aggregation Procedure	Research Question 3 Catalog name: (CBC Ratio, F Ratio)	Research Question 4 Database name: (CBC Ratio, F Ratio)	Research Question 5 Catalog/database name: (CBC Ratio, F Ratio)
A	* YA: (.391, .43) MM: (.341)	* MW: (.426, .93) AV: (.359)	* MW: (.426, .24) YA: (.391)
B	* MM: (.398, .11) YA: (.353)	* MW: (.402, .68) AV: (.304)	* MW: (.402, .00) MM: (. 398)
C	* YA: (.319, .43) MM: (.241)	* MW: (.476, 1.70) AV: (.311)	* MW: (.476, 1.66) YA: (.319)
D	* YA: (.499, .69) MM: (.384)	* AV: (.463, .30) MW: (.399)	* YA: (.499, .07) AV: (.463)
E (low)	* YA: (.624, .29) MM: (.555)	* MW: (.501, .96) AV: (.383)	* YA: (.624, .96) MW: (.501)
E (medium)	* MM: (.356, .03) YA: (.335)	* MW: (.500, 1.59) AV: (.342)	* MW: (.500, .35) MM: (.356)
E (high)	* YA: (.213, .95) MM: (.111)	* AV: (.353, .46) MW: (.277)	* AV: (.353, 1.49) YA: (213.)
F (low)	* MM: (.615, 1.16) YA: (.472)	* MW: (.536, .11) AV: (.496)	* MM: (.615, .35) MW: (.536)
F (medium)	* YA: (.375, .37) MM: (.299)	* MW: (.314, .27) AV: (.249)	* YA: (.375, .19) MW: (.314)
F (high)	* YA: (.297, .57) MM: (.202)	* MW: (.367, .04) AV: (.344)	* MW: (.367, .27) YA: (.297)

Note. * = winning Internet IR strategy; YA = Yahoo; MM = Medical Matrix; MW = Medical World Search; AV = Alta Vista; low = low cognitive difficulty; medium = medium cognitive difficulty; high = high cognitive difficulty; CBC = Content-bearing Click.

Research question 3: "In terms of Internet IR performance, which web catalog employs the more effective strategy for the manual monothetic/hierarchical classification of WWW resource collections?" In data shows that the domain-independent web catalog (Yahoo) performed more effectively than the domain-specific web catalog (Medical Matrix) for 7 of the 10 aggregation procedures. Hypothesis 3, in which domain-specific web catalogs were predicted to perform significantly better than domain-independent web catalogs, is rejected.

Research question 4: "In terms of Internet IR performance, which web database employs the more effective strategy for automatically indexing WWW resource collections?" The domain specific web database (Medical World Search) performed more effectively than the domain-independent web database (Alta Vista) for 8 of the 10 aggregation procedures, but these results were not statistically different. Hypothesis 4, in which domain-specific web databases were predicted to perform significantly better than domain-independent web databases, is rejected.

Research question 5: "In terms of Internet IR performance, is a web catalog more effective than a web database?" The web databases performed more effectively than the web catalogs for 6 of the 10 aggregation procedures. Hypothesis 5, in which web catalogs were predicted to perform significantly better than web databases, is rejected.

Additional Findings

The design of the CBC Ratio enables additional findings that indicate how well an Internet IR strategy performs in retrieving relevant information at various levels of objective and subjective cognitive difficulty. Table 4.3 show the percentage of answers found by each subject and the time on task for locating that information using their assigned Internet IR strategy. Questions were subdivided according to low, medium and high levels of objective cognitive difficulty.

Table 4.3

Percentage of Answers found within Five Minutes and Average Time on Task for each Internet IR Strategy for Questions by Objective Cognitive Difficulty

Internet IR Strategy	Low %; TOT	Medium %; TOT	High %; TOT
Medical Matrix	77.8% 198 seconds	61.1% 237 seconds	16.7% 283 seconds
Yahoo	94.4% 170 seconds	61.1% 234 seconds	33.3% 260 seconds
Medical World Search	83.3% 167 seconds	94.4% 180 second	55.6% 241 seconds
Alta Vista	77.8% 177 seconds	55.6% 204 seconds	72.2% 184 seconds

NOTE: % = percentage of questions with found answers within 5 minutes; TOT = average time on task.

Table 4.4 show the percentage of answers found by each subject and the time on task for locating that information using their assigned Internet IR strategy. Questions were subdivided according to low, medium and high levels of subjective cognitive difficulty.

In summary, the results of the measurement study show that reliable measurement is theoretically possible in the Internet IR environment. However, the results of the demonstration study indicate that the proposed modification of the classical model requires further research to determine whether non-significant measurement results were due to the measurement model, itself, or due to the fact that the Internet IR strategies were actually performing so similarly that the CBC Ratio was unable to determine which was better. These results are further discussed in the concluding chapter.

Table 4.4

Percentage of Answers found within Five Minutes and Average Time on Task for each Internet IR Strategy for Questions by Subjective Cognitive Difficulty

Internet IR Strategy	Low	Medium	High
	%; TOT	%; TOT	%; TOT
Medical Matrix	85%	33.3%	61.1%
	188 seconds	256 seconds	256 seconds
Yahoo	85%	61.1%	66.7%
	204 seconds	226 seconds	240 seconds
Medical World Search	86.9%	83.3%	83.3%
	175 seconds	230 second	198 seconds
Alta Vista	93.7%	61.1%	83.3%
	148 seconds	212 seconds	200 seconds

NOTE: % = percentage of questions with found answers within 5 minutes; TOT = average time on task.

CHAPTER 5

DISCUSSION

Introduction

This chapter discusses the results of the study. It is divided into two sections. The first section discusses results of the measurement study and the second section discusses the results of the demonstration study.

Discussion of Measurement Study Results

Table 4.1 from the previous chapter shows that the procedure for aggregating human relevance judgements makes a difference when calculating the reliability of the Content-bearing Click (CBC) Ratio. Following the aggregation approach of the classical measurement model, in which human relevance judgments are aggregated across all users, an alpha of .813 was calculated. Also, by using a modified classical approach, in which human relevance judgments are aggregated over the cognitive class of question difficulty, whether subjectively or objectively determined, higher alphas were recorded (.883 and .902). However, the procedure for aggregating human relevance judgments across cognitive class of judge did not meet the standard.

An important task of the scientist is to theoretically account for empirical results that depart in potentially significant ways from previous research. In this dissertation, the measurement study involved the testing of a modified classical IR measurement model that proposed several procedures of aggregating human relevance judgments across questions by cognitive class of difficulty rather than across judges. The data show that such a modification results in higher reliability coefficients. To account for this result, this section discusses the theoretical shortcomings of the classically conceived IR metrical space and suggests a relativistic IR metrical space as a possible alternative.

A metrical space is a theoretical device used in mathematics to describe the properties of the space in which acts of measurement take place. A classic example was Euclid's development of axioms and postulates for the geometrical measurement of the phenomena of our everyday world. An Euclidean space is a metrical space in which Euclid's axioms and postulates are the rules for measurement. The Euclidean space is flat, which corresponds to our common sense interpretation the three-dimensional space in which we reliably measure using rigid rods, such as yardsticks.

An advance in the understanding of physical measurement was made by the proposal of a relativistic metrical space by Einstein (Hawking, 1988). Specifically, the General Theory of Relativity rejected the idea of absolute space and time, which in turn meant that the proposition that the Euclidean metrical space was *universally* applicable was false. However, through the application of Riemannian geometry, Euclidean space remains a viable theoretical construct in terms of *local* measurement (Reichenbach, 1958).

In a similar fashion, Brookes (1980b) attempted to advance an understanding of IR measurement by presenting an alternative to the metrical space of the classical IR measurement model. In his paper, Brookes wrote of the problem that the classical (Cranfield) measurement model faced when attempting to apply a physical counting approach, such is used when counting paper documents, to human relevance judgments. In Brookes' view, treating relevance judgments as discrete entities raised a problem concerning the loss of information about individual relevance judgements. His suggestion was to use a log scale of perspective to preserve more of the context surrounding relevance and judgments about relevance. This perspectival IR metrical space would produce a landscape-like representation of a field of inquiry within which relevance-based quantitative measurement could take place. While not successful in modifying the classical measurement model, Brookes' effort to theorize about the IR metrical space provided a useful way for considering variation in human relevance judgments in the context of measurement theory.

Stephen Harter (1996) has also written about the theoretical problems of the classical measurement model. In his paper, stated that it is not his goal to derail the classical IR measurement model, but rather to emphasize that concerns about variations in human relevance judgments must be directly and explicitly accounted for during IR measurement. As discussed in Chapter 2, he conjectured that it was possible to maintain that human relevance judgments vary *and* that quantitative IR evaluation was possible. To accomplish this, he suggested that human relevance judgments vary predictably according to a combination of two factors: "differentiated relevance judgments according to cognitive state" and "topical relevance assigned in advance to a document." He reasoned that when combined, these factors would produce a reliable relevance-based quantitative IR measurement model. Underlying Harter's theoretical approach is the implication that the classical IR metrical space's use of objective human relevance judgments is too static, which prevents the metrical space from providing a sufficiently rich environment for reliably measuring the true phenomenon underlying IR strategy performance.

Based on Brookes' and Harter's work and the data of the measurement study, there is basis for rejecting the classical IR, or "Cranfieldean," metrical space because of its use of static human relevance judgments as the criterion for measuring IR strategy performance. As an alternative, the measurement study results suggest that in order to reliably measure, we need only reject the notion of *universal validity* that underpins the classical model's static human relevance judgment construct and replace it with the idea of *locally valid* human relevance judgments. The metrical space of the modified classical model suggests that a local Cranfieldean metrical space can exist, but only if human relevance judgments are aggregated across questions based on either an objective or subjective determination of its cognitive difficulty, rather than across judges. This is supported empirically by the achievement of higher levels of reliability for those aggregation procedures, and as a result, the CBC Ratio, as an example of a modified classical measurement model, was able to reliably measure using dynamic human relevance judgments as the criterion. Thus, the CBC Ratio is able to determine the best Internet IR strategy in terms of how well it retrieves relevant information that is also cognitively accessible relative to the capabilities of endusers.

In summary, the results of the measurement study enable options for selecting the best way to control for measurement error when evaluating the performance of Internet IR strategies. For example, when using aggregation procedures E or F, differences in CBC Ratios are more likely to be due to real differences in Internet IR strategy performance, rather than due to measurement error. However, caution must be used in applying the results of measurement study to the demonstration study. This measurement study evaluated the reliability of the CBC Ratio when measuring Internet IR performance independent of the actual strategy used. The concern of the measurement study was to indicate how reliability was influenced by the aggregation procedure applied to human relevance judgments.

Discussion of Demonstration Study Results

Table 5.1 shows the null hypotheses tested the demonstration study for each aggregation procedure. Each sub-section below presents discussion of the evaluation of each null hypothesis. Tables containing data <u>and ANOVA summary information</u> are located in Appendix E.

Table 5.1

<u>Null Hypotheses Tested in Demonstration Study for Each Aggregation Procedure</u>

$H3_0$: μ(Medical Matrix) $\leq$ μ(Yahoo!)
$H4_0$: μ(Medical World Search) $\leq$ μ(Alta Vista)
$H5_0$: μ(Medical Matrix) $\leq$ μ(Medical World Search)

<u>Discussion of Measurement Demonstration Results</u>

<u>Using Aggregation Procedure A</u>

The measurement demonstration results using aggregation procedure A (Tables E1 and E2), in which human relevance judgments were aggregated across judges regardless of cognitive class, were that Yahoo and Medical World Search had higher mean CBC Ratio scores than Medical Matrix and Alta Vista respectively. When compared against each other (Table E3), Medical World Search scored better than

Yahoo, though the difference in the mean CBC Ratio is not statistically significant. Based the alpha of .813 that was calculated in the measurement study above for this procedure for aggregating human relevance judgments, the conclusion can be made that the CBC Ratios were reliably measured.

Discussion of Measurement Demonstration Results

Using Aggregation Procedure B

The measurement demonstration results using aggregation procedure B (Tables E4 and E5), in which human relevance judgments are aggregated across cognitive class of relevance judge (medical students), were that Medical Matrix and Medical World Search have higher mean CBC Ratio scores than Yahoo and Alta Vista respectively. When compared against each other (Table E6), Medical World Search scored better than Yahoo, though the difference in the mean CBC Ratio is not statistically significant. However, based the alpha of .076 that was calculated in the measurement study above for this procedure for aggregating human relevance judgments, the conclusion cannot be made that the CBC Ratios were reliably measured.

Discussion of Measurement Demonstration Results

Using Aggregation Procedure C

The measurement demonstration results using aggregation procedure C (Tables E7 and E8), in which human relevance judgments are aggregated across cognitive class of relevance judge (family practice residents), were that Yahoo and Medical World Search have higher mean CBC Ratio scores than Medical Matrix and Alta Vista respectively. When compared against each other (Table E9), Medical World Search scores better than

Yahoo, though the difference in the mean CBC Ratio is not statistically significant. However, based the alpha of .697 that was calculated in the measurement study above for this procedure for aggregating human relevance judgments, the conclusion cannot be made that the CBC Ratios were reliably measured.

Discussion of Measurement Demonstration Results

Using Aggregation Procedure D

The measurement demonstration results using aggregation procedure D (Tables E10 and E12), in which human relevance judgments are aggregated across cognitive class of relevance judge (family practice clinician), were that Yahoo and Alta Vista have higher mean CBC Ratio scores than Medical Matrix and Medical World Search respectively. When compared against each other (Table E12), Alta Vista scores better than Yahoo, though the difference in the mean CBC Ratio is not statistically significant. However, based the alpha of .690 that was calculated in the measurement study above for this procedure for aggregating human relevance judgments, the conclusion cannot be made that the CBC Ratios were reliably measured.

Discussion of Measurement Demonstration Results

Using Aggregation Procedure E (Low)

The measurement demonstration results using aggregation procedure E (low) (Tables E13 and E14), in which human relevance judgments are aggregated across questions by objective cognitive difficulty rating, were that Yahoo and Medical World Search have higher mean CBC Ratio scores than Medical Matrix and Alta Vista respectively. When compared against each other (Table E15), Yahoo scores better than

Medical World Search, though the difference in the mean CBC Ratio is not statistically significant. Based the alpha of .902 that was calculated in the measurement study above for this procedure for aggregating human relevance judgments, the conclusion can be made that the CBC Ratios were reliably measured.

Discussion of Measurement Demonstration Results

Using Aggregation Procedure E (Medium)

The measurement demonstration results using aggregation procedure E (medium) (Tables E16 and E17), in which human relevance judgments are aggregated across questions by objective cognitive difficulty rating, were that Medical Matrix and Medical World Search have higher mean CBC Ratio scores than Yahoo and Alta Vista respectively. When compared against each other (Table E18), Medical World Search scores better than Medical Matrix, though the difference in the mean CBC Ratio is not statistically significant. Based the alpha of .902 that was calculated in the measurement study above for this procedure for aggregating human relevance judgments, the conclusion can be made that the CBC Ratios were reliably measured.

Discussion of Measurement Demonstration Results

Using Aggregation Procedure E (High)

The measurement demonstration results using aggregation procedure E (high) (Tables E19 and E20), in which human relevance judgments are aggregated across questions by objective cognitive difficulty rating, were that Yahoo and Alta Vista have higher mean CBC Ratio scores than Medical Matrix and Medical World Search respectively. When compared against each other (Table E21), Alta Vista scores better than Yahoo, though the difference in the mean CBC Ratio is not statistically significant. Based the alpha of .902 that was calculated in the measurement study above for this procedure for aggregating human relevance judgments, the conclusion can be made that the CBC Ratios were reliably measured.

Discussion of Measurement Demonstration Results

Using Aggregation Procedure F (Low)

The measurement demonstration results using aggregation procedure F (low) (Tables E22 and E23), in which human relevance judgments are aggregated across questions by subjective cognitive difficulty rating, were that Yahoo and Alta Vista have higher mean CBC Ratio scores than Medical Matrix and Medical World Search respectively. When compared against each other (Table E24), Alta Vista scores better than Yahoo, though the difference in the mean CBC Ratio is not statistically significant. Based the alpha of .883 that was calculated in the measurement study above for this procedure for aggregating human relevance judgments, the conclusion can be made that the CBC Ratios were reliably measured.

Discussion of Measurement Demonstration Results

Using Aggregation Procedure F (Medium)

The measurement demonstration results using aggregation procedure F (medium) (Tables E25 and E26), in which human relevance judgments are aggregated across questions by subjective cognitive difficulty rating, were that Yahoo and Alta Vista have higher mean CBC Ratio scores than Medical Matrix and Medical World Search respectively. When compared against each other (Table E27), Alta Vista scores better than Yahoo, though the difference in the mean CBC Ratio is not statistically significant. Based the alpha of .883 that was calculated in the measurement study above for this procedure for aggregating human relevance judgments, the conclusion can be made that the CBC Ratios were reliably measured.

Discussion of Measurement Demonstration Results

Using Aggregation Procedure F (High)

The measurement demonstration results using aggregation procedure F (high) (Tables E28 and E29), in which human relevance judgments are aggregated across questions by subjective cognitive difficulty rating, were that Yahoo and Medical World Search have higher mean CBC Ratio scores than Medical Matrix and Alta Vista respectively. When compared against each other (Table E30), Medical World Search scores better than Medical Matrix, though the difference in the mean CBC Ratio is not statistically significant. Based the alpha of .883 that was calculated in the measurement study above for this procedure for aggregating human relevance judgments, the conclusion can be made that the CBC Ratios were reliably measured.

In summary, the CBC Ratio was developed as an example of a modified classical measurement model. The demonstration study presented various ways in which the CBC Ratio can be used to compare how well Internet IR strategies perform in retrieving relevant information as judged by endusers.

Discussion of Additional Findings

An important application of the CBC Ratio concerned it ability to measure how well an Internet IR strategy performed in retrieving relevant information as various levels of objective and subjective cognitive difficulty. In addition to this capability, the time component of the CBC Ratio allowed for a simulation of the working environment of the subjects. A five-minute limit was placed on the primary care subjects in order to simulate the time normally available for information retrieval during patient care.

As shown in Table 4.3, web catalogs and web databases were able to consistently facilitate the retrieval of information verifying answers to questions with low objective cognitive difficulty within the five minute limit set for this study, with Yahoo performing the best. For questions of medium objective cognitive difficulty, the domain-specific web database, Medical World Search, performed the best. For questions of high objective cognitive difficulty, the domain-independent web database, Alta Vista, performed the best. In addition, web databases facilitated faster retrieval of information verifying answers to questions with medium and high objective cognitive difficulty than web catalogs.

As shown in Table 4.4, web catalogs and web databases were able to consistently facilitate the retrieval of information verifying answers to questions with low subjective cognitive difficulty, while web databases were more consistent than web catalogs for questions with high subjective cognitive difficulty. In addition, web databases facilitated faster retrieval of information verifying answers to questions with low and high subjective cognitive difficulty than web catalogs. Finally, Medical Web Search consistently facilitated high levels of retrieval performance when facilitating the retrieval of answers to questions of all levels of subjective cognitive difficulty.

In summary, the design of the CBC Ratio allowed for additional findings to be presented concerning the ability of Internet IR strategies to retrieve relevant information within the length of time normally allow for subjects of the domain under study.

Conclusions from these results are outlined in the final chapter.

CHAPTER 6

CONCLUSION

This study developed, evaluated and demonstrated the Content-bearing Click (CBC) Ratio, whose theoretical basis was a modification of the classical measurement model, for use in the Internet IR environment. Results showed that a relevance-based quantitative measurement approach is possible that exceeds the standard for reliable measurement. It was also shown that the CBC Ratio is flexible enough to answer a variety of questions about Internet IR strategy performance, including measuring how well strategies perform against each other and how well they perform in retrieving relevant information at various levels of cognitive difficulty.

The CBC Ratio was developed in order to meet two problems of measuring Internet IR strategies: variation human relevance judgments and fundamental unit of measure. This study tested various procedures for aggregating human relevance judgments suggested by Information Science relevance research from both the objective and subjective perspectives. The objective relevance perspective treats variation in human relevance judgments as part of random measurement error that is potentially controllable in experimental design using appropriate statistical means. The subjective relevance perspective treats variation in human relevance judgments as integral to the evaluation of information retrieval and views relevance judgments as part of the true score variation of measurement. The objective relevance perspective underlies the classical measurement model, and its widespread use since the early days of IR evaluation is attributable to its ease of application to quantitative experimental designs. The subjective relevance perspective, on the other hand, has been less successful so far in developing a quantitative measurement model. This study filled this void by exploring a different procedures for aggregating human relevance judgments. Rather than aggregating across judges, as in the classical model, this study aggregated across questions according to the class of cognitive difficulty determined objectively by subject experts and subjectively by individual users.

For the second problem, this study investigated a new fundamental unit of measure appropriate for the cyberspace of Internet IR. The proposed solution was the hypertext click. A case was made that its evolution from static to dynamic prevents the document, as used by the classical IR model, from serving as the fundamental unit of Internet IR measure. The ubiquity and functionality of hypertext clicks, which serve as the mechanism for initiating dynamic document retrieval, provided the rationale for its choice as the replacement for the static document.

From the results of this study, the conclusion can be drawn that dynamic human relevance judgments and dynamic documents can be integrated into the classical measurement model to produce a highly reliable modified classical model. A theoretical explanation was presented that this achievement also justifies a relativistic interpretation of the IR metrical space in which quantitative measurement is possible in local Cranfieldean metrical spaces.

In conclusion, this study points to new procedures for aggregating human relevance judgments when quantitatively measuring the performance of Internet IR strategies.

Implications and Recommendations for Future Research

The results of this study suggest clinical and information science implications, as well as several avenues for future research.

Clinical Implications

There are several implications of this study that are of interest to the clinical community. The value of Internet IR in the clinical environment was shown through the testing of strategies with a research design that used real clinical questions and simulated the time constraints of the average physician. In particular, high levels of success were shown when subjects sought information verifying answers to clinical questions with low objective or subjective cognitive difficulty. This was regardless of whether they chose to use web catalogs or web databases. In addition, web database use was indicated for use when seeking information verifying answers to clinical questions with high subjective cognitive difficulty. These results give specific advice to clinical endusers. For example, if a physician wanted to retrieve information verifying an "easy" question to determine whether the standard of care had changed or simply to double-check his or her hunch, then successful use of Internet IR is possible within the time constraints of clinical practice. Similarly, if a question arises that is perceived as having a high level of subjectively cognitive difficulty, the results of this study show that information can be retrieved within the time constraints of clinical practice.

There are also implication for those who train physicians to use the Internet. When educating physicians about Internet IR, trainers should not overlook the importance of domain-independent web catalogs, such as Yahoo. Though the content of Yahoo contains non-medical information, it breadth of organizational structure and large numbers of information resources enabled the retrieval of clinically useful information within the five minute window of this study. Medical library and informatics personnel need to be aware of the potential of such Internet IR approaches when designing their Internet training curricula. Also, those physicians who have not received formal training should be made aware of the potential benefits of domain-independent web catalog use.

Information Science Implications and Future Research

There are several implications of this study that are of interest to the Information Science community. The modification of the classical measurement model took the form of the CBC Ratio. The CBC Ratio's accounting for dynamic relevance judgments and dynamic documents produced a reliable measurement model. This suggests that future research in Information Science should consider the need for units of measure, such as recall and precision, that are not simple ratios of document counts, because the nature of the document has fundamentally changed due to their dynamic potential. The CBC Ratio basis in the hypertext click was presented as one such measuring unit.

The next step for this research program is to investigate the reasons for a lack of statistical significance in the demonstration study. At this point, it is not known whether the Internet IR strategies were performing in with very similar characteristics or whether the CBC Ratio is in need of further development as a measurement model. Approaches to investigating the CBC Ratio include replicating this study with either clinicians, resident, or medical students alone, replicating the study and compare only two Internet IR strategies, or replicating the study and increasing the number of questions used to test each Interne IR strategy.

Another area of research is to investigate the qualitative aspects of the CBC Ratio to determine good CBC Ratio levels. Data concerning individual perception of search success must be collected and matched to CBC Ratio scores. This research would enable better interpretation of future measurement using the CBC Ratio.

It would also be of interest to investigate the optimal length of Internet IR searches. This means determining how long an enduser should search in order to maximize the likelihood of retrieving relevant information. The CBC Ratio is a measurement model that can treat time as a variable in order to test for the optimal search length. An understanding of the optimal search time is an important next step for this research program.

Finally, this study should be replicated for other clinical specialties, as well as for other domains in order to determine whether the theoretical proposition of a relativistic IR metrical space is justified and to apply the CBC Ratio and its possible analogs in other Internet IR settings.

APPENDIX A

PRETEST QUESTIONNAIRE

PRETEST QUESTIONNAIRE

To perform as a subject in this study, you must be able to answer the following questions. If you are unsure of any, please tell the investigator prior to beginning assigned information task:

1. What is a hyperlink and what purpose does it serve in the World Wide Web environment?

____

2. State the difference between a web site and a web page.

____

3. State the difference between a web site catalog (e.g., Yahoo! and Medical Matrix) and a web page database (e.g., Alta Vista and Medical World Search).

____

4. Briefly state quality issues important for evaluating Internet-based medical information.

____

____

5. State how web-based information (i.e., web sites and web pages) differs from traditional bibliographic retrieval (e.g., MEDLINE).

____

____

APPENDIX B

INSTRUCTIONS TO PARTICIPANTS

INSTRUCTIONS TO PARTICIPANTS

Thank you for participating in this study.

We would like for you to help us test World Wide Web (WWW) information retrieval strategies. Your knowledge of medicine is not being evaluated.

Contained in this packet are six multiple-choice questions, one per page, for you to work sequentially. When signaled by the investigator, turn the page and begin. For each question, please do the following:

1. Read the question.

2. Select an answer by marking it with an "X" in the space provided (if you do not know the answer, please make a response based on your best judgement).

3. Rate the question for level of difficulty by placing a hashmark representing the confidence you have in the correctness of your answer. For example:

   |______________________________|

   I am confident of my answer   I do not know the answer

4. Attempt to verify your answer by locating information on the WWW using your assigned search method. You will have up to five minutes to do so for each question. Please alert the investigator when you begin your search and when you have found a verification of your answer (or when you find a contradiction of your answer -- see step 5 below). The investigator will time your session in order to determine how long it takes (up to five minutes) for you to find the information that supports or changes your answer.

5. If on any of the six questions you find information that changes your initial answer to that question, please note this on your answer sheet. Please do not change your answer unless you find information that changes it.

6. As you search for information, investigator will keep track of each "false positive," defined as each time that you "click" on a hypertext link expecting to find the answer to your question but are dissatisfied with the resulting information and are required to continue searching. If you are unsure of this concept, please alert the investigator prior to beginning you session.

APPENDIX C

QUESTION SET

1. Question

A 45-year-old woman says that she has been sick for several weeks. First, she had a viral-like illness with malaise, anorexia, and low-grade fever. As she was slowly recovering from this illness, she noted pain in her neck and a return of malaise and fever. She has noticed tenderness over her lower neck on both sides. She wonders if there is something wrong because her heart constantly beats fast and sometimes appears to be irregular. Examination reveals a painful, slightly enlarged thyroid gland, tachycardia, and a fine tremor of both hands. Which of the following diagnoses is most closely associated with this case?

_____ A. Graves' disease

_____ B. Subacute (painful) thyroiditis

_____ C. Hashimoto's thyroiditis

_____ D. Toxic multinodular goiter

_____ E. Nontoxic diffuse goiter

2. Place a hash mark across the line below to rate the difficulty level of this question:

|______________________________|

I am confident of my answer I do not know the answer

[Begin searching now--please signal investigator when you are ready to start. You will have up to five minutes to locate information verifying the answer specified above.]

3. Did you find information verifying your answer to this question?

_____ No _____ Yes

4. False Positives

5. Change in Answer

Did the *result* of your web search cause you to change your initial answer to this question?

___ No ___ Yes (if yes, what is your new answer?: ___ A ___ B ___ C ___ D ___ E

2. Question

A 60-year-old man tells you that, over the course of the previous three months, he has experienced fatigue and has been losing weight. He has a history of cigarette smoking but quit about 10 years ago. Suspecting a malignancy, you order a chest x-ray as part of the investigation. It shows a widened mediastinum but no evidence of discrete pulmonary lesions. A screening panel of blood chemistry tests is normal except for the serum level of calcium which is increased (11.8 mg/dL [normal, 8.5 mg/dL to 10.5 mg/dL]). The serum levels of phosphorous, albumin, and creatinine are normal. Additional studies show that the serum levels of both parathyroid hormone (PTH) and parathyroid-hormone-related-peptide (PTH-RP) are low, whereas the serum level of 1,25-dihydroxyvitamin D (1,25-[OH]$_2$-D is markedly elevated. Based on the available information, which of the following would be the most likely diagnosis?

_____ A. Squamous cell carcinoma of the lung

_____ B. Primary hyperparathyroidism

_____ C. Lymphoma

_____ D. Multiple myeloma

_____ E. Metastatic renal cell carcinoma

2. Place a hash mark across the line below to rate the difficulty level of this question:

|______________________________|

I am confident of my answer I do not know the answer

[Begin searching now--please signal investigator when you are ready to start. You will have up to five minutes to locate information verifying the answer specified above.]

3. Did you find information verifying your answer to this question?

_____ No _____ Yes

4. False Positives

5. Change in Answer

Did the *result* of your web search cause you to change your initial answer to this question?

___ No ___ Yes (if yes, what is your new answer?: ___ A ___ B ___ C ___ D ___ E

3. Question

In late November, a small, commercial fishing boat working out of Maine developed engine trouble at sea. At the time, wave heights were eight to ten feet and the boat began taking on water and started to sink. The three-man crew was forced to abandon the boat and spend the next 48 hours in a small life raft before being rescued by the Coast Guard. On e of the men complained that even thought he was wearing seaboots, his feet had gotten wet. They now felt "tingly" and were found to be pale, cool, damp, and slightly edematous. What cold-associated diagnosis is the most probably diagnosis?

_____ A. Immersion foot

_____ B. Raynaud's phenomenon

_____ C. Hypothermia

_____ D. Frostbite

_____ E. Chilblains

2. Place a hash mark across the line below to rate the difficulty level of this question:

|______________________________|
I am confident of my answer   I do not know the answer

[Begin searching now--please signal investigator when you are ready to start. You will have up to five minutes to locate information verifying the answer specified above.]

3. Did you find information verifying your answer to this question?

_____ No _____ Yes

4. False Positives

5. Change in Answer

Did the *result* of your web search cause you to change your initial answer to this question?

___ No ___ Yes (if yes, what is your new answer?: ___ A ___ B ___ C ___ D ___ E

4. Question

A 24-year-old woman who has had multiple sexual partners develops a copious, frothy, yellowish-green vaginal discharge. She says that she has pruritus of the external genitalia, mild abdominal discomfort, and dysuria. The pH of the vaginal secretions is >5. A potassium hydroxide (KOH) preparation creates a strong fishy odor. What is the vulvovaginal infection most closely associated with this clinical description?

_____ A. Trichomonal vaginitis

_____ B. Bacterial vaginosis

_____ C. Vulvovaginal candidiasis

_____ D. Chlamydial vaginitis

_____ E. Gonorrhea

2. Place a hash mark across the line below to rate the difficulty level of this question:

|___________________________|

I am confident of my answer I do not know the answer

[Begin searching now--please signal investigator when you are ready to start. You will have up to five minutes to locate information verifying the answer specified above.]

3. Did you find information verifying your answer to this question?

_____ No _____ Yes

4. False Positives

5. Change in Answer

Did the *result* of your web search cause you to change your initial answer to this question?

___ No ___ Yes (if yes, what is your new answer?: ___ A ___ B ___ C ___ D ___ E

5. Question

A 55-year-old woman has pain in her left ear. Examination reveals a vesicular eruption in the outer portion of the left external auditory canal and evidence of a left-sided facial weakness. Since paralysis of the facial muscles can be caused by any of several neurologic lesions, these lesions can usually be differentiated from each other on the basis of the clinical history and physical examination. Which neurologic lesion is most closely associated with this clinical description?

_____ A. Idiopathic facial palsy (Bell's palsy)

_____ B. Right cerebral hemisphere stroke

_____ C. Infarction in the lower pons/upper medulla

_____ D. Infarction in the midbrain

_____ E. Presumed herpes zoster infection of the geniculate ganglion

2. Place a hash mark across the line below to rate the difficulty level of this question:

|____________________________|

I am confident of my answer I do not know the answer

[Begin searching now--please signal investigator when you are ready to start. You will have up to five minutes to locate information verifying the answer specified above.]

3. Did you find information verifying your answer to this question?

_____ No _____ Yes

4. False Positives

5. Change in Answer

Did the *result* of your web search cause you to change your initial answer to this question?

___ No ___ Yes (if yes, what is your new answer?: ___ A ___ B ___ C ___ D ___ E

6. Question

A 25-month-old child was seen in your office because of an acute otitis media for which you prescribed a 10-day course of amoxicillin. On the day that she was given her last dose of amoxicillin, she had a peanut butter and jelly sandwich. By the following day, her mother noticed "a few hives" on the child's trunk. Twenty-four hours later, the rash had spread to the child's extremities with mild swelling of the hands and feet, and 12 hours later, the mother tells you that the child has a "fever" and is refusing to walk. Your physical examination reveals that the child was somewhat irritable with a temperature of 40.3C (104.6F); respiratory rate, 24; blood pressure, 100/68 mm Hg; pulse rate 100 beats/min. A generalized, morbilliform rash and scattered urticarial lesions are present on the trunk and extremities, and the hands, feet, knee joints, and elbow joints are swollen. However, there are no mucosal lesions, and the remainder of the physical examination is unremarkable. The most likely diagnosis is:

_____ A. anaphylaxis secondary to peanuts

_____ B. Steven-Johnson syndrome

_____ C. a viral exanthem

_____ D. serum sickness-like reaction

_____ E. hereditary angioedema

2. Place a hash mark across the line below to rate the difficulty level of this question:

|_____________________________|

I am confident of my answer   I do not know the answer

[Begin searching now--please signal investigator when you are ready to start. You will have up to five minutes to locate information verifying the answer specified above.]

3. Did you find information verifying your answer to this question?

_____ No _____ Yes

4. False Positives

5. Change in Answer

Did the *result* of your web search cause you to change your initial answer to this question?
___ No ___ Yes (if yes, what is your new answer?: ___ A ___ B ___ C ___ D ___ E

APPENDIX D

DATA AND CRONBACH’S ALPHA CALCULATION

FOR MEASUREMENT STUDY

DATA AND CRONBACH'S ALPHA CALCULATION

FOR MEASUREMENT STUDY

This appendix contains data and Cronbach's alpha calculations for the measurement study.

For each procedure for aggregating relevance judgments, Cronbach's alpha (α) is calculated as follows (Friedman and Wyatt, 1997):

$$\alpha = 1 - \frac{SS_{\text{error}} / (n_i - 1)\,(n_j - 1)}{SS_{\text{objects}} / (n_{\boldsymbol{i}} - 1)}$$

where n_i is the total number of objects and n_j is the total number of observations. Before obtaining SS_{error}, the calculation for the total sum of squares (SS_{total}) was made:

$$SS_{\text{total}} = \sum X^2_{ij} - \frac{(\sum X_{ij})^2}{n_i\, n_j}$$

Along with the sum of squares for objects (SS_{objects}):

$$SS_{\text{objects}} = \frac{\sum (\sum X_{ij})^2}{n_j} - \frac{(\sum X_{ij})^2}{n_i\, n_j}$$

and the sum of squares for observations ($SS_{\text{observations}}$):

$$SS_{\text{observations}} = \frac{\sum (\sum X_{ij})^2}{n_j} - \frac{(\sum X_{ij})^2}{n_i\, n_j}$$

From these three quantities, the sum of squares for error (SS_{error}) is computed:

$$SS_{\text{error}} = SS_{\text{total}} - SS_{\text{objects}} - SS_{\text{observations}}$$

and is used to calculated Cronbach's alpha above.

Table D1

Matrix for α Calculation based on the Aggregation of CBC Ratio Scores across All Relevance Judges Regardless of Cognitive Class

Quest. #
|
|

Judge #

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
1	.33	0	.25	.2	0	0	.33	0	1	.5	1	0	.5	1	.5	.5	1	0	1	1	.14	0	.33	.5	1	.33	.25	0	1	0	1	.2	.5	1	1	0
2	0	0	1	.5	0	0	1	0	0	.25	.2	1	0	0	0	0	0	0	0	.5	.33	0	.33	.25	0	0	.5	0	.5	0	0	0	.25	.5	.25	0
3	1	.33	0	1	1	.33	.5	0	.33	.33	.2	.5	0	0	0	1	.33	.5	.33	1	.33	.33	0	1	1	.5	.5	0	1	.33	.25	0	1	1	.5	1
4	1	1	.5	.33	.5	1	.25	0	1	0	.5	.5	.33	1	1	.33	.25	.33	1	1	.2	.33	.2	0	.5	.33	1	1	1	1	1	.33	.5	.25	.25	0
5	0	1	.25	0	.25	.25	.33	.5	1	0	0	0	0	0	0	.5	0	.5	.25	.33	.16	.33	0	0	0	0	1	.5	.33	1	.5	.25	.5	.25	.25	1
6	0	0	.5	0	1	0	.5	0	.5	.5	0	0	0	0	0	0	0	.5	0	1	0	0	1	1	0	0	0	0	1	.33	.33	.33	.25	.33	.25	.5

Relevance Judgments Aggregated across all Judges Regardless of Cognitive Class

Table D1 contains the data matrix for calculating Cronbach's alpha when relevance judgments are aggregated over all judges regardless of cognitive class. Table D2 gives the results of each sum of squares calculation used to determine the alpha.

Table D2

Sum of Squares and Cronbach's Alpha Calculations for Relevance Judgments Aggregated across all Judges Regardless of Cognitive Class

Calculation	Result
n_i	6
n_j	36
SS_{total}	30.50
$SS_{objects}$	3.29
$SS_{observations}$	5.62
SS_{error}	21.59
Cronbach's alpha (α)	.813

Relevance Judgments Aggregated across Judges by Cognitive Class

Tables D3 through D8 contain the data matrices and sum of squares calculations used to determine Cronbach's alpha for relevance judgments aggregated across judges by cognitive class.

Table D3

Matrix for α Calculation based on the Aggregation of CBC Ratio Scores across Cognitive Class of Relevance Judge: Medical Students

Question. #	Judge # 1	2	3	4	5	6	7	8	9	10	11	12
1	.33	0	1	1	0	0	0	0	.33	1	1	0
2	.25	1	0	.5	.25	.5	.2	.5	1	.33	0	0
3	0	0	1	.5	.25	1	0	0	.33	1	.25	0
4	.33	1	.5	.25	.33	.5	0	0	0	0	.5	0
5	1	0	.33	1	1	.5	.5	.25	.33	0	0	.5
6	1	.2	.2	.5	0	0	0	1	.5	.5	0	0

Table D4

Sum of Squares and Cronbach's Alpha Calculations for Relevance Judgments Aggregated across Cognitive Class of Judge: Medical Students

Calculation	Result
n_i	6
n_j	12
SS_{total}	10.06
$SS_{objects}$	.79
$SS_{observations}$	1.24
SS_{error}	8.03
Cronbach's alpha (α)	.076

Table D5

Matrix for α Calculation based on the Aggregation of CBC Ratio Scores across Cognitive Class of Relevance Judge: FP Residents

Question. #	Judge #											
	1	2	3	4	5	6	7	8	9	10	11	12
1	.5	0	0	.33	0	0	1	0	0	1	0	0
2	.5	0	0	1	0	0	.5	0	1	.33	.5	0
3	1	0	.33	.25	0	0	0	0	.5	.33	.5	.5
4	1	0	.33	1	.25	0	1	.5	1	1	.33	1
5	.143	.33	.33	.2	.16	0	0	0	.33	.33	.33	0
6	.33	.33	0	.2	0	1	.5	.25	1	0	0	1

Table D6

Sum of Squares and Cronbach's Alpha Calculations for Relevance Judgments Aggregated across Cognitive Class of Judge: FP Residents

Calculation	Result
n_i	6
n_j	12
SS_{total}	9.82
$SS_{objects}$	1.78
$SS_{observations}$	2.10
SS_{error}	5.94
Cronbach's alpha (α)	.697

Table D7

Matrix for α Calculation based on the Aggregation of CBC Ratio Scores across Cognitive Class of Relevance Judge: FP Clinicians

Question Number	Judge Number											
	1	2	3	4	5	6	7	8	9	10	11	12
1	1	0	1	.5	0	0	.33	0	.5	.33	0	0
2	.25	.5	.5	1	1	0	0	0	0	1	.5	0
3	1	.5	1	1	.33	1	0	0	.33	1	1	.33
4	1	0	.25	1	.5	.33	.2	0	0	.33	.25	.33
5	.5	.25	1	.5	.5	.25	1	.5	1	.25	.25	.33
6	1	.25	.5	.25	.25	.25	0	0	1	0	1	.5

Table D8

Sum of Squares and Cronbach's Alpha Calculations for Relevance Judgments Aggregated across Cognitive Class of Judge: FP Clinicians

Calculation	Result
n_i	6
n_j	12
SS_{total}	10.25
$SS_{objects}$	1.87
$SS_{observations}$	1.98
SS_{error}	6.40
Cronbach's alpha (α)	.690

Relevance Judgments Aggregated across Questions by Objective Cognitive Difficulty

Prior to calculating Cronbach's alpha for relevance judgments aggregated across question by objective difficulty rating, data were collected to reflect the consensus expert opinion of three teaching physicians with five or more years of clinical practice experience who were not part of the study sample. The data in Table D9 represent their determination of the objective cognitive difficulty of each question classified as low, medium, and high level of difficult.

Table D9

Consensus Expert Opinion concerning Objective Difficulty Rating for each Question

Question Difficulty Rating 1 (Low Difficulty)	Question Difficulty Rating 2 (Medium Difficulty)	Question Difficulty Rating (High Difficulty)
Questions #3 and #4	Questions #1 and #5	Questions #2 and #6

Inter-rater reliability = 0.67

Table D10

Matrix for α Calculation based on the Aggregation of CBC Ratio Scores across Questions by Objective Classification of Difficulty

Objective Difficulty Rating Classification
|
|

Question Answer #1-#36

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
1	.25	.33	1	1	1	.33	.33	.5	0	0	1	.5	.5	.33	.5	.5	0	.33	0	.33	0	1	1	.33	1	.5	1	.33	0	0	1	.33	1	0	1	.2
2	1	1	.33	1	.5	1	1	.5	.2	1	.5	0	0	0	1	.25	1	0	1	0	.14	.33	0	0	1	.5	0	.25	0	.5	.5	0	.2	.33	.33	1
3	0	.25	0	0	0	0	.25	0	0	.5	0	0	0	0	0	.33	0	0	.2	0	1	0	.5	.5	1	0	.25	1	.33	0	.5	.5	0	0	0	.25

Table D10 (continued)

Objective Difficulty Rating Classification
|
|

Question Answer #37-#72

	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72
1	.33	.25	1	.5	0	.33	.5	1	1	.33	1	1	1	.25	.5	.25	1	.2	1	.33	1	1	.33	.5	.33	.33	.5	.2	0	1	.25	0	0	1	.5	1
2	.33	.5	.25	0	.25	1	.25	.5	1	.33	0	1	.5	.25	0	1	.25	.5	.25	.33	.5	1	.5	0	0	0	0	.33	.25	0	0	0	.16	0	0	0
3	.25	1	.5	0	1	0	0	0	.33	.33	1	.5	0	0	1	0	1	0	0	0	.5	0	0	.25	0	0	.33	0	.5	0	0	.5	.33	.5	0	0

Tables D10 and D11 contain the data matrix and sum of squares calculations used to determine Cronbach's alpha for relevance judgments aggregated across questions by objective cognitive difficulty.

Table D11

Sum of Squares and Cronbach's Alpha Calculations for Relevance Judgments Aggregated across Questions by Objective Cognitive Difficulty

Calculation	Result
n_i	3
n_j	72
SS_{total}	30.50
$SS_{objects}$	2.76
$SS_{observations}$	8.62
SS_{error}	19.12
Cronbach's alpha (α)	.902

Relevance Judgments Aggregated across Questions by Subjective Cognitive Difficulty

Prior to calculating Cronbach's alpha for relevance judgments aggregated across question by subjective difficulty rating, data were collected to reflect individual judge's rating of the self-perceived cognitive difficulty level of each question. These data are in Table D12

Tables D13 and D14 contain the data matrix and sum of squares calculations used to determine Cronbach's alpha for relevance judgments aggregated across questions by subjective cognitive difficulty.

Table D12

CBC Ratio Scores for Subjective Cognitive Difficulty Rating Level 1

Question Number	Cognitive Difficulty Rating (Raw Score)	Cognitive Difficulty Rating (Class*)	CBC Ratio
4	9	1	1
11	15	1	1
19	21	1	.2
22	1	1	.33
23	10	1	0
27	15	1	1
28	11.5	1	.5
29	14	1	.25
34	15	1	1
35	16.5	1	.25
39	18	1	.5
40	9.5	1	.25
41	1	1	.33
46	4.5	1	0
49	19	1	1
51	11	1	.33
53	9.5	1	1
57	17.5	1	.33
70	22	1	.5
73	16	1	.5
75	22	1	0
76	18	1	.33
82	18	1	1
88	6.5	1	1
91	20	1	.5
95	1	1	.5
97	10.5	1	1
100	12	1	.25
101	3.5	1	0
102	22	1	0
106	8	1	.33
107	19	1	.5
108	18	1	.5
109	21	1	1
110	10.5	1	0
112	2	1	1
113	20	1	.25
118	8.5	1	1
119	9.5	1	.33
120	8.5	1	1
128	3	1	0
130	19.5	1	.33
131	20	1	.33
136	16.5	1	.2
148	5.5	1	.5
151	5	1	.33
154	4.5	1	.33
155	14.5	1	0
160	19	1	1
161	13	1	1
166	7.5	1	1
177	16.5	1	.33
178	17.5	1	1
181	3.5	1	1
182	7	1	0
183	10.5	1	.25
184	8	1	1
185	7	1	.5
190	22.5	1	.33
191	4	1	.25
196	4	1	.5
197	13.5	1	.5
199	19	1	1
201	13	1	1
203	2	1	.25
205	3	1	1
206	6	1	.25
208	1	1	.25
209	6.5	1	.25
210	6	1	.25
213	16	1	1
215	6	1	1

* These CBC scores correspond with the top 72 raw cognitive difficulty rating scores (rating =1)

Table D12 (continued)

CBC Ratio Scores for Subjective Cognitive Difficulty Rating Level 2

Question Number	Cognitive Difficulty Rating (Raw Score)	Cognitive Difficulty Rating (Class*)	CBC Ratio
3	44	2	1
7	46	2	0
9	45	2	.33
15	33	2	0
16	24	2	.5
20	34	2	.5
24	48	2	0
25	25	2	0
26	39	2	0
30	24	2	1
37	44	2	.33
43	37	2	0
44	49	2	0
47	24	2	.5
50	31.5	2	0
58	39.5	2	0
59	31	2	0
69	29.5	2	.5
71	30	2	0
72	33	2	0
77	43	2	0
78	41	2	0
79	35.5	2	1
80	39.5	2	0
81	39	2	0
85	24	2	.5
87	23.5	2	0
89	31.5	2	0
92	34.5	2	0
93	47.5	2	1
94	23.5	2	.33
98	23.5	2	0
99	38.5	2	.33
105	31	2	.5
115	31	2	1
121	42	2	.143
123	34	2	.33
124	24	2	.2
127	34.5	2	0
129	27	2	.33
132	47	2	0
134	47	2	.33
138	49	2	1
142	39.5	2	0
144	45	2	1
145	32.5	2	1
146	45	2	0
147	23.5	2	1
150	45.5	2	0
152	33	2	0
153	25	2	.5
156	40	2	0
157	31	2	.25
159	26	2	.5
163	28.5	2	0
164	24	2	0
165	36	2	0
167	28.5	2	.5
172	31.5	2	1
173	43.5	2	.33
174	47	2	1
175	27	2	0
179	28	2	1
186	42	2	.33
188	34.5	2	0
194	29	2	.25
195	43	2	1
202	36	2	.25
204	41	2	.33
212	49	2	0
214	24	2	0
216	45	2	.5

*These CBC scores correspond with the middle 72 raw cognitive difficulty rating scores (rating =2)

Table D12 (continued)

CBC Ratio Scores for Subjective Cognitive Difficulty Rating Level 3

Question Number	Cognitive Difficulty Rating (Raw Score)	Cognitive Difficulty Rating (Class*)	CBC Ratio
1	81	3	.33
2	93	3	0
5	68	3	0
6	61	3	0
8	86	3	0
10	76	3	1
12	95	3	0
13	56	3	.25
14	65	3	1
17	55	3	.25
18	71	3	.5
21	81	3	1
31	50	3	0
32	91	3	0
33	79	3	.33
36	95	3	0
38	75	3	1
42	82	3	.5
45	61	3	0
48	58	3	0
52	50	3	1
54	57	3	.5
55	62	3	.5
56	99	3	.25
60	77	3	.5
61	97	3	1
62	53	3	.2
63	97	3	.2
64	50	3	.5
65	79	3	0
66	76.5	3	0
67	69	3	0
68	60.5	3	1
74	68	3	0
83	59.5	3	0
84	69.5	3	0
86	55	3	0
90	53	3	0
96	62.5	3	0
103	58.5	3	0
104	60.5	3	0
111	58	3	.33
114	56.5	3	0
116	61	3	.5
117	62.5	3	1
122	77	3	.33
125	70.5	3	.16
126	58.5	3	0
133	82	3	.33
135	84	3	0
137	85.5	3	0
139	55	3	.5
140	72.5	3	.25
141	57.5	3	1
143	63.5	3	0
149	75	3	0
158	87.5	3	.5
162	75	3	0
168	53	3	0
169	51	3	1
170	87	3	.5
171	86	3	1
176	93	3	0
180	77.5	3	.33
187	74	3	.2
189	62	3	0
192	64	3	.33
193	63.5	3	.5
198	59.5	3	.25
200	63	3	.5
207	49.5	3	.5
211	55	3	0

* These CBC scores correspond with the bottom 72 raw cognitive difficulty rating scores (rating =3)

Table D13

Matrix for α Calculation based on the Aggregation of CBC Ratio Scores across Questions by Subjective Classification of Difficulty

Subjective Difficulty Rating Classification
|
|

Question Answer #1-#36

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
1	1	1	.33	1	.5	1	1	.5	.2	1	0	.25	0	0	.25	.33	1	1	1	.33	.33	.5	0	.33	.25	1	.5	0	.33	.5	1	1	.33	1	1	1
2	.5	0	0	0	1	.25	1	0	1	0	.14	.33	0	0	0	.25	0	0	.5	0	0	0	0	0	.33	0	0	0	1	.5	.5	.33	.5	.5	0	.33
3	0	1	.5	0	.25	0	.5	.5	0	.2	.33	.33	1	.2	0	1	0	.5	.5	1	0	.25	1	.33	0	.5	.5	0	0	0	.25	.5	1	.33	0	0

Table D13 (continued)

Subjective Difficulty Rating Classification
|
|

Question Answer #37-#72

	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72
1	.25	.5	.25	1	.2	1	.33	1	1	.33	.5	.33	.33	.5	.25	0	.25	1	.25	.5	1	.33	0	1	.5	.25	0	1	.25	.5	.25	.33	.25	1	.5	0
2	0	.33	0	1	1	.33	1	.33	0.5	.2	0	1	.25	0	0	.5	1	.5	0	0	0	0	.33	1	0	0	0	.33	.33	1	.5	0	0	1	0	1
3	1	.33	1	0	1	.2	1	0.5	1	.25	0	0	0	.16	0	0	0	0	0	0	.5	0	0	.25	0	0	.33	0	.5	0	0	.5	.33	.5	0	0

Table D14

Sum of Squares and Cronbach's Alpha Calculations for Relevance Judgments Aggregated across Questions by Subjective Cognitive Difficulty

Calculation	Result
n_i	3
n_j	72
SS_{total}	30.50
$SS_{objects}$	2.26
$SS_{observations}$	9.43
SS_{error}	18.81
Cronbach's alpha (α)	.883

APPENDIX E

DATA TABLES AND SUMMARY ANOVA INFORMATION

FOR DEMONSTRATION STUDY

DATA TABLES AND SUMMARY ANOVA INFORMATION

FOR DEMONSTRATION STUDY

This appendix contains data tables and summary ANOVA information used in the discussion of the results of the demonstration study.

Table E1

Demonstration of Measurement of Web Catalog Performance for Aggregation Procedure A with Summary ANOVA Information

Web Catalog	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1,106)
Domain-specific: Medical Matrix	51.9%	164.0	.34074	
Domain-independent: Yahoo	63.0%	155.5	.39056	.43

Note. Ans. = Answers; Sec. = Seconds.

	Summary ANOVA				
Source	*SS*	*df*	*MS*	*F*	*p*
Between	.0670	1	.0670	.43	.5141
Within	16.5742	106	.1564		
Total	16.6412	107			

Table E2

Demonstration of Measurement of Web Database Performance for Aggregation Procedure A with Summary ANOVA Information

Web Database	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1,106)
Domain-specific: Medical World Search	77.8%	165.3	.42598	.93
Domain-independent: Alta Vista	68.5%	137.3	.35944	

Note. Ans. = Answers; Sec. = Seconds.

	Summary ANOVA				
Source	*SS*	*df*	*MS*	*F*	*P*
Between	.1195	1	.1195	.93	.3383
Within	13.6967	106	.1292		
Total	13.8162	107			

Table E3

Demonstration of Comparison of Best CBC Ratios for Aggregation Procedure A with Summary ANOVA Information

Top Performing Strategies	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1,106)
Domain-independent web catalog: Yahoo	63.0%	155.5	.39056	
Domain-independent web database: Medical World Search	77.8%	165.3	.42598	.24

Note. Ans. = Answers; Sec. = Seconds.

Table E3 (cont.)

	Summary ANOVA				
Source	*SS*	*df*	*MS*	*F*	*p*
Between	.0339	1	. 0339	.24	.6274
Within	15.1531	106	. 1430		
Total	15.1870	107			

Table E4

Demonstration of Measurement of Web Catalog Performance for Aggregation Procedure B (Medical Students) with Summary ANOVA Information

Web Catalog	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1, 34)
Domain-specific: Medical Matrix	61.1%	165.5	.39778	.11
Domain-independent: Yahoo	61.1%	173.3	.35333	

Note. Ans. = Answers; Sec. = Seconds.

	Summary ANOVA				
Source	*SS*	*df*	*MS*	*F*	*P*
Between	.0178	1	.0178	.11	.7452
Within	5.6303	34	.1656		
Total	5.6480	35			

Table E5

Demonstration of Measurement of Web Database Performance for Aggregation Procedure B (Medical Students) with Summary ANOVA Information

Web Database	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1, 34)
Domain-specific: Medical World Search	66.7%	179.2	.40222	.68
Domain-independent: Alta Vista	61.1%	140.2	.30444	

Note. Ans. = Answers; Sec. = Seconds.

	Summary ANOVA				
Source	*SS*	*df*	*MS*	*F*	*P*
Between	.0860	1	.0860	.68	.4158
Within	4.3101	34	.1268		
Total	4.3962	35			

Table E6

Demonstration of Comparison of Best CBC Ratios for Aggregation Procedure B (Medical Students) with Summary ANOVA Information

Top Performing Strategies	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1, 34)
Domain-specific web catalog: Medical Matrix	61.1%	165.5	.39778	
Domain-independent web database: Medical World Search	66.7%	179.2	.40222	.00

Note. Ans. = Answers; Sec. = Seconds.

Table E6 (cont.)

	Summary ANOVA				
Source	*SS*	*df*	*MS*	*F*	*p*
Between	.0002	1	.0002	.00	.9737
Within	5.4718	34	.1609		
Total	5.4720	35			

Table E7

Demonstration of Measurement of Web Catalog Performance for Aggregation Procedure C (FP Residents) with Summary ANOVA Information

Web Catalog	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1, 34)
Domain-specific: Medical Matrix	33.3%	190.3	.24056	
Domain-independent: Yahoo	61.1%	225.0	.31889	.43

Note. Ans. = Answers; Sec. = Seconds.

	Summary ANOVA				
Source	*SS*	*df*	*MS*	*F*	*P*
Between	.0552	1	.0552	.43	.5169
Within	4.3761	34	.1287		
Total	4.4313	35			

Table E8

Demonstration of Measurement of Web Database Performance for Aggregation Procedure C (FP Residents) with Summary ANOVA Information

Web Database	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1, 34)
Domain-specific: Medical World Search	83.3%	151.3	.47628	1.70
Domain-independent: Alta Vista	61.1%	153.1	.31111	

Note. Ans. = Answers; Sec. = Seconds.

	Summary ANOVA				
Source	*SS*	*df*	*MS*	*F*	*P*
Between	.2455	1	.2455	1.70	.2008
Within	4.9058	34	.1443		
Total	5.1513	35			

Table E9

Demonstration of Comparison of Best CBC Ratios for Aggregation Procedure C (FP Residents) with Summary ANOVA Information

Top Performing Strategies	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1, 34)
Domain-independent web catalog: Yahoo	61.1%	225.0	.31889	
Domain-independent web database: Medical World Search	83.3%	151.3	.47628	1.66

Note. Ans. = Answers; Sec. = Seconds.

Table E9 (cont.)

Source	*Summary ANOVA*				
	SS	*df*	*MS*	*F*	*p*
Between	.2230	1	.2230	1.66	.2060
Within	4.5560	34	.1341		
Total	4.7827	35			

Table E10

Demonstration of Measurement of Web Catalog Performance for Aggregation Procedure D (FP Clinicians) with Summary ANOVA Information

Web Database	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1, 34)
Domain-specific: Medical Matrix	61.1%	178.7	.38389	
Domain-independent: Yahoo	66.7%	132.2	.49944	.69

Note. Ans. = Answers; Sec. = Seconds.

Source	*Summary ANOVA*				
	SS	*df*	*MS*	*F*	*p*
Between	.1202	1	.1202	.69	.4136
Within	5.9643	34	.1754		
Total	6.0845	35			

Table E11

Demonstration of Measurement of Web Database Performance for Aggregation Procedure D (FP Clinicians) with Summary ANOVA Information

Web Database	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1, 34)
Domain-specific: Medical World Search	83.3%	171.0	.39944	
Domain-independent: Alta Vista	83.3%	123.7	.46278	.30

Note. Ans. = Answers; Sec. = Seconds.

	Summary ANOVA				
Source	*SS*	*df*	*MS*	*F*	*p*
Between	.0361	1	.0361	.30	.5889
Within	4.1237	34	.1213		
Total	4.1598	35			

Table E12

Demonstration of Comparison of Best CBC Ratios for Aggregation Procedure D (FP Clinicians)

Top Performing Strategies	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1, 34)
Domain-independent web catalog: Yahoo	66.7%	132.2	.49944	.07
Domain-independent web database: Alta Vista	83.3%	123.7	.46278	

Note. Ans. = Answers; Sec. = Seconds.

Table E12 (cont.)

Source	*Summary ANOVA*				
	SS	*df*	*MS*	*F*	*p*
Between	.0121	1	.0121	.07	.7901
Within	5.7157	34	.1681		
Total	5.7276	35			

Table E13

Demonstration of Measurement of Web Catalog Performance for Aggregation Procedure E (Low) with Summary ANOVA Information

Web Catalog	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1, 34)
Domain-specific: Medical Matrix	77.8%	198.3	.55500	
Domain-independent: Yahoo	94.4%	169.9	.62389	.29

Note. Ans. = Answers; Sec. = Seconds.

Source	*Summary ANOVA*				
	SS	*df*	*MS*	*F*	*p*
Between	.0427	1	.0427	.29	.5932
Within	4.9919	34	.1468		
Total	5.0346	35			

Table E14

Demonstration of Measurement of Web Database Performance for Aggregation Procedure E (Low) with Summary ANOVA Information

Web Database	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1, 34)
Domain-specific: Medical World Search	83.3%	166.8	.50111	.96
Domain-independent: Alta Vista	77.8%	176.9	.38278	

Note. Ans. = Answers; Sec. = Seconds.

	Summary ANOVA				
Source	*SS*	*df*	*MS*	*F*	*P*
Between	.1260	1	.1260	.96	.3347
Within	4.4749	34	.1316		
Total	4.6010	35			

Table E15

Demonstration of Comparison of Best CBC Ratios for Aggregation Procedure E (Low) with Summary ANOVA Information

Top Performing Strategies	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1, 34)
Domain-independent web catalog: Yahoo	94.4%	169.8	.62389	.96
Domain-independent web database: Medical World Search	83.3%	166.8	.50111	

Note. Ans. = Answers; Sec. = Seconds.

Table E15 (cont.)

	Summary ANOVA				
Source	*SS*	*df*	*MS*	*F*	*p*
Between	.1357	1	.1357	.96	.3334
Within	4.7902	34	.1409		
Total	4.9259	35			

Table E16

Demonstration of Measurement of Web Catalog Performance for Aggregation Procedure E (Medium) with Summary ANOVA Information

Web Catalog	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1, 34)
Domain-specific: Medical Matrix	61.1%	237.4	.35611	.03
Domain-independent: Yahoo	61.1%	234.1	.33500	

Note. Ans. = Answers; Sec. = Seconds.

	Summary ANOVA				
Source	*SS*	*df*	*MS*	*F*	*p*
Between	.0040	1	.0040	.03	.8682
Within	4.8765	34	.1434		
Total	4.8805	35			

Table E17

Demonstration of Measurement of Web Database Performance for Aggregation Procedure E (Medium) with Summary ANOVA Information

Web Database	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1, 34)
Domain-specific: Medical World Search	94.4%	179.9	.49961	1.59
Domain-independent: Alta Vista	55.6%	204.0	.34222	

Note. Ans. = Answers; Sec. = Seconds.

	Summary ANOVA				
Source	*SS*	*df*	*MS*	*F*	*p*
Between	.2229	1	.2229	1.59	.2165
Within	4.7795	34	.1406		
Total	5.0024	35			

Table E18

Demonstration of Comparison of Best CBC Ratios for Aggregation Procedure E (Medium) with Summary ANOVA Information

Top Performing Strategies	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1, 34)
Domain-specific web catalog: Medical Matrix	61.1%	237.3	.35611	
Domain-independent web database: Medical World Search	94.4%	179.8	.49961	1.35

Note. Ans. = Answers; Sec. = Seconds.

Table E18 (cont.)

Source	*Summary ANOVA*				
	SS	*df*	*MS*	*F*	*p*
Between	.1853	1	.1853	1.35	.2534
Within	4.6673	34	.1373		
Total	4.8527	35			

Table E19

Demonstration of Measurement of Web Catalog Performance for Aggregation Procedure E (High) with Summary ANOVA Information

Web Catalog	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1, 34)
Domain-specific: Medical Matrix	16.7%	283.8	.11111	
Domain-independent: Yahoo	33.3%	260.8	.21278	.95

Note. Ans. = Answers; Sec. = Seconds.

Source	*Summary ANOVA*				
	SS	*df*	*MS*	*F*	*P*
Between	.0930	1	.0930	.95	.3361
Within	3.3217	34	.0977		
Total	3.4148	35			

Table E20

Demonstration of Measurement of Web Database Performance for Aggregation Procedure E (High) with Summary ANOVA Information

Web Database	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1, 34)
Domain-specific: Medical World Search	55.6%	241.3	.27722	
Domain-independent: Alta Vista	72.2%	184.7	.35333	.46

Note. Ans. = Answers; Sec. = Seconds.

	Summary ANOVA				
Source	*SS*	*df*	*MS*	*F*	*p*
Between	.0521	1	.0521	.46	.5009
Within	3.8290	34	.1126		
Total	3.8811	35			

Table E21

Demonstration of Comparison of Best CBC Ratios for Aggregation Procedure E (High) with Summary ANOVA Information

Top Performing Strategies	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1, 34)
Domain-independent web catalog: Yahoo	33.3%	260.83	.21278	
Domain-independent web database: Alta Vista	72.2%	184.72	.35333	1.49

Note. Ans. = Answers; Sec. = Seconds.

Table E21 (cont.)

Source	*Summary ANOVA* *SS*	*df*	*MS*	*F*	*p*
Between	.1778	1	.1778	1.49	.2305
Within	4.0546	34	.1193		
Total	4.2324	35			

Table E22

Demonstration of Measurement of Web Catalog Performance for Aggregation Procedure F (Low) with Summary ANOVA Information

Web Catalog	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1, 31)
Domain-specific: Medical Matrix	84.6%	188.38	.61462	1.16
Domain-independent: Yahoo	85.0%	204.10	.47200	

Note. Ans. = Answers; Sec. = Seconds.

Source	*Summary ANOVA* *SS*	*df*	*MS*	*F*	*p*
Between	.1602	1	.1602	1.16	.2888
Within	4.2644	31	.1376		
Total	4.4247	32			

Table E23

Demonstration of Measurement of Web Database Performance for Aggregation Procedure F (Low) with Summary ANOVA Information

Web Database	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1, 37)
Domain-specific: Medical World Search	86.9%	175.17	.53565	.11
Domain-independent: Alta Vista	93.7%	148.94	.49625	

Note. Ans. = Answers; Sec. = Seconds.

	Summary ANOVA				
Source	*SS*	*df*	*MS*	*F*	*P*
Between	.0146	1	.0146	.11	.7457
Within	5.0753	37	.1372		
Total	5.0900	38			

Table E24

Demonstration of Comparison of Best CBC Ratios for Aggregation Procedure F (Low) with Summary ANOVA Information

Top Performing Strategies	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1, 34)
Domain-specific web catalog: Medical Matrix	84.6%	188.38	.61462	.35
Domain-independent web database: Medical World Search	86.9%	175.17	.53565	

Note. Ans. = Answers; Sec. = Seconds.

Table E24 (cont.)

Source	***Summary ANOVA***				
	SS	*df*	*MS*	*F*	*P*
Between	.0518	1	.0518	.35	.5569
Within	5.0023	34	.1471		
Total	5.0541	35			

Table E25

Demonstration of Measurement of Web Catalog Performance for Aggregation Procedure F (Medium) with Summary ANOVA Information

Web Catalog	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1, 40)
Domain-specific: Medical Matrix	33.3%	256.09	.29909	
Domain-independent: Yahoo	61.1%	225.95	.37450	.37

Note. Ans. = Answers; Sec. = Seconds.

Source	***Summary ANOVA***				
	SS	*df*	*MS*	*F*	*P*
Between	.0596	1	.0596	.37	.5475
Within	6.4751	40	.1619		
Total	6.5346	41			

Table E26

Demonstration of Measurement of Web Database Performance for Aggregation Procedure F (Medium) with Summary ANOVA Information

Web Database	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1, 28)
Domain-specific: Medical World Search	83.3%	229.69	.31408	.27
Domain-independent: Alta Vista	61.1%	212.06	.24941	

Note. Ans. = Answers; Sec. = Seconds.

	Summary ANOVA				
Source	*SS*	*df*	*MS*	*F*	*P*
Between	.0308	1	.0308	.27	.6104
Within	3.2490	28	.1160		
Total	3.2798	29			

Table E27

Demonstration of Comparison of Best CBC Ratios for Aggregation Procedure F (Medium) with Summary ANOVA Information

Top Performing Strategies	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1, 31)
Domain-independent web catalog: Yahoo	61.1%	225.95	.37450	.19
Domain-specific web database: Medical World Search	83.3%	229.69	.31408	

Note. Ans. = Answers; Sec. = Seconds.

Table E27 (cont.)

Source	*Summary ANOVA*				
	SS	*df*	*MS*	*F*	*p*
Between	.0288	1	.0288	.19	.6658
Within	4.6890	31	.1513		
Total	4.7177	32			

Table E28

Demonstration of Measurement of Web Catalog Performance for Aggregation Procedure F (High) with Summary ANOVA Information

Web Catalog	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1, 31)
Domain-specific: Medical Matrix	61.1%	256.21	.20158	
Domain-independent: Yahoo	66.7%	240.43	.29714	.54

Note. Ans. = Answers; Sec. = Seconds.

Source	*Summary ANOVA*				
	SS	*df*	*MS*	*F*	*P*
Between	.0736	1	.0736	.54	.4663
Within	4.1935	31	.1353		
Total	4.2672	32			

Table E29

Demonstration of Measurement of Web Database Performance for Aggregation Procedure F (High) with Summary ANOVA Information

Web Database	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1, 37)
Domain-specific: Medical World Search	83.3%	198.39	.36667	.04
Domain-independent: Alta Vista	83.3%	199.67	.34429	

Note. Ans. = Answers; Sec. = Seconds.

	Summary ANOVA				
Source	*SS*	*df*	*MS*	*F*	*P*
Between	.0049	1	.0049	.04	.8403
Within	4.3595	37	.1178		
Total	4.3644	38			

Table E30

Demonstration of Comparison of Best CBC Ratios for Aggregation Procedure F (High) with Summary ANOVA Information

Top Performing Strategies	% Ans. Found Within 5 Min	Mean Time on Task (Sec.)	Mean CBC Ratio	F (1, 30)
Domain-independent web catalog: Yahoo	66.7%	240.43	.29714	
Domain-specific web database: Medical World Search	83.3%	198.39	.36667	.27

Note. Ans. = Answers; Sec. = Seconds.

Table E30 (cont.)

Source	***Summary ANOVA***				
	SS	*df*	*MS*	*F*	*p*
Between	.0381	1	.0381	.27	.6080
Within	4.2665	30	.1422		
Total	4.3046	31			

REFERENCES

Belkin, N. J. (1978). Information concepts for information science. Journal of Documentation, 34, 55-85.

Bishop, A. P. & Star, S. L. (1996). Social informatics of digital library use and infrastructure. In M. E. Williams (Ed.), Annual review of information science and technology, (pp. 301-400). Medford, NJ, Information Today.

Brookes, B. C. (1980a). The foundations of information Part I. Philosophical aspects. Journal of Information Science, 2, 125-133.

Brookes, B. C. (1980b). Measurement in information science: Objective and subjective metrical spaces. Journal of the American Society for Information Science, 31, 248-255.

Bruce, H. (1998). User satisfaction with information seeking on the Internet. Journal of the American Society for Information Science, 49, 541-556.

Cleverdon, C. W. (1962). Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. Unpublished manuscript.

Cleverdon, C. W., Mills, J., and Keen, E .M. (1967). Factors determining the performance of indexing systems. Unpublished manuscript.

Core Content Review of Family Medicine Executive Committee. (1998). The Core Content Review of Family Medicine, 1997-98 Edition. Connecticut and Ohio Academies of Family Physicians.

Covi, L. & Kling, R. (1996). Organizational dimensions of effective digital library use: Closed rational and open natural systems models. Journal of the American Society for Information Science, 47, 672-689.

Crocker, L. & Algina, J. (1986). Introduction to Classical and Modern Test Theory. New York: Holt, Rinehart, & Winston.

Cuadra, C. A. & Katter, R. V. (1967). Opening the black box of “relevance”. Journal of Documentation, 23, 291-303.

Dervin, B. (1997). Given a context by any other name: methodological tools for taming the unruly beast. In Vakkari, P., Savolainen, R. & Dervin, B. (Eds.), Information Seeking in Context. Proceedings of the International Conference on Research in Information Needs, Seeking and Use in Different Contexts, (pp. 13-38). London: Taylor Graham.

Dervin, B. (1994). Information-democracy: An examination of underlying assumptions. Journal of the American Society for Information Science, 45, 369-85.

Dervin, B. (1992). From the mind's eye of the "user:" The sense-making qualitative/quantitative methodology. In: J.D. Glazier & RR. Powell (Eds.), Qualitative research in Information Management, (pp. 61-84), Englewood, CO: Libraries Unlimited.

Dervin, B. (1977). Useful theory for librarianship: communication, not information. Drexel Library Quarterly, 13, 16-32.

Dervin, B. & Dewdney, P. (1986). Neutral questioning: a new approach to the reference interview. RQ, 50, 6-13.

Dervin, B. & Nilan, M. (1986). Information needs and uses. In M. E. Williams (Ed.), Annual review of information science and technology, (pp. 3-33). White Plains, NY: Knowledge Industry.

Ding, W. & Marchionini, G. (1996). A comparative study of web search service performance. Proceedings of the 59th American Society for Information Science Annual Conference 33, 136-142.

Ellis, D. (1996). The dilemma of measurement in information retrieval research. Journal of the American Society for Information Science, 47, 23-36.

Ellis, D. (1994). Paradigms in information retrieval research. In A. Kent (Ed.). Encyclopedia of Library and Information Science, (Vol 54, pp. 275-290). New York: Marcel Dekker.

Ellis, D. (1984). Theory and explanation in information retrieval research. Journal of Information Science, 8, 25-38.

Eisenberg, M. B. (1988). Measuring relevance judgments. Information Processing and Management, 24, 373-389.

Fidel, R. (1987). What is missing in research about online searching behavior? The Canadian Journal of Information Science, 12, 54-61.

Friedman, C. P. & Wyatt, J. (1997). Evaluation Methods in Medical Informatics. New York: Springer.

Harmon, D. K. (1995). Overview of the Second Text Retrieval Conference (TREC-2). Information Processing and Management, 31, 271-289.

Harter, S. P. (1996). Variations in relevance assessments and the measurement of retrieval effectiveness. Journal of the American Society for Information Science, 47, 37-49.

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

Harter, S. P. & Hert, C. A. (1997). Evaluation of information retrieval systems: approaches, issues, and methods. In M. E. Williams (Ed.), Annual review of information science and technology, (pp. 3-94). Medford, NJ: Information Today.

Hawking, S. W. (1988). A brief history of time: From the big bang to black holes. New York: Bantam Books.

Hersh, W., Pentecost, J. & Hickam, D. (1996). A task-oriented approach to information retrieval evaluation. Journal of the American Society for Information Science, 47, 50-56.

Hilz, S. R. & Johnson, K. (1989). Measuring acceptance of computer-mediated communication systems. Journal of the American Society for Information Science, 40, 386-397.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. Journal of Documentation, 52, 3-50.

Karamuftuoglu, M. (1998). Collaborative information retrieval: toward a social informatics view of IR interaction. Journal of the American Society for Information Science, 49, 1070-1080.

Kranz, D. H., Luce, D. R., Suppes, P., & Tversky, A. (1971). Foundations of measurement, (Vol. 1). New York: Academic Press.

Lamb, R. (1995, June). Using online information resources: Reaching for the *.*’s. Paper presented at the 1995 Digital Libraries Conference, Austin, TX. Available: http://csdl.tamu.edu/DL95/papers/lamb/lamb.html [1999, March 26].

Leighton, H. V. & Srivastava, J. (1997). Precision among world wide web search services (search engines): Alta Vista, Excite, Hotbot, Infoseek, Lycos. Unpublished paper. Available: http://www.winona.msus.edu/is-f/libraryf/webind2/webind2.htm [1999, March 26].

Lesk, M. E. & Salton, G. (1968). Relevance assessments and retrieval system evaluation. Information Storage and Retrieval, 4, 343-359.

MacRoberts, M. H. & MacRoberts, B. R. (1989). Problems of citation analysis: A critical review. Journal of the American Society for Information Science, 40, 342-349.

Marchionini, G., Barlow, D. & Hill, L. (1994). Extending retrieval strategies to networked environments: Old ways, new ways, and a critical look at WAIS. Journal of the American Society for Information Science, 45, 561-564.

Neill, S. D. (1992). Dilemmas in the study of information. New York: Greenwood Press.

Paisley, W. J. (1980). Information and work. In: B. Dervin & M.J. Voigt (Eds.) Progress in the Communication Sciences. (Vol. 2, pp. 114-165). Norwood, NJ: Ablex.

Poulter, A. (1997). The design of World Wide Web search engines: A critical Review. Program, 31, 131-245.

Rees, A. M. & Saracevic, T. (1966). The measurability of relevance. Unpublished manuscript, Western Reserve University.

Reichenbach, H. (1958). The philosophy of space and time. Toronto: Dover.

Roland, M. O. Bartholomew, J., Courtenay, M. J. F., Morris R. W. & Morrell D. C. (1992). The "five minute" consultation: Effect of time constraint on verbal communication. BMJ, 292, 874-876.

Rowland, F., McKnight, C. & Meadows, A. J. (1995). Project ELVYN: An experiment in electronic journal delivery: Facts, figures and findings. London: Bowker-Sauer.

Salton, G. (1992). The state of retrieval evaluation. Information Processing and Management, 28, 441-449.

Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. Journal of the American Society for Information Science, 26, 321-343.

Schamber, L., Eisenberg, M. B. & Nilan, M. S. (1990). A re-examination of relevance: Toward a dynamic, situation definition. Information Processing and Management, 26, 755-776.

Schwartz, C. (1998). Web search engines. Journal of the American Society for Information Science, 49, 973-982.

Sparck Jones, K. (1995). Reflections on TREC. Information Processing and Management, 31, 291-314.

Tague-Suttcliffe, J. M. (1996). Some perspectives on the evaluation of information retrieval systems. Journal of the American Society for Information Science, 47, 1-3.

Van House, N. A., Butler, M. H., Ogle, V., & Schiff, L. (1996, February). User-centered interative design for digital libraries [Online]. D-Lib Magazine. Available: http://www.dlib.org/dlib/february96/02vanhouse.html [1999, March 26].