AN EXAMINATION OF EVALUATION METHODS FOR COMPARING TWO
INFORMATION RETRIEVAL SYSTEMS SUPPORTING TEACHER
PERFORMANCE

Steven Craig Schatz

Submitted to the faculty of the School of Education in partial fulfillment of the
requirements for the degree
Doctor of Philosophy
In the Department of Instructional Systems Technology

April, 2004
Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Thomas M. Schwen, Ed.D.

Curtis J. Bonk Ph.D.

Javed Mostafa Ph. D.

Martin A. Siegel Ph. D.

April 22, 2004
ACKNOWLEDGEMENTS

Here I stand with my piccolo
but the orchestra plays much more sweetly…
My thanks is humbly offered:

- To my committee – Tom, Javed, Curt, and Marty, who have guided me through the years leading up to this process and through this long, messy labor. Like midwives everywhere, I made the problems and you were there to help me push, ‘though you rarely got more than cursing for your efforts.

- To Mike for many moments of calm guidance. Your words helped me learn to be a student and a scholar.

- To Bill and Bob, who went there first.

- To Gene, for two very well timed and well placed boots.

- To Sasha, who got me started on the right track.

- To Rob, for wicked humor and deep understanding. I miss you.

- To Lee, for understanding my ramblings and making these things work.

- To Harvi for starting me out and Javed for helping me understand and continually sharing insights while keeping me going.

- To Tom, for friendship, guidance, too many books, and many hours of wonderful chats – I guess you finally got me hooked on this scholarship thing.

- Dr. Ajit Mukherjee - for guiding me through the statistics and accomplishing the impossible – helping me understand what the statistics say.

- To William Cooper for an idea too long kept waiting.

- To Mom and Dad, for continued love, support, guidance, and the belief that I could do this, even when I doubted it. You each made this possible.

- To Susan, who has been through so much for so many years…with great love. There ARE a million things to be – aren’t there?

- And finally, to Paul, who came all the way to Indiana for no good reason except so we could stay together…who was there through the tensions, fears, frustrations, and silences with your quiet, solid, love. You have been entirely patient with me & incredibly good. ..I don’t think two people could have been happier than we have been. S.
<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation</td>
<td>89</td>
</tr>
<tr>
<td>Chapter 6: Conclusions</td>
<td>91</td>
</tr>
<tr>
<td>Chapter 7: Implications</td>
<td>99</td>
</tr>
<tr>
<td>Limitations</td>
<td>99</td>
</tr>
<tr>
<td>Future Studies</td>
<td>102</td>
</tr>
<tr>
<td>A New Era of Evaluation</td>
<td>103</td>
</tr>
<tr>
<td>Evaluation Tool</td>
<td>105</td>
</tr>
<tr>
<td>Considerations of Tag-Based Search</td>
<td>106</td>
</tr>
<tr>
<td>Applications to the Semantic Web</td>
<td>109</td>
</tr>
<tr>
<td>References</td>
<td>112</td>
</tr>
</tbody>
</table>
Chapter One: Introduction

Situating the Research I:

Learning Objects – Beneath the Sizzle, a Technology for More Efficient Search

Learning objects have stirred great controversy in the educational community over the past few years. Some embrace the idea as a new wave in education, allowing customization of instruction on the fly. Others flatly reject the concept of objects. If learning is constructed (either internally or socially) (Burton, Brown, & Fischer, 1999; Hogan & Tudge, 1999; Rogoff, 1990; Rogoff & Lave, 1984), then the idea of objectified learning contradicts over two decades of research (more if we go back to the writings of Dewey and Vygotsky’s original writings). Instead of the widely accepted notions of constructivist learning environments, learning objects seem to be a throwback to programmed instruction. However, learning objects are not the problem themselves. It is their intended use.

What is a Learning Object

The bane and blessing of learning objects is that the term seems self-explanatory. However, there are many different definitions of various degrees of vagueness in play. The IEEE LOM defines learning objects as, “Any entity, digital or non-digital, which can be used, re-used or referenced during technology supported learning” (IEEE_LOM, 2002 para.1). Wiley says, “Learning objects are elements of a new type of computer-based instruction grounded in the object-oriented paradigm of computer science” (Wiley, 2001 p.3). Cisco systems says, “a learning object is based on a single learning or performance objective, built from a collection of static or interactive content and instructional practice activities” (Ying, 2002 para.3).
At the 2003 ELearn conference in Phoenix, there were 22 sessions that mentioned learning objects. These sessions seemed to fall into three categories. The first were those involved with specifications. These groups focused on developing new sets of tags for learning objects that allowed for new applications (such as access for disabilities) and building libraries of tagged objects. These groups worked with prototypes and test beds of a limited number of objects to prove that it was possible to show the same information in different ways depending on the needs of the user or that it was possible to have teachers tag objects and put them in a repository. These groups did not address the pedagogic challenges to learning objects, nor the disconnect between the time necessary to understand tag sets and to tag objects and the reality of teachers with limited time. In one session, a spokesperson for a granting agency that was encouraging teachers to tag objects admitted, “There is a huge gap between the learning object repository and the reality of the classroom.” The second group of presentations was those I will call marketers. These presenters were using learning objects as the latest “hot topic du jour”. Just as everything a few years ago was suddenly multimedia and, more recently, had an elearning component, these people were touting what were essentially libraries of templates and some content as learning objects. The final group of presentations was reports by practitioners. These people are using learning objects to build curriculum. They are using learning objects as pieces of instruction, just as one would use a video, a journal article, a web site, or any other media. The context is provided by the teacher and the student. The learning object is information.

This final use provides an answer to the tension poised by one participant who said, “The power of learning objects is that they can be decontextualized. This is what
allows them to be used in different places. As educators, we know that context is
everything. So, there is a fundamental tension between the power of learning objects and
their use”(McGee, 2004). If the teacher or instructional designer provides the context, the
learning object is merely a tool, a bit of information. The pedagogic problem lies in the
claim that objects may be somehow strung together automatically based on some criteria.
The problem is with the intended use. It is not with the object – any more than there is a
problem with locating information in a book or a journal article.

What is in a name? A learning object is nor hand, nor foot, nor arm, nor face, nor
any other part belonging to a teaching (Apologies to Shakespeare). If the problem is with
intended use, let us look for a more useful, perhaps less explosive definition. Instead of
defining by use, let us define by form. Very simply, we may define a learning object (or
information object, as I prefer to call them, to further remove the emotional baggage of
the term), as an object that may be accessed online by means of some kind of tags. The
learning is directed by the learner and the instructor. The distinguishing feature of an
information object intended to support performance or learning is that someone has added
some kind of information (called meta-tags or meta-data, which is simply data about data
– some kind of description of the object) to aid in the search and retrieval of the object. In
other words, if one removes the dubious pipe dream of automatic construction of
instruction, the efforts surrounding learning objects are to one end – making it faster and
easier to find and retrieve useful materials that can support instruction. This in no way
demeans the efforts toward developing meta-tags and tagging objects. Anyone who has
tried to find useful information via search engines knows the frustration of information
overload. If, by spending time and energy on adding meta-data to objects, it is possible to
reduce the time required to sift through the chaff of a web search, a great service is provided.

So, if a learning object is an object that someone has tagged to make it easier to find so it may be used by educators, then all of the learning object initiatives are efforts to make it easier to find and use information available. If we focus on form, toward the end of enhancing finding and retrieving, then it behooves us to explore different methods of tagging objects and other methods to improve search and retrieval.

Situating the Research II:

Human Performance Technology and Performance Support Systems

Researchers in instructional systems technology have long studied more than just instruction. A particularly robust sub domain is human performance technology (HPT). Performance improvement focuses on results, not activities. Training is merely one class of intervention. The goal in any intervention is to improve performance. The result provides the measure of success. Tom Gilbert (Gilbert, 1996) outlines three areas to turn to when seeking to improve performance – information, instrumentation, and motivation.

To engineer performance, it makes sense to try to use strategies that have the greatest impact for the least cost. He suggests looking at these three variables in the environment.

Begin with the data. Ask if they are a sufficient, informative, and reliable guide both to how one should perform and to how well one has performed. Next, examine the tools and materials people have to work with. Next, look at incentives. Finally, though not least in importance: look at training as a means to achieve greater competence. It is often a powerful strategy, but usually very expensive. (p. 91)

So, according to the HPT view, one may have a greater impact on improved performance by providing better information during the course of performance. Rossett (Rossett, 1996) points out that training tries to build capacity, occurring before a need
arises. Performance interventions are used as needed. As information provided just in
time offers a powerful tool for improving outcomes, a class of performance interventions
has concerned itself with building computer-based systems that can provide just in time
information. This class of interventions, variously called electronic performance support
systems (EPSS), performance support systems (PSS), and performance support tools (of
which a performance portal is a subset), have been in use since the 1980s. One may view
these systems on a continuum of how embedded they appear. A classic view of a highly
embedded or intrinsic EPSS is the Turbo Tax software, where the system asks a series of
questions and does a series of calculations to help the user complete a tax form. Help
functions, both ones that attempt to respond based on user actions and those that provide
cue cards and other help only upon request are a more extrinsic PSS. Performance
Portals, which provide a single point of access to training, information, and support tools,
are completely extrinsic, stand-alone systems. Saul Carliner (Carliner, 2002), introducing
a special issue on EPSSs, defined them as, “part online help, part online tutorial, part
database, part application program, and part expert system...EPSSs quickly and easily
provide answers to the questions workers have when performing a job, and address
workers’ concerns” (p. 400).

While performance systems are designed to support the goal of improved
performance, they certainly support learning, too. “In a well-designed PS (performance
support) system, learning is likely, desirable, even inevitable, but it’s not the point.
Performance is the point.” (Dickelman, 2000  p.8). Rosenberg makes the point even more
strongly, saying that EPSS is a paradigm shift, bringing learning and work
together(Rosenberg, 1995). He says the people developing EPSSs:
...wanted to obliterate the line between learning and work so that, in reality, learning is work, and work is learning...We see an integration of information and learning, and gain an understanding that knowledge may sometimes be more effectively and efficiently delivered...than through instruction...It is not appropriate to say that EPS will eliminate the need for education and training, for there will always be a need for new knowledge and continuous learning. But using EPS instead of training for disseminating facts and procedures frees up training resources for more sophisticated efforts in areas...better suited to the strengths of the educational model. (p. 92-93)

While there is a great diversity of functionalities across EPSS applications, one of the most common functions is a database of information that may be searched. In more embedded systems, this search is less obvious, often completed by the system based on context. However, in the more extrinsic systems, the search/retrieval function is an obvious and essential part of the system. Quesenbery identified four technologies of performance support (Quesenbery, 2002), intelligent agents, information visualization, search engines, and collaborative filtering (which is a variation of information retrieval using methods other than text search). An article in the Communications of the ACM (Cole, Fischer, & Saltzman, 1997) states: “The primary design goal of an EPSS is that the knowledge it contains be easily retrievable by the users at the time they need it” (p. 50). A survey of readers of CBT Solutions magazine conducted in 1996 listed “searchable reference” as the most common feature of EPSSs either built or being built for their companies. (Benson, 1997). The same is true with performance portals. Nearly all of the over 50 demonstration sites listed by the portal software maker Plumtree (www.plumtree.com) contain a search function in a prominent location.

However, while the search/retrieval function or system has been shown to be an important function of performance systems, there is little known about within the field of instructional technology about evaluation of information retrieval systems. This is a
problem, as there have been ongoing attempts to improve search systems. In 1992, Carr wrote, “Despite years of work, our methods for retrieving information from data bases remain relatively rigid and primitive” (Carr, 1992 p. 35). Much of the work concerning meta tags (Dublin core, IMS, SCORM, GEM and more) is predicated on the idea that adding meta tags can improve retrieval systems. Unfortunately, without being able to measure the effect of different search/retrieval systems, there is no way to know if efforts to improve practice are yielding results. The goal of this research is to see how the constructs and measures of information retrieval evaluation may inform the research agenda of performance support and if new constructs and measures may enrich and expand the theory.

Evaluation of information retrieval systems (whose practitioners are usually located in schools of library and information science) is a relatively mature field, with a literature dating to the mid 1960's (further if one considers pre-computer methods). While this literature offers a rich foundation, the constructs and measures underlying the theory are thin. For one wishing to evaluate information retrieval within performance support systems, additional constructs and measures that reflect the unique needs of performance support are necessary.

The traditional method for evaluating information retrieval systems relies on the relevance based measures, recall and precision. To accomplish this type of evaluation requires:

- a collection of documents
- a collection of questions (queries) to be asked of the document collection
- a set of judgments of which documents are relevant to each question.
To evaluate the system, one queries the document collection (ask the questions and see what documents are retrieved) and calculates recall and precision. Recall is the measure of how many relevant documents were actually retrieved. For example, if one question was 'How many angels can dance on the head of a pin?' and upon studying the document collection, it was judged that there were 50 documents that were relevant to that question, and a system retrieved 20 of those documents when it searched for 'How many angels...?', then the recall rating for that system would be the number of relevant documents retrieved divided by the total number of relevant documents expressed as a percentage. (In this case 20/50 or 40%.) Precision is the measure of how many relevant documents were retrieved divided by the total number of retrieved documents. It is a measure of how “noisy” the results are. It would be easy to retrieve the 50 relevant documents – just retrieve all the documents in the collection. A system that retrieved 30 documents of which 20 were relevant to the question is of more value than a system that retrieves 100 documents of which 30 are relevant. While the recall of the second system is higher (60% vs. 40%), the precision is lower (System 1 precision = 20/30 = 66% and System 2 precision = 30/100 = 30%). Both measures are reported in the traditional evaluation paradigm. They are generally assumed to be inversely proportional – the better the recall (the more relevant items retrieved) the lower the precision (the “noisier the result).

The traditional method of evaluation is system-centric. It is possible to test several information retrieval systems using a batch mode – using a program that runs all the queries against a system and calculates the recall and precision of the system. There is no consideration of users. The focus is on the system.
Statement of Traditional theory

While the theory underlying information retrieval is rather thin, it has been the basis of hundreds of research papers. The theoretical constructs, measures, relationships between constructs and assumptions of relevance-based evaluation are:

Assumptions

- Relevance judgments are valid and meaningful indicators of the effectiveness of an information retrieval system. (Harter & Hert, 1997)
- Relevance is a meaningful construct on which to base the evaluation process.
- Systems can be evaluated using the combined measures of recall and precision.
- A successful retrieval system retrieves the greatest number of relevant documents with the least extraneous documents.

Constructs

- Relevance – a judgment of whether a particular object answers a particular question.
- Precision – how many extraneous or non-relevant documents are retrieved for a particular question by a particular system. A measure of accuracy or quality of retrieval.
- Recall – how many documents drawn from a particular set of documents relevant to a particular question are retrieved with a particular system. A measure of scope or quantity of retrieval.

Relationships between Constructs
• Recall and Precision have an inverse relationship. Generally, the greater the recall (number of relevant documents retrieved), the lower the precision (more noise).

Measures

• Relevance judgments can be made through a variety of methods, including automatic text matching and expert judgments.

• Recall is measured by dividing the number of documents that have been judged relevant to a particular query in a document set by the number of relevant documents retrieved for that query.

• Precision is measured by dividing the number of documents retrieved for a query by the number of relevant documents retrieved.

It is an indication of the state of the theory of evaluation that recall, precision, and relevance are both constructs and measures. This encourages confusion. In this research, I will explicitly refer to the recall measure, the recall construct, the precision measure, the precision construct, the relevance construct, and the relevance measure. To provide an example of the confusion inherent in these discussions, we may look at the relevance construct. Relevance is a construct that underlies the way the measures recall and precision are calculated. However, there are fundamental concerns about the relevance measure. All the discussions that follow detailing concerns with relevance judgments are actually concerns about the relevance measure. These concerns, then affect the interpretation of recall and precision constructs. There is not a problem with the measures of recall and precision, the actual formulas make sense. However, the constructs are brought into question because of questions surrounding the relevance measures.
There are additional problems with the theory of evaluation when it is used to evaluate performance systems. “The main assumption behind the use of measures (constructs) such as recall and precision is that the average user is interested in retrieving large amounts of relevant materials..., while at the same time rejecting a large proportion of the extraneous items ...” (Salton, 1992 p. 442). For a performance system that has the goal of providing a fast, specific answer to a problem, the best result is a small number of items with high precision. This is a fundamental difference between an information retrieval system (like a web search engine) and a performance support system.

Research questions

The current state of theory for the evaluation of information retrieval systems is not rich enough to easily map to performance systems. Some have said it does not map all that well to information retrieval systems (Cooper, 1981; Harter, 1996; Harter & Hert, 1997; Schamber, 1994). In this study, I have used a case study to examine two new constructs and their measures as well as the traditional constructs and their measures to attempt to obtain a richer view of comparative evaluation of two systems. The research questions are:

- Can evaluation of information retrieval be improved through the use of multiple constructs?

- Can user-centric constructs and measures provide time and cost effective comparative evaluation of two information retrieval systems?

- Are the non-relevance based constructs of information need and utility useful for evaluation of information retrieval systems?
While information retrieval is a necessary part of design and development of a performance support system, it is not an area of research common to instructional systems. The research traditions within instructional systems can bring insight into the evaluation of information retrieval systems. The tradition of research of evaluation of information retrieval systems certainly provides pertinent guidance for instructional/performance systems designers. However, as this is a study within the purview of instructional systems, a bit of background in the field of information retrieval is necessary. Following are two sections. The first provides an overview of what search engines do and the second details three eras of research and development involving information retrieval.

What are search engines

We use search engines daily. A search engine is an information retrieval system. However, to put this study in context, it is useful to explain briefly some of the workings of a search engine. First, let us distinguish between types of document sets. The most common engines most of us use are web-based systems, such as Google or Yahoo or Excite. These search engines search the web, which is an information or document set that is large and constantly growing. The growth of the web is not controlled by the search engines or system designers. This is in contrast to information systems such as ERIC or Lexus/Nexus that have documents or document summaries that have been added into a database by someone associated with the service. While the web is often referred to as an entity, it is an uncontrolled source. Services such as ERIC or Lexus/Nexus are controlled – the contents and access are controlled by systems designers.
A basic difference between web-based and controlled search/retrieval systems is the source of their documents. One of the challenges for web-based systems is keeping abreast of new web-based documents. Web search systems have one part called a spider or crawler. This is an automated program that goes out on the web, following links and reading web pages. Controlled systems do not find their document set in this manner; they are added to the systems by the system maintainers based on the system’s set of criteria.

When the crawler or a controlled system reads a document, the system then breaks the documents into words or phrases and constructs an index. This index contains a list of vocabulary words and a list of all documents that contain that word (or phrase). So, when you search (for example for “horse”), the system is not actually searching the documents, it is searching its own index. This is much faster.

The third part of a search engine is a set of decisions about weighting or ranking of documents. If you search for horse, the engine may return 1,000,000 records. How does it decide which ones to display first? Perhaps you have entered “horse race” and are interested in how to become a jockey. The words horse and race may return documents on horse breeding, horse blankets, dude ranches, race relations, and many other items you do not want. The reason you will get different results using different search engines is that they use different decision sets to index the documents and to weight the documents, deciding which documents display toward the top of the list. It is good to keep in mind that these are all automatic processes, designed to handle large numbers of documents.

So, when evaluating information retrieval systems, what we are evaluating, particularly in the traditional methodology, is how the system divides the documents
(parses the document), how it constructs the index, and the decision set for ordering the results.

History of Information Retrieval Evaluation

The focus of this section is on electronic information retrieval. Of course, there has been information retrieval of a sort since there have been stockpiles of information (libraries). Saracevic writes that “Information science is a field and a subject that is concerned with problems arising in communication of knowledge in general and with records in such communication in particular (Saracevic, 1975 p.322). Information systems emerged as tools of scientific communication. With the ability to search and retrieve information came the desire to know if one method of search and retrieval is better than another. In this paper, I delineate three conceptual eras in the history of search/retrieval systems. While the boundaries of these three eras are not stark, this lens helps to understand the history. For each era, there are two foci – technical and theoretical. While the technological foci do follow a linear, temporal course, the boundaries of the theoretical considerations are not as clear. None of these theoretical foci has a clear point in time where one may point and say, “It began here.” None of the theoretical foci has ended, they are just reduced in primacy. However, with the reader’s indulgence, this division should prove to be worthwhile in making sense of a short summary of 50 years of research and development. The purpose of this history is to attempt to illuminate the backdrop against which the theoretical manifestations take place.

The modern age of evaluation dates back to the late 1950s. Sparck Jones references the 1958 International Conference on Scientific Information as the “Beginning
of a new era in information processing” (Sparck Jones, 1981). Others point to the experiments by Cleverdon at Cranfield, known as Cranfield 2, that were published in the early to mid 1960s (Cleverdon, 1962; Cleverdon, Mills, & Keen, 1966).

Cranfield established several important foci, many of which are still used today. These include a system view for evaluation, a reliance on quantitative, experimental methods, and the foundational belief that the measure of success is the retrieval of the most relevant documents. The system is viewed as a black box. To evaluate systems, you need a set of questions, a set of documents, and a set of judgments as to the relevancy of the documents to the questions. If you have those things, you can write a program that uses different systems to query the documents. Then, it is possible to calculate the recall and precision measures for each system. These constructs and measures test the efficacy of the system – did it retrieve as many as possible relevant documents with as few extraneous documents as possible? Relevance was often decided by judges (as opposed to automatic word matching) and was always done prior to testing.

This traditional theoretical view continues to hold sway in much of the information retrieval community. The development of the theory was driven by technological advances in data storage and data processing. With the advent of relatively reliable computers, it was possible to store large amounts of data. However, the challenges inherent in creating these systems led naturally to the system centric view of evaluation. The question, on the “bleeding edge” of technology was, “Can this be done?”, not “How do we make this work for users?” The explorers into these realms unknown, like trappers of old, were rarely particularly interested in or good at the niceties of human
interaction. The questions they focused on were the needs of the system. Those who used the system were really viewed as no more than a type of input/output device.

Just as explorers inevitably had to give way to townspeople, who could never survive without stores, power, and utilities, a new focus was necessary as systems become used, to meet the needs of a new type of users who were less hardy and resilient. During this time, when computer systems were relatively new, most information systems were not available to the general public. A user went to a special “temple” and talked to the “priest” who talked to the system or computer and brought back their information. While these systems were computer-based, the data in them was rarely full text and almost never anything but text. The documents were represented by abstracts, indices, or phrases representing documents. Storage was still expensive and online versions of journals were rare. So, most systems returned references to paper journals.

The second conceptual era was, like the first, theory based, but was driven by technology. Beginning in mid 1970s, with continued ebbs and flows of interest (the latest having been in the mid-1990s), the hallmark of this era was a focus on the user and the result of the search/retrieval – expanding the box to encompass the user as more than an input/output devise, and investigating the inner workings of the box. Continuing the exploration metaphor, this era was the time of the homesteaders. The places are on the maps. Some people came in to settle, to see how the town (or system) worked for them. The question was no longer whether or not the system could work, but how well it worked for the user. The shift was away from the tool and toward the user. The research of this era was often qualitative in method and tended toward a constructionist epistemology. The discussions of the dynamic nature of relevance belong to this era.
(Note that while the author is using the past tense, this era is still very much viable, with user-centric research still being conducted and published.) Technically, this era was the time of larger databases, more linkages, and more storage. As time went on, information could be retrieved from a great number of sources and those sources were beginning to interconnect. In addition, more of the information was full text. This meant that the information retrieval systems could search more exactly – indexing the entire document, not just an abstract or other document representation. More systems were also open to use by neophytes, if not the general public. More computers were available to more people.

The third era is driven by technology – specifically, the web. Since the 1980’s personal computers have become increasingly available. Mosaic, the first widely distributed browser, was released in 1994 and both surfing the web and putting up web sites became increasingly common. By the late 1990s, most K-12 schools, nearly all universities and many community libraries had computers and high-speed access to the internet. The result of this dissemination of technology has been greater access by greater numbers of people to more and more information available. Most who read this dissertation can recall the experience of searching the net in the mid-1990s and being pleased to find a site that had some information relevant to a search. Now, the usual question is not if something will be found, but how to pick through the thousands of hits to find what you want. In addition, fundamentally different kinds of systems are being developed. In addition to the web, there are web-based databases of information not publicly available (sometimes called the hidden web). Databases (both public and “hidden”) of millions of full text records that grow by thousands of records a week exist.
Theoretically, there is a growing diversity of interests as the field of evaluation expands and matures. Researchers involved in studying information retrieval have come to specialize in fundamentally different areas, often with little common ground between specializations. For example, this study, involving very small documents sets (in the range of 500), operates under a completely different set of interests and constraints than a researcher developing retrieval tools for PubMed or some other huge database where it is essential to develop mechanisms to automatically index documents, because it is not possible to physically review the torrent of new information. I have found that there are crucial differences in the framing of research questions and the approach to systems development and evaluation between those working on small and huge systems. However, at this point, these differences are rarely mentioned or recognized. Petroski, an engineer, writes of the problem of scale, “Perhaps no principle in design is so well known and yet so frequently forgotten as the effect of size or scale on performance (Petroski, 1994 p.29)”. Boats and bridges that worked fine at one size suddenly fall apart when “supersized”. This is a pervasive problem in instructional design – studying the effect of an intervention on a single person or small group and assuming that the intervention may then be scaled up with no further study (Schwen, 2001). In this study, I shall be explicitly studying small systems. The design concepts I hope to evaluate would not be tenable for large systems.

In addition to the difference between large and small systems, there are an increasing number of custom systems that are designed for specific groups of users and specific types of information. Some systems, such as popular web search engines have “one size fits all” interfaces that access a huge and growing number of documents. Others
are directed toward specific users who share a vocabulary and information need and require the use of thesauri and/or training for effective use. Researchers are exploring new methods for accessing information through not only the internal workings of the systems, but through different approaches to user interface. As this era proceeds, there is little doubt that more systems that are fundamentally different will be developed. It will be interesting to see how evaluation theory grows to encompass this plethora of systems and diversity of research agendas.
Chapter Two: Review of Pertinent Literature

In this chapter, I will explore the breadth of research into information retrieval evaluation. In addition, I will look to the literature that provided direct guidance for the methods used in this study.

Traditional Information Retrieval Evaluation

Salton makes the point that to properly evaluate a system, one must look to the assumptions underlying judgments of success for that system (Salton, 1992). The assumptions of success established for the large scale evaluation projects in Cranfield, England in the 1960s were that a “good” system retrieved as many documents as possible and that the returned set of documents had mostly documents that pertained to the question asked (Cleverdon et al., 1966; Sparck Jones, 1981). The best system would return all possible documents that had something to do with the question and no documents that were extraneous.

The above theoretical assumptions led to the constructs of recall (how many of the documents that had something to do with the question were retrieved) and precision (how many of the retrieved documents actually do have something to do with the question). Underlying these constructs is the construct of relevance – a judgment as to which documents do actually pertain to each question.

As mentioned before, the theoretical tradition does not well distinguish between constructs and measures. Indeed, the constructs of recall and precision are called relevance measures in the literature. While I will discuss these concepts in more detail later in the chapter, let me clarify the constructs and the measures here so that during the discussion that follows, the distinctions may be more obvious. (I will do so a few times
throughout this report to try to keep the terms separate.) The construct Relevance is the idea that a particular document may be judged to pertain to a particular query. Relevance may be measured in many different ways, including text matching and expert judging. In the literature, this is called relevance judgments and is an area of controversy (Who, for what, and how). So, when the literature talks of relevance judgments (Harter, 1996; Hersh, 1994; Mizzaro, 1997; Park, 1994), this is a discussion of the measure relevance.

The construct Precision is the idea of looking at the set of retrieved documents and seeing how many of them are relevant. While there are disagreements over how (or if) documents may be judged (measured) relevant, the construct of precision is not controversial of itself. The Precision is measured by dividing the number of documents returned by the number of relevant documents returned. Again, the controversy is with the measure of relevant documents.

The construct Recall is slightly more controversial, as it ties to a theoretical assumption. The construct Recall looks at how many of the possible relevant documents are retrieved. Underlying recall is a theoretical assumption that it is good to retrieve as many of the relevant documents in a document set as possible. This is not a problem with the measure, but a problem with the construct. In addition, as with the construct Precision, the measure of relevance of documents besmirches the construct Recall. The measure Recall is, as the measure of Precision, straightforward - divide the number of relevant documents in the document set by the number of relevant documents returned.

Traditional evaluation uses a systems approach – the system is viewed as a black box with no need of user input for evaluation. A batch program runs queries against the test document collection and ratings of recall and precision are calculated. This approach
to evaluation is rooted in an objectivist epistemology (Crotty, 1998). There is an external reality. Documents may be judged relevant or not. A clear measurement of effectiveness of a system may be calculated. Systems are evaluated using experimental research methods.

**Problem with the traditional model**

The traditional model of evaluation, introduced in the 1960s is still very much in use today. However, there are several fundamental questions about this model. The question that has been explored most extensively examines issues of relevance. Questions include: What is a definition of relevance? Who should judge relevance? and Is relevance a static or a dynamic judgment? In addition to relevance issues, some researchers are taking a more user-oriented view, looking at task orientation and interactivity.

**What is relevance**

While very simple on first view, relevance (does a document answer a query) has proven to be a very difficult construct, mostly because of the difficult in developing an accepted relevance measure. As definitions often lay the groundwork for a measure, the only part of a definition that can be agreed upon is that relevance is a relation between two entities. (Mizzaro, 1997). Schamber, writing the first chapter solely focused on relevance in the Annual Review of Information Science and Technology (a significant milestone in the field), used as a foundation definition “the relationship between a user’s information problem or need and the information that could solve the problem. (Schamber, 1994). It is interesting to note that in defining the construct, she placed the user within the definition, indicating that measures that do not address user needs are ineffective. As an advocate of relevance measures that depends on the user, the situation,
and the information need, Schamber immediately distanced herself from the traditional, systemic measures of the construct.

Saracevic, a leading writer in the field, proposes five different measures of relevance (Saracevic, 1996). They are: System, Topical, Cognitive, Situational and Motivational. System relevance is assigned automatically by the system – usually by matching words based on some algorithm. Topical relevance is assigned by a judge through some assessment of query and document. This is the type of relevance used in the traditional evaluations, such as Cranfield. Cognitive relevance depends on the mental processes of the user as they view the information. Researchers who look at cognitive relevance often look at the information need of the user. Some writers call this pertinence. Harter is the writer most associated with this type of relevance, but Spink, Hersh, Barry and Schamber, and Park have all taken a similar view (Barry & Schamber, 1998; Harter, 1996; Hersh, 1994; Park, 1994; Schamber, 1994; Spink, Greisdorf, & Bateman, 1998). Situational relevance involves the task at hand. Barry and Schamber wrote of relevance decisions of weather information from multiple sources by construction managers (Barry & Schamber, 1998). These managers had to decide how to staff a site based depending on the weather. Whether or not a piece of information was relevant was very task dependent. Similarly, Hirsch writes of relevance decisions of medical students responding to detailed questions. Their relevance decisions are based on the task (Hersh, 1994).

Finally, Saracevic introduces the concept of motivational relevance, which he describes as the “relation between the intents, goals, and motivations of a user, and texts
retrieved by a system...Satisfaction, success, accomplishment, and the like are criteria for inferring motivational relevance” (Saracevic, 1996 p. 4).

In a recent article, Borlund suggests two main classes of relevance measures – objective or system-based and subjectivity or human (user)-based (Borland, 2003). System-based relevance considers relevance judgments that do not involve a human judgment, depending on the matching of words or word combinations. This is intuitively a weaker method of ascertaining relevance. However, when one considers the question of scale, the importance of system relevance becomes clear. Most of the new development of information retrieval systems involves methods that attempt to handle document collections of staggering proportions. For example, PubMed is a service of the National Library of Medicine. It provides access to nearly 13 million citations. Thousands of new citations per week are added. When dealing with in such a scale, it is not possible to make human relevance judgments for the entire collection. Yet, in order to make a calculation of the recall of an IR system, one must divide the total number of relevant documents by the number of relevant documents retrieved. In order to evaluate such a system, one is forced to use a subset of the document set, use system-based relevance, or turn to some other type of evaluation that either uses normalized precision (Quiroga & Mostafa, 2002) or relative recall and precision (Gordon & Pathak, 1999; Su, 2003a) or does not use relevance measures.

Borlund details three types of relevance judgments within the class of user-based relevance – intellectual topicality, pertinence, and situational. Intellectual topicality judgments are those such as done in the TREC experiments, where some judge or judges decide prior to the experiment which documents are relevant. Pertinence measures
“...represents the intellectual relation between the intrinsic human information need and the information objects as currently interpreted or perceived by the cognitive state of an assessor or user. This allows for the existence of a *dynamic* information need” (Borland, 2003 p. 915). Situational relevance highlights the dynamic nature of relevance depending on the task. This is often know as task relevance (Reid, 2000; Vakkari, 2003).

System-based relevance and topical relevance (or intellectual topicality) reflect an objectivist epistemology, assuming that an object can be judged by someone (or a set of algorithms) to be relevant or not. The user-based measures reflect a constructionist epistemology, where the user, and only the user, in situ, is competent to make a relevancy judgment. At the heart of this tension lies the question, “Relevant when, to who?” There is an extensive literature that examines the personal, dynamic nature of relevance (Harter & Hert, 1997; Hersh, 1994; Mizzaro, 1997; Park, 1994; Schamber, 1994; Spink et al., 1998). Different users find items relevant, depending upon their own information need and their existent knowledge, preferences, and understandings. Researchers have found differences in relevance judgments depending on the document’s representation (title only, title and summary, index items, full document). While Barry and Schamber have begun to construct a set of shared relevance criteria for relevance judgments (Barry & Schamber, 1998), they agree that each person uses the criteria to make different judgments. In addition, depending on the specific task and the situation (lack of time, level of interest) a person will judge relevancy differently (Barry & Schamber, 1998; Harter, 1996; Hersh, 1994; Mizzaro, 1997). This contention is supported by research. It also makes sense if one reflects upon one’s own search behavior. When doing a web
search, we use different and changing methods for deciding whether or not to look at documents returned by a search.

The problem with user-constructed measures of relevance arises when using it for the recall measure. If it is not possible to assign relevance to all documents in the document set, it is not possible to calculate recall (the measure of how many of the total relevant documents are returned). The traditional methodology uses the paired measures of precision and recall. Relying only on precision would give higher marks to a system that returned one relevant document, but ignored 200 other relevant documents. Relying on recall only would give higher marks to a system that always returned the entire document set, because that would assure 100% recall. The traditional method relies on both precision and recall to give a richer view. Some researchers have constructed measures of recall and precision based only on the documents returned. Called either normalized (Quiroga & Mostafa, 2002) or relative measures (Gordon & Pathak, 1999; Su, 2003a), these measures ask users to make relevance judgments for the set of returned documents, then calculates precision and recall from that number. This approach has had limited application.

The people involved with building systems seek a nice, clean measure…a number by which they can compare systems. It is important to keep in mind that this is the purpose of evaluation – a comparison…a comparison of two systems, or a comparison of alternative manifestations of a single system. Information retrieval systems are typically large and expensive to build. Developers and funding agencies want some method to easily compare systems and manifestations. The traditional method provides that method, while research using user defined relevance measures, although they give a richer view of
searches does not. A recent pair of articles by Su (Su, 2003a, 2003b) present the first research this author is aware of that compares systems with user-centric measures. For that study, Su used relative recall and relative precision measures (among several other new constructs and measures), having users judge the relevance of a document set of 20 retrieved documents. Most studies that explicate user-based relevance measures study those measures while evaluating a single system or focus on user behavior to explore the new measure. Even if traditional relevance measures rests are questionable, the constructs of recall and precision are a universally accepted means of reporting system performance and continues to be used for lack of another, similarly clean, alternative. This is analogous to the debate in education about using standardized tests as a measure of success or a predictor of future success. While there are definite questions about the validity of test scores, they continue to be put forth by elected officials as useful indicators. This is because it is easy to talk about and compare test scores. If scores go up, that is good. If they go down, that is bad. Such news easily reduces to a headline or sound bite. The intricacies of the assumptions upon which the test development rests are not so easy to understand. In the same way, while there are fundamental problems with the measures in the traditional information retrieval evaluation methodology, recall and precision are ubiquitous because they are simple to do and easy to understand and compare.

To this researcher, the tension between the traditional paradigm and those interested in understanding and questioning relevance measures represents the tension between two research epistemologies – objectivism and constructionism. Not a dichotomous shift, there is evidence of a move along a continuum from objectivism
toward constructionism. In addition, there is a clear move toward adopting more qualitative methodologies within the field.

The objectivist thinkers see the Cranfield model as providing a clear measure of success. Of course, external judges can assess relevancy. The experimental methods allow batch testing – so each system is tested numerous times with numerous queries, with variables well controlled. To an objectivist view, this gives the best possible evaluation of a system. For those moving toward a more constructionist epistemology, these methods are as the house built on sand – “The rain will fall, and the floods will come, and the winds will blow and beat against that house, and it will fall” (Matthew 7:26-27). However, it is clear that while some researchers have constructivist leanings – looking at a user-centric construction of relevance rooted in an individual, a time, a task, a situation – they have not given up the hope to be able to provide a means to evaluate systems, clearly an objectivist goal.

These “closet constructivists” have certainly expanded the methods of research, investigating task-based, naturalistic studies (Vakkari, 2003). A rich research literature going back at least 30 years points to inherent weaknesses with using relevance based measures (the constructs Recall and Precision), yet all IR systems are still evaluated using Cranfield-type experiments. Two groups, two different research traditions, one undergoing an epistemological crisis of faith exist within the field with little understanding or communications between the two sides, and mot much realization that there even is a problem. So, like a bad marriage, there is tension. One group studies and questions the traditional constructs and measures. The other group continues to read their morning paper, mumbling, “Yes dear.” while continuing to use Cranfield-like
experiments for evaluation. The frustration is particularly evident in the writings of those involved in relevance research. In 1996, Harter wrote,

...the reaction to this research and criticism from experimental researchers who use relevance assessments to conduct Cranfield-like experiments on information retrieval systems, has been mostly silence...It is as though the enormous experimental and naturalistic research literature discussing factors that affect relevance assessments did not exist, or that the critics of the use of relevance as a foundation for retrieval testing are publishing in remote fields or live in distant lands with which we have no communication. The critics have not been refuted, or even acknowledged. Mostly they have simply been ignored. Researchers conducting experimental work in information retrieval using test collections and relevance assessments assume that Cranfield-like evaluation models produce meaningful results. But there is massive evidence that suggest the likelihood of the contrary conclusion (Harter, 1996 p.43).

In Instructional Systems, a similar tension over methodologies played out over the past two decades, with qualitative methods growing in influence and use. The resultant blending of methodologies (Denzin & Lincoln, 2000a, 2000b) provides guidance in this current research. The current problem in evaluation is that on one hand, relevance-based measures using experimental methods provide too thin a picture of a system’s efficacy, but on the other hand, alternative constructs and measures have either not provided potential for evaluation, not provided potential for comparative evaluation of systems or have required too many resources to complete. In this research, I seek to apply Yin’s (Yin, 2002) case study methods (a qualitative method with an objectivist foundation) to provide a new view of evaluation. By using a case study approach to evaluation, it is possible to examine a smaller number of users in greater detail, attempting to obtain a richer understanding of users’ views of two systems. This move away from experimental,
batch testing methods allows investigation of new constructs and measures for comparative evaluation.

**What to study?**

In order to obtain a richer comparison, this study looks at four constructs and their attendant measures – the paired traditional constructs of precision and recall and two measures that do not use relevance judgments – information need and utility. The tradition constructs of recall and precision provide a benchmark. Using the traditional constructs requires a document set, a set of queries or tasks, and relevance judgments about which documents match which tasks.

I have not explored constructs that utilize user-based relevance measures in this study. These constructs have been extensively researched elsewhere. However, even though these constructs and measures address the problems inherent in the traditional measure of relevance, they do not address the underlying theoretical assumption that “success” means retrieving as many relevant documents as possible. As I am attempting to specifically examine retrieval in a performance support context, where more is not better, I feel a more effective investigation will explore non-relevance based constructs and measures, with the traditional constructs providing a comparison.

I follow Hersh’s lead and look to the outcomes experienced by users, not the system for the focus of evaluation (Hersh, 1994). Hersh works in the area of medical information retrieval and is at the forefront of task – based evaluation. In medical information, there are huge quantities of information that is constantly changing. It is important that information found using these systems be not only relevant, but also
correct. Hersh argues that we need to develop new constructs as well as measures of the relevance construct.

...systems should ultimately be judged by how well the help users in their task for which they consult the system, whether it is to make correct decisions or lead to some improved state in what they do with the information. An outcomes-based approach provides a different perspective on the topical versus situational relevance debate, indicating that neither is adequate for a complete assessment of retrieval systems. (p.202)

In addition to the traditional method of evaluation, I turn to evaluation models that use different initial assumptions of success. In this research, I have used measures based on Spink’s Information Need and Cooper’s Utility. Using these methods in addition to traditional relevance measures within a case study methodology will, it is hoped, provide a richer comparison of two systems.

Different assumptions for success

Spink (Spink, 2002) has developed a methodology which includes the construct of shift or change in information need. Spink has subjects self-rate themselves before and after a search as to which stage in the information search process they feel they are. These stages are taken from Kuhlthau’s work (Kuhlthau, 1997) and are:

1. initiation
2. selection
3. exploration
4. formulation
5. collection
6. presentation
These rankings seem to be rather gross measures for rating the efficacy of a search. They also do not map well to a performance support system, where the problem is already established and presentation is not applicable. So, while the specific measure is limited in this context, the construct of information need and the idea of measuring the change before and after the search is certainly useful. By asking users to self-rate information need before and after a search with a system without attaching Kuhlthau’s labels, we have a useful measure for the construct.

The gross measure used by Spink highlights an unstated assumption of search behavior in most studies that is in contrast to the behavior in a performance support system. Usually, when searching, a person wants to answer a question, or get information to accomplish a goal. It is not unusual for a user to go through Kuhlthau’s six steps within a single search session. When doing this, several questions will be formulated, asked, and answered. However, in a performance support system, there is usually a single question to be answered or task that needs to be accomplished. Consequently, the searches are much more direct, simple, one-dimensional. This is why, after consultation with committee members, I decided to have users complete two searches for each system for this study.

William S. Cooper’s View of Evaluation

Finally, I draw from the writings of William Cooper (Cooper, 1973a, 1973b, 1981). Cooper was at UC, Berkeley in the 1970s and 1980s and challenged some of the foundational assumptions of information retrieval. In an article in 1973, that won the “Best Article of the Year” award from the most important journal in the field (The Journal of the American Society for Information Science), Cooper questioned the basic
assumptions of retrieval evaluation. He proposed a new construct - utility, which is based on the value a user puts on a document retrieved. How valuable or useful are the documents retrieved by a system for the user? That, Cooper says, should be the judgment of success. Relevance is a good construct for information scientists focused on designing and testing systems. It is easy to design experimental tests that do not require human subjects. However, this leads to systems designed to optimize the goals of information scientists (a systems design version of teaching to the test). Cooper feels this is an unworthy direction. “Systems should be designed so as to optimize the satisfaction of their users, not the information scientists.” (Cooper, 1973a p. 92). One must look at utility, the usefulness of documents for the user, not relevance, even if relevance is the easier way to judge a system. He suggests using only relevance is like a mechanic who only knows how to do wheel alignment and so judges the performance of a car on the basis of a wheel alignment test. This limited construct was patently ridiculous when Cooper wrote in 1973. Unfortunately, thirty years later, with larger and much more complex “cars”, the system mechanics are still only checking the alignment to evaluate systems. Relevance is not a bad construct; it merely offers a partial view.

In a seminal book on information retrieval experiments published in 1981 (Sparck Jones, 1981), Cooper used his chapter to poke holes at the idea of doing experimentation at all. He views experimentation as expensive, not grounded in any theory, not generalizable, and not particularly useful.

Perhaps partly in recognition of this paucity of theory, many researchers have turned to experimentation, and especially laboratory experimentation. As might be expected, the classical experimental approach has been fairly theory-independent, consisting essentially in the trying out of various competing retrieval schemes...to see which seem to work best...It is hardly surprising, then, that on the rare
occasions when the methodology used in a large-scale retrieval test has been subjected to careful independent scrutiny, the results have been far from reassuring" (Cooper, 1981) pp. 201-202).

In place of most experimentation, Cooper suggests theory based thought experiments. In place of relevance, he suggests the “utile” as a measure of the utility construct. The utile is measured by asking users to rate individual documents based on how much they value them. The entire system’s utility may be measured by averaging the utile of each document. Cooper is much cited when those who question relevance are suggesting different measures for evaluation. However, I have not found any research that attempts to use utility as a measure. Cooper himself is also not aware of any such research (Cooper, 2003). It is my belief that the reason Cooper has not been used is that the measure is more amenable to qualitative methods, and in the 1970s, such an approach was not be considered good science in the field of information retrieval evaluation.

Indeed, Cooper not only attempted (in a second article) to apply a quantitative methodology to this underlying theory (Cooper, 1973b), he also called it a “naïve evaluation methodology” and spent two thirds of the paper suggesting objections and answering them. With the current general acceptance of qualitative methods in IST and the beginnings of acceptance in the field of information retrieval evaluation, the time has come to use Cooper’s naïve methodology to inform the methods of retrieval evaluation.

The goal of Cooper’s construct is to “measure somehow the retrieval systems’ ultimate worth’ to its users” (Cooper, 1973a p.88), or to develop a means to measure this construct. He proposes a series of questions to ascertain the measure of value (or utile). “Since there is probably no better way of getting the value of a system than by questioning its users (and possibly other people) directly, the meaning of the phrase
‘ultimate worth’ is most readily clarified in terms of some kind of questioning or elicitation procedure” (Cooper, 1973a p.88). The goal of this questioning is to find the quantifiable “utile”. In his article, Cooper used money as the utile. At that time, searching was not generally available. It was common to have to pay for expert searchers. So, using money as the valuation of a search was natural.

Cooper suggests the following method for measuring the utility of a system. A search is done and the user reviews the first document. The researcher asks the user to make a judgment as to whether that document has been helpful or not, and then to quantify that judgment. If it has been helpful, how much would they be willing to pay to have had this experience with the document? If the document was not helpful to the user, the question becomes how much the user would pay to avoid such an experience. If the user feels nothing was gained or lost, the researcher records a zero. Cooper even suggests questions to elicit this type of information. This number, either positive or negative is the “document-utility”. The user then goes on to evaluate the next document. Cooper suggests that the utility of the second document is dependent on the first. For example, if the second document is a near duplicate of the first, its utility will be low or even negative.

Cooper defines the end of the evaluation process as the point when the user has either “satisfied his information need or given up the search”. So, after each document is evaluated, the user is asked whether they are satisfied, want to give up, or want to continue. This is significantly difference from relevance-based measures of evaluation. The number of documents evaluated is not necessarily the number of documents the system returns. It depends upon the wishes of the user. Once the information need is
satisfied or abandoned, the evaluation is complete. This can mirror the activity of the user when using a performance system, where the user seeks an answer and stops once an acceptable answer is found. This is user-centric instead of system-centric evaluation.

When the user does not want to continue, the measure of system utility is calculated as follows. The researcher totals the document-utility figures obtained for each document in the sequence. This sum is Search-utility. This procedure is repeated for every system use or search with both systems. The researcher then averages the search utilities for each system, providing a measure he calls the system-utility or utility of each system. Cooper suggests the following methodology:

Finally, he subtracts the utility of system A from the utility of system B and computes a confidence interval around the difference in accordance with accepted statistical practice. The result of his investigation as he submits it to his sponsors has therefore the form of a single number with accompanying confidence limits; e.g. a final result of $3.5+\cdot2$ at the 99% confidence level would be interpreted to mean that the managers of the system could have that degree of statistical confidence that system B was worth, on the average, between $3.30 and $3.70 more per search than system A, in the judgment of the users (Cooper, 1973a p.90).

In this research, I define the utile by time, not money. So, instead of taking the actual monetary value, I ask users to rate how much time they have saved by seeing the useful documents on a scale from 1 to 7 (more detail is provided in the next chapter). The performance system evaluated in this research is designed for teachers and, having interviewed teachers during the development of the system (discussed in detail in Chapter Four), a common complaint was lack of time. In today’s online environment, where a large amount of information is freely available, the cost is rarely actual money, but the time necessary to sift the wheat from the chafe. Cooper agrees that time is a useful measure in this case (Cooper, 2003). With this change, Cooper’s naïve methodology
reflects the needs of evaluation in a performance support system. As it uses the mean of
the utility of each document, the number of documents retrieved is not as important as
how useful they are. In this methodology, a system that returns a very few highly useful
documents would be rated higher than a system that returns a great many documents of
questionable usefulness.

So, in order to get a richer comparative evaluation, I begin with the traditional
method, using the quantitative measures that date to Cleverdon’s work in Cranfield in the
early 1960s. I then turn to Spink’s Information Need. Finally, I operationalize Cooper’s
naïve methodology from 1973. In the next chapter, I shall detail the methods of this
combined evaluation.
Chapter 3: Methods

Traditional information retrieval evaluation is objectivist, experimental and quantitative. Cooper (Cooper, 1981) has questioned the efficacy of this model, suggesting that “full-scale retrieval tests are difficult, expensive, unreliable, and often inconclusive”. A reason for this is because of the lack of a strong theoretical basis.

In fact, in the search for a general theory it is hard to do much better than to give some elaboration of the vague rule that a system should retrieve for the user those documents most likely to satisfy him. As scientific theories go this truism is not very impressive, but it is the only wisp of general theory we have. What was said of a recent political candidate can be said of document retrieval theory: Deep down inside it’s shallow. (p. 201)

The current state of information retrieval evaluation is not only theoretically shallow; there are some theoretical problems, specifically with the construct of relevance and how it is measured. Relevance based constructs of recall and precision do not take into account the cognitive processes of the users because of the way relevance is measured. Recall and precision do not connect the user and the evaluation of results. However, recall and precision have been used in hundreds of studies. They are still the standard evaluative tool for information systems and have been for over 50 years. So, in this study I am going to look at two constructs that do look at the consequences of cognitive processes to see if those measures may give a richer view for evaluation.

Theoretical Assumptions

Traditional Theory

As stated in chapter one, the theory that underlies the traditional method of evaluation includes constructs, relationships between constructs, measures of constructs and assumptions. For the traditional theory, these are:
Assumptions

• Relevance judgments are valid and meaningful indicators of the effectiveness of an information retrieval system. (Harter & Hert, 1997)

• Relevance is a meaningful construct on which to base the evaluation process.

• Systems can be evaluated using the combined measures of recall and precision.

• A successful retrieval system retrieves the greatest number of relevant documents with the least extraneous documents.

Constructs

• Relevance – a judgment of whether a particular object answers a particular question.

• Precision – how many extraneous or non-relevant documents are retrieved for a particular question by a particular system. A measure of accuracy or quality of retrieval.

• Recall – how many documents drawn from a particular set of documents relevant to a particular question are retrieved with a particular system. A measure of scope or quantity of retrieval.

Relationships between Constructs

• Recall and Precision have an inverse relationship. Generally, the greater the recall (number of relevant documents retrieved), the lower the precision (more noise).

Measures
• Relevance judgments can be made through a variety of methods, including automatic text matching and expert judgments.

• Recall is measured by dividing the number of documents that have been judged relevant to a particular query in a document set by the number of relevant documents retrieved for that query.

• Precision is measured by dividing the number of documents retrieved for a query by the number of relevant documents retrieved.

There are some additional considerations of these constructs and their measures. Recall and precision are constructs that look at number of documents. They do not evaluate content. In most research, when relevance is measured, it is a binary measure – a document is either relevant or not. Most studies do not define measures of relevance that can address questions of degree of relevance of retrieved documents (Su’s recent paper (Su, 2003a, 2003b) is an exception, using a 3 level measure of relevance). So, for example, recall can tell us that a particular system retrieved 50% of the relevant documents. Recall cannot provide information about the retrieval of the (for example) 10% of really significant documents. Neither recall nor precision is concerned with measuring degrees of relevance or quality of document. A document that has been judged relevant counts as one relevant document toward the calculation of recall and precision – no more and no less.

New Constructs

This study seeks to look at new constructs and new measures that are not relevance-based, freeing evaluation from the paradigm requiring relevance judgments, which makes situ research on authentic search tasks impossible. One new construct is
Utility, based on Cooper’s writing (Cooper, 1973a, 1973b). The measure of utility is based on the utile – what the searchers value. Originally, the utile was money. In this research, the utile is time. Each document’s utility is measured in utiles and the total System Utility is simply a mean of the individual document utility scores. The second construct is Information Need – the notion that a the amount of information a user desires to answer the query they bring to a search tool changes between the time they begin to search and the time they stop searching. Information need is measured by self-rating of time needed to respond to a task both before and after a search and, in this research, is then mapped to a 7 point scale (discussed in more detail later). The use of these measures is examined in a case study, as detailed by Yin (Yin, 2002). This research examines whether these new constructs may provide additional insight into evaluation of information retrieval systems. In this specialized sense, I am generalizing to theory. The new constructs studied here are independent of each other, so do not have relationships between the constructs (as recall and precision do). The new constructs, measures of constructs and assumptions examined for this research are:

**Assumptions**

- End user consideration of document content can provide a useful measure of evaluation.
- Usefulness for the user of documents retrieved is an important way of ascertaining the effectiveness of a system.
- Usefulness is relative, a judgment of each individual user.
- A self-assessed measure of information need, taken before and after search/retrieval can provide a useful measure of effectiveness of a system.
**Constructs**

- System utility - a user-based judgment of how useful the documents retrieved are for a particular user for a particular task.
- Utile – the measure of utility. In Cooper’s original work, the utile was money – cost of retrieval. In this research, the utile is time – time spent or saved by finding a particular document.
- Information Need – the amount of time required to be ready to fulfill a task. By measuring the change in information need from before to after a search, one can ascertain the effectiveness of a system.

**Measures**

- System Utility can be measured by asking users to rate the usefulness of individual documents and averaging the ratings of all documents retrieved that the user wishes to examine.
- Information Need can be measured by asking users to self rate their information need measured both in time and on a 7 point scale before and after searching. The difference between the two assessments is the change in information need.

**A Case Study**

This is case study research based on an objectivist epistemology and a postpositivist methodology (Crotty, 1998; Phillips & Burbules, 2000). This study attempts to capture a naturalistic setting. Denzin and Lincoln say, “…qualitative research involves an interpretive, naturalistic approach to the world. This means that qualitative researchers study things in their natural settings, attempting to make sense of, or to
interpret, phenomena in terms of the meanings people bring to them” (Denzin & Lincoln, 2000a p.3). This study is not a true naturalistic setting. It does not involve observation of people using information retrieval systems with “real” information needs. However, neither is it an experimental setting, as is the format of the traditional method of evaluation. Participants searched for information to meet tasks. Each person picked tasks from a list of 36, selecting the ones most like an actual task they might undertake. A concentrated effort toward authenticity drove the development of the tasks. The tasks are based on real needs expressed by teachers during interviews used in the development of the tag-based system (detailed in chapter four). The tasks were further reviewed for authenticity by a former elementary school science teacher. In addition, participants were asked how realistic the task was for them. In the continuum from naturalistic to experimental, this study may not have crossed the border into naturalistic, but it lingers amiably at the gates.

Collecting the data

Studies that evaluate using the traditional method use a large number of queries and a large number of searches. As there is no need to involve humans in this method, the systems are tested with batch programs that can run queries hundreds of times for little more cost than the first experiment. On the other hand, user-focused studies investigating relevance decisions often use as little as nine respondents. Case studies often use small numbers of respondents, trading more respondents for a richer understanding of each respondent. This study attempts to bridge these methods. The foundation of the study is a comparison of evaluation constructs used to examine a comparison of information retrieval systems. Each respondent evaluated documents retrieved by two different
systems, searching twice using each system, for a total of four searches per respondent. Thirty-four educators participated in the research. The goal of the research is a richer view of evaluation through a deeper understanding of theoretical constructs and measures.

This is a three x two study – three types of evaluation (traditional, information need, and utility) and two cases (a tag based and a text based search tool). Participants were recruited via email, posting on Listservs and bulletin boards, and referrals by associates. Sixty three educators agreed to participate, 34 finished the questionnaire. Participants were all educators. Most were classroom teachers (details in the results section).

This research used a web-based data collection. This allowed a larger number of respondents from a wider geographic area (including educators from Florida, California, Oklahoma, and Massachusetts). The field of information retrieval is slowly moving toward a greater acceptance of qualitative research. However, there is still a very strong tradition of quantitative methods. For one used to seeing traditional evaluation methods, a sample of less than 30 might be so unusual as to strain the credibility, causing the work to be devalued. While many of the user-based studies (Barry & Schamber, 1998; Chen, Houston, Sewell, & Schatz, 1998; Harter & Hert, 1997) used small sample sizes (9-15), the traditional studies that use experimental research designs use larger samples.

Finally, it is apparent that one of the reasons that the traditional evaluation method has continued to be used is that it can be applied relatively quickly and easily. System designers can understand how to use that method to evaluate the system and can easily understand the results. If, because of this research, a web-based research protocol tool is
shown to provide a useful view of the evaluation process, then this research will have the additional impact of providing an easily replicated method for others to use.

Participant Process

Participants completed a web-based questionnaire (Appendix C). The process was as follows. Each participant conducted four searches – two for each of the two systems. Participants were randomly assigned to start with either tag or text based search. After filling out some demographic questions, the participant selected a task from a list of 36 tasks. They rated their information need (Pre test for information need), both on a 1-7 scale and by estimating the time they would require to meet their information need to complete the task (which was converted to a 1-7 scale during analysis).

Next, participants viewed an instruction page that detailed how to use the search tool (participants were randomly assigned to start with either the tag or the text search tool). The participant searched as long and as many times as they wished. For returned documents, both search tools (tag and text) displayed the title of the document and a short description. Participants could mark a document they wished to keep and then continue searching. Participants could keep a maximum of seven documents. Searching continued until the participant clicked a button saying they were satisfied.

The questionnaire then asked them to review the documents they had selected, one at a time. For each document, they were first asked if it had been useful (Yes/No) and then, considering the time they had spent searching, if they were better or worse off. If the document was useful and they were better off, then they were asked to quantify how useful (1-7) and how much time they had saved by finding this document (1-7). These answers were averaged to compute a document utility score. Finally, they were asked if
they felt they had enough information – if they would stop looking at documents if this were not a research project. This answer was used in the system utility calculation. If a participant indicated they would stop, no more document utilities were calculated. However, the participants were asked to review all documents they initially selected, for the calculation of information need and in order to produce a richer data set for further research.

After reviewing all documents, participants were asked the two information need questions (detailed above for pre test) to provide a post intervention measurement. In addition, two questions of user satisfaction were asked – satisfaction with results and satisfaction with the system. While satisfaction is reported in the results chapter, it is not considered in this research as an evaluation method. Su does include satisfaction as a measure (Su, 2003a), but it was decided upon consideration that for this research it is not a robust enough measure to be included alongside information need and utility.

Participants were then asked to repeat the entire process with the first search tool in order to mitigate differences in results that might have been caused by being unfamiliar with the search tool. Results were averaged between the two searches, resulting in a 1-7 ranking score for traditional method, information need, and utility. Participants were then asked to repeat the process with the second search tool. The entire process took about one hour. Participants were able to leave the questionnaire and return at a later time. Most took advantage of this feature.

Screen captures of the two search tools and a sample set of retrieved documents may be found in Appendix B. Screen shots of a walk through of the questionnaire may be found in Appendix C.
Overview of Methods

- Construct two very different systems (to provide a clear contrast between systems – one basic text search tool (the Swish free engine www.swish.org) and one tag-based, controlled vocabulary system (detailed in the chapter four).
- Develop a set of realistic tasks based on teacher interviews from study detailed in chapter four. (Reviewed by a teacher for authenticity)
- Gather 500 education related documents from all over the web.
- Tag all documents for the tag-based retrieval system.
- Make topical relevance judgments for all documents.
  - Have teacher review all relevance judgments.
- Recruit 30 educators (teachers, media/library, and administrators).
- Each participant will use both search tools, but is randomly assigned to begin with one of the two systems.
- With a web based questionnaire:
  - Instruct how to search using the first tool.
  - Have the respondent choose a task.
  - Respondent ranks task as to “authenticity”.
  - Respondent answers questions to ascertain information need. (Information need pre-test)
  - Respondent searches. No time limit. Respondent can search and refine search as often as they wish.
  - Return is fixed so that a maximum of seven items are retrieved.
  - Note time taken to complete search.
• Note items returned for traditional recall and precision. (Traditional evaluation method)

• User answers utility questions. (Utility evaluation method)

• User asked if they would continue to look for information. If they respond no, no more utility judgments are made.

• Process repeated for all seven objects.

• Afterwards, user asked post-information need questions. (Information need evaluation method – post test)

• User asked satisfaction questions.

• User repeats entire process with second task.

• User repeats entire process with second search tool.

• Respondents were able save their research protocol and do it in more than one session.

• Evaluate data

Questions asked of the data

• Compare the recall/precision of the two systems.
  
  o Is the difference significant?

• Compare the system utility of the two systems.
  
  o Is the difference significant?

• Compare the change in information need using the two systems.
  
  o Is the difference significant?

• Compare the time taken to finish searching with the two systems.
  
  o Is the difference significant?
• Is there a correlation between high traditional precision, high user precision, high utility, and high satisfaction?

Areas of discussion

Document Collection

Documents were found in over 300 sites. The researcher started with the list of queries and used different search engines (mostly Google) to locate web-based resources that seemed to related to the queries. Often the documents retrieved from one search had other pages linked to it. In addition, several sites that were lesson plan repositories provided several objects. Locating and tagging the documents took eight weeks.

Documents of all qualities were captured. As a tag-based system has an advantage over the text-based system when seeking documents that are all or mostly graphics, and the researcher did not want to unfairly disadvantage the text search tool, nearly all of the documents were largely text documents. Documents with graphics that illustrated a point made in text were used. In addition, the great majority of objects were relevant documents. There was no effort to “test” the text engine by providing documents with irrelevant homonyms.

Once located, the URLs of the documents were entered into the project database, along with a short description. This allowed tagging of objects and prevented documents from disappearing or changing after being tagged.

Task collection

Participants picked their tasks from a list of 34 tasks (complete list in Appendix A). The tasks were developed from two sources – extensive interviews with teachers and administrators during the development of the tag-based system and the Indiana Academic
standards. Thirteen tasks involved teaching, technology, or assessment and were applicable to both teachers and administrators. Three tasks for each of the grade levels 1-3, 4-5, 6-8, and 9-12 were identified, mostly taken from science or math standards. Nine tasks were directed at science/math teachers for grades 10-12.

How people selected

In order to recruit the target population of 30 respondents, several methods were employed. These included:

- A web site promoting the study, encouraging people to contact me to take part.
- Postings to listserves and educational web sites.
- Contact with known associates of myself and colleagues.

Over 900 email messages were sent out. Three different science educator sites gave permission for me to post a message in newsgroups. One science educator group in Indiana allowed me to post a message to their listserv, which sent the message out to over two hundred subscribers. One national site allowed me to post a message and conduct an online lecture on search evaluation (although that lecture did not result in any respondents.) One national teacher professional development site gave me permission to send emails to their entire registered user base (about 500). Finally, throughout the recruitment phase, I spent many hours online, searching for educator email addresses in a variety of sources, including lesson plan repositories and school web sites. The result was that 63 educators agreed to participate. Of those 53, 17 started but did not finish – many due to technical problems. 11 never started (4 emailed saying they did not have time). Eight finished, but were lacking 1 of the 4 searches because of technical problems. These
participants were sent an email asking them to complete an additional search. Only 1 did so, so the original 8 has been reduced to 7. These seven were included in the statistics, but results of one of the systems were not based on an average between two searches. The purpose of having two searches was to diminish any learning effect – a participant getting better at searching as they searched a second time. In all cases, the search not recorded by the system was the first search, so these participants did do two searches and the results of their second search were recorded. Of these seven, four were missing one tag search and three were missing one text search. One participant finished, but was lacking 2 searches – this data was not included in the statistics. 27 participants finished all 4 searches. So, analysis was completed using data from 34 participants.

Relevance judgments

As discussed previously, a question raised in many studies of relevance is who decides what is relevant and when do they decide – the measure of relevance. With the traditional construct of relevance (often called topical relevance), an “expert” makes a judgment before users search.

For this study, those making relevance judgments used the following definition: An object that provides, in your judgment, useful information toward the completion of the task is a relevant object.

The method for the measure of relevance for this study was: The researcher reviewed all the documents in the document set, comparing them with the set of tasks and made judgments as to the relevance of each document. A former teacher also reviewed the documents with the set of tasks to validate the judgments. A document was judged
relevant if judged relevant by both judges. A total of 201 relevant documents were identified by one of the judges. Of these, 146 were judged relevant by both judges.

The measures of the three constructs

A brief discussion of each of the three evaluation measures follows. The complete questionnaire is included in Appendix C.

1. Traditional Method

When a participant indicated that they were finished searching by clicking the “I’m satisfied” button, the system recorded the documents (up to 7) that they had selected. These were compared to the documents judged to be relevant for that task, allowing computation recall (# of relevant documents returned divided by # of relevant documents) and precision (# of relevant documents returned/# of documents returned). Because each respondent completed two searches per system, the two measures of recall and precision for the system were averaged, resulting in a score of recall and precision for each participant. This was repeated for the second system. So, each participant ended with a percentage score of recall and precision for each system.

In order to easily compare these percentage scores, the researcher also mapped each individual’s recall and precision system scores onto a 1-7 scale as follows: 0-14%=7, 15-29%=6, 30-49%=5, 50-69%=4, 70-89%=3, 90-99%=2, 100%=1. Then the recall and precision scores were averaged for a new measure of the traditional construct called P/R. This P/R measure could then be compared to the other methods.

2. Information Need

Information Need is not relevance-based measure. This construct is measured by reporting the change in the participant’s information need based on interaction with the
system. As such, it requires ascertaining pre and post levels of information need. These are self-ratings by respondents, as opposed to external tests or measures of understanding of the task. The pre and posttest asked the same two questions:

1. Thinking just about your information needs (what information you need to complete the task), not physical preparation such as gathering materials, just considering your information needs, how ready are you to complete your task?

Rate from 1 – 7 1= I have no information need. I’m ready to complete my task right now. 7 = Large information need. I have little knowledge of this and very little, or no supporting information.

2. Thinking about your information need, how much time do you think it would take for you to get all the information you need to complete your task. _____ hours _____ minutes.

To measure the first question required a small calculation. There was a score of 1 – 7 that each respondent had rated at the beginning and the end of each search. As with the other measures, 1 was the best possible, so in this case, if the participant had no information need, they would select 1. In most cases, participants had a rather high number for the pre test. This makes sense, as it is no surprise that someone would have a greater information need at the beginning of a search. Indeed, one would expect that the pre score would be high and the post score would be low, indicating that the information need had decreased because of searching and looking at the documents retrieved. This is,
after all, the reason for searching. The highest rating possible would result from a participant who rated their information need as a 7 in the pre test and a 1 in the post test.

However, two extra calculations were required. As the final score is the result of a subtract (post test minus pre test), the resultant score must be flipped. For example, although 1 is the highest score, a “perfect” performance of pre test 7 and post test 1 would result in 7-1=6. A score of 6 indicates a very bad performance. So, the scores had to be remapped – 1 = 7, 2=6, 3=5, 4=4, 5=6, and 7=1.

In addition, it was necessary to add 1 to the pre test score before doing the subtraction. Again, consider the “perfect performance” (pre test 7 and post test 1). The result of the subtraction would be 6, which would map to a score of 2. So, because of the subtraction, the highest score possible would be 2. This was rectified by adding 1 to the pre test score before the subtraction.

The resultant score for information need for each search (two per system) was averaged, creating a score for question one information need for the system.

For the second question – change in time required to be ready to complete the task, the post test time was subtracted from the pre test time. The resultant numbers from all respondents were mapped to a 7-point scale that was developed by looking at the range of time reported from the data. The scale used was: less than 0 minutes (more need than the user started with) = 7, 0 minutes = 6, 1-15 minutes improvement = 5, 16 – 30 minutes improvement = 4, 31 – 60 minutes improvement = 3, 1 – 2 hours improvement = 2, more than 2 hours improvement = 1.

The scores for the second question was averaged across the two systems, resulting in a question 2 information need score for each respondent. Then the information need
score for questions 1 and 2 were averaged for each person, resulting in a single information need score for each person. This was repeated for the second system the respondent used to search, resulting in an information need score for each person for each system.

3. Utility

Based on Cooper’s “naïve methodology” detailed in chapter two, the measure of the utility of the system was quantified for each respondent. When Cooper originally proposed his methodology, he used money as the measure of utility (the utile). In the 1970’s, it was expensive to access information systems. Now that many systems are free to access, the coin of realm is time – how long does it take to sift through the chafe to find the wheat? I started with Cooper’s suggested question: “Are you better or worse off having had the experience of looking at this document?” If the answer was worse, then the document was rated as zero. Cooper proposed a simplifying assumption that may be summarized as bad is bad, there is no good purpose served in trying to quantify how bad. So, if the experience with the document was negative, in this study, the document was assigned zero. To quantify positive interactions, I used the following two questions.

1. We are trying to evaluate – to quantify how much you value this document – how useful it is for you working on this task. It takes time to find and look at a document. For this particular document…for you…for this task – How much time would you be willing to spend searching in order to find this document?

1=over an hour; 2=31-60 minutes; 3=16-30 minutes; 4=11-15 minutes; 5=5-10 minutes; 6=1-5 minutes; 7= about a minute or less
2. How much time do you think seeing this document has saved you toward completing your task?

1=More than 2 hours; 2=1 – 2 hours; 3=30 minutes to an hour; 4=15 – 30 minutes; 5=5- 15 minutes; 6=2- 5 minutes; 7= a minute or two

As with the other measures of constructs, the respondents were looking at the documents they selected during their search for information to help them complete a task that they had selected. Respondents searched four times in total – two times using each of two systems. For each document set they retrieved, they answered questions from which were derived scores for information need and utility (the traditional measure was calculated automatically).

To measure utility, each respondent looked at each document they had chosen. If looking at it was positive, they answered the two questions above, which resulted in two utility scores for each document. The two numbers were averaged to derive a document utility score.

After the first document was rated, the user was asked if they now had enough information to complete their task, or if they would like to continue looking at documents. If they wished to continue, they moved to the next document, examined it, and answered the three questions (better or worse and the two measures), again creating a document utility score. When the participant either indicated that they would stop looking at documents or came to the end of the documents selected during the search, search utility is calculated by taking the mean of the document utility scores. As with the other measures, the participant repeated the process. The two system utility scores were averaged with the first search to create the measure – system utility.
4. Satisfaction

Some researchers have worked to develop measures of satisfaction, notably Bruce and Chen et. al (Bruce, 1998; Chen et al., 1998). Hersh (Hersh, 1994), who studies physicians retrieving medical information is doubtful of the construct satisfaction. Just because the user “likes” the information does not mean that the information is “good for them”. I did not use satisfaction as an evaluation construct, feeling it was not as robust as the other measures. However, a measure of satisfaction was taken to provide some additional insight. This proved to be useful (as detailed in the results section).

I used two questions (ranking 1-7) as a satisfaction measure. The first asked how satisfied the user was with the results of the system (focus on the actual documents retrieved). The second asked how much they liked using the system (focus on the system – interface and functionality). As with the other measures, the resulting scores for each participant were averaged for each search and then across the two searches for each system, resulting in a system satisfaction rating for each participant.
Chapter Four: Developing Teacher Based Tags

Introduction

This research is focused on developing a richer theoretical basis for the evaluation of information systems by comparing two different search tools. One, a text based search, is familiar. Most web search engines (Google, Yahoo, Alta Vista) are text based search tools. The other search tool is based on meta tags. The reader will recall that in chapter one, I described the significant differences in tag-based searching – that tags filter documents. For example, if a searcher does not select any tags, the entire document set is returned, because nothing has been filtered away. Conversely, if nothing were typed in the search box in a text search, nothing would be returned. Text systems match and tag systems filter.

If one is to use a tag-based strategy, what tags should be used? Remember that the real purpose of tags is to aid in search and retrieval. There are several tagging schemas in existence, including the IMS and the Dublin Core. Which of these schemas should be used? Would unique tags help search/retrieval? The eventual goal of the research started in this dissertation is to be able to answer such questions. For now, however, I compare two very different systems – one tag-based and the other text-based.

In this chapter, I explore the reasons and methods for developing a unique tagging schema in order to develop tags that specifically support the performance needs of a particular population – in this case, P-12 educators.

Two problems with Existing Tagging Schemas

The work of those involved in promoting the use of tagging schemas is a great boon to those in education. They have provided a means to cope with the tsunami of
online information. However, two problems threaten the usability, and so, the
implementation of tagging as a general practice. The first is an object orientation for tags.
The development schemas (what tags should be included), is usually focused on
describing the objects that will be tagged in many different ways – the title, author,
media, order of use, etc. The focus is on objects, not on use – what problems a user may
want to solve. The problem with attempting to describe an object is that there are so
many different features of an object; the result is often a very elaborate schema with a
great number of tags. The problem with a large tag set is one of time and money. Each
tag takes time to add. Each tag requires a judgment call by someone who is
knowledgeable enough to be able to make that judgment call. Each tag decreases
uniformity, as each person will make different tagging judgments. In other words, the
more tags, the greater the cost, the more time it takes, and the fewer the number of people
who have the expertise to be able to tag. So, tagging becomes a very specialized pursuit,
accomplished by experts, not a common task, generally done by everyone who adds an
object to a repository. It certainly limits the number of tagged objects available.

The second problem is compromise. A driving goal with most efforts is to
provide a single, universal specification to facilitate easy exchange of information
(interoperability). There are several schemas already in use and more being created. Most
of these schemas can “map” to each other – sharing some similar tags. Previous attempts
to create universal standards for tagging failed. In the current effort, most schemas either
are variants of or make an effort to map to the IMS specifications (imsproject.org), which
are an outgrowth of the IEEE LOM (learning objects meta-data group- www.ltsc.iee.org).
In order to become “universal”, it is necessary that a specification meets the needs of a
great many people and groups. In order to get groups to “sign on” to supporting a certain specification, the group developing the specification must take note of (if not reflect) the needs, vocabulary, and agendas of key players.

For example, one of the key players in the development of meta-data standards is ADL.net (www.adlnet.org), a group that brought together software developers, content providers, and content consumers to develop their SCORM (shared content object reusability model). SCORM uses the IMS specifications. Software developers and content providers wanted to be able to declare their products “SCORM compliant”. They wanted a simple, stable tag set. Because of this, suggestions of developing means to incorporate unique tag sets focused on smaller groups were initially decried. One may guess that there will be an ongoing tension between groups who have adopted specifications, who do not want to be constantly changing their products and do not want to continuously invest time and money into training their staff to stay abreast of a moving target, and the evolution of a relatively new technology that must change as it meets and adapts to new needs and challenges.

Two additional tensions continue to shape the evolution of meta-data specifications, requiring continuing compromises. The first is how many tags should be included in the schema. The Dublin Core schema is the only tag set that has been adopted as a standard (a “notch above” a specification) The Dublin Core (www.dublincore.org) is a very simple tag set, with less than 20 tags. A tension exists between how simple, but generic, a tag set should be versus how complex, but specific, is useful. This tension cannot be easily nor universally answered. For example, consider a web site on which there is a photograph of a famous statue. Who is the author - the sculptor, the
photographer, the web site designer, the site owner? Having a simple tag set opens the
door to such questions. However, having a tag set that includes tags for ‘created by’,
‘original creator’, and the like adds to the cost and complexity of tagging, reducing the
likelihood that objects will be tagged at all.

The final tension considered here is between universality and uniqueness. The
previous tension concerns the degree of specificity – the number of tags in the set. This
tension considers the nature of the tags. If one builds a tag set for a specific group, such
as teachers, perhaps one can create a tag set that more closely answers the needs of the
target population. It may also be possible to limit some of the tags to a certain list of
words – a “controlled vocabulary”. This would certainly increase uniformity in tagging.
However, such a thing would not be possible for a tag set that tried to be all things to all
people, even with a great number of tags. The tension is between a more targeted
approach, with much more limited appeal versus a more generic approach with broad
application. For example, including a tag for pedagogy would make sense in a schema
designed for teachers, but would have no place in a universal schema. Dublin core is a
very generic, simple, universal tag set. It is used for learning objects, museum pieces,
web documents and more. The IMS specification is a more targeted tag set, with more
specific, detailed tags, aimed at tagging objects that focus in some way on learning.

A Model for Developing Unique Tags

Most of the work done with meta-tags has focused on tags that serve a large target
population. In this research, I have continued the move toward specificity taken by IMS
that developed a specification focused on learning objects. What kind of tag set would be
useful for a group with relatively homogenous information needs – in this case, K-12
science and math educators? The research in this chapter studies a model that was used to develop tags that describe possible user needs (user-centric) instead of using an object-centric approach to development. This was the process by which the tags in the tag-based system were developed.

In order to develop tags for the purpose of aiding search/retrieval of documents/objects to answer questions or problems posed by a group with homogeneous information needs (in this case, P-12 educators), I used HPT (human performance technology) and Sensemaking for theoretical guidance.

**Human Performance Technology**

HPT is a field within instructional design, which offers guidance in developing interventions that increase performance. When there is a “performance gap” – a difference between targeted performance and actual performance – a traditional instructional systems approach will suggest that the solution lies in a training intervention. However, HPT looks to other possible causes for the gap. For example, when one considers the problem of drivers exceeding the speed limit, the problem is not one of training. Drivers know how to read the signs (one hopes) and know how to operate a vehicle at different speeds. This is a problem of motivation.

Rossett classifies causes of performance problems into four areas: Lack of skill and/or knowledge, lack of motivation, flawed incentives, and flawed environment (Rossett, 1996). Gilbert highlights the importance of the supporting environment in seeking ways to improve performance (Gilbert, 1996). Gilbert looks at information, instrumentation, and motivation in the environment and in a person’s repertory of behavior.
In order to study the person, their performance, and their environment, HPT has developed a rich set of methods and tools. These include observation, interviews, focus groups, questionnaires, brainstorming and more (Hackos & Redish, 1998; Jonassen, Tessmer, & Hannum, 1999).

The use of HPT techniques within this model grounds the data collection with users, their tasks, and their environment. This shifts the focus from trying to describe the specifics of an object to gathering information about user needs. Developing tags must start with users – interviewing them about the problems they have as well as observing them in the environments in which they work.

Sensemaking

Sensemaking is a theoretical perspective that examines the cognitive processes undertaken by individuals and groups when faced with a situation or information that does not fit into expectations – a “jolt”. Sensemaking has been used to explore how organizations interpret their environment and how this interpretation process influences strategic behavior (Schneider, 1997), the behavior of airplane pilots during events proceeding a crash (Weick, 2001), and the process of creativity, both individually and organizationally (Drazin, Glynn, & Kazanijian, 1999). “The sensemaking process both draws upon cognitive schemas as a guide for action and updates these cognitive schemas in making sense of experience” (Morrison, 2002). The emphasis in Sensemaking is on “understanding the processes through which individuals and organizations develop systems of meaning about creative action (Drazin et al., 1999) Weick identifies seven attributes of the Sensemaking process: 1. Grounded in identity construction, 2.

Using Sensemaking as a theoretical lens to both gather and analyze data roots the activity in examining the purpose the user has for seeking information as well as the type of information the user wants and expects. Our work centers on performance support systems – systems that provide answers to problems, as opposed to generalized searches for information. While one occasionally has time to “cast a wide net” to find tangential information about a topic, one more often searches for objects because of a need or a problem. This type of searching may be viewed as a Sensemaking activity – finding information and using it to develop the user’s frame of reference to be able to solve the problem or fill the need that initially stopped the smooth flow of performance. For example, a teacher, tasked with teaching a lesson in the phases of the moon may not have a ready explanation or may need supporting illustrations. The teacher has a need, has a perception of what they want, and begins to search to find those objects that can provide information to “make sense” of the task – filling in the gaps in their understanding. The teachers is engaged in a process that includes “place of items into frameworks, comprehending, redressing surprise, (and) constructing meaning…” (Weick, 1995).

So, we need to explore what problems bring the target population to search for information, what types of information they believe they need, and to what use they will put that information. If one develops tags that support this process, the search may be more direct and more effective.
Specifics of the model

To help the reader understand the application of this model, I offer a brief walkthrough of my considerations as I use the model.

Using HPT methods (observation, interviews, document analysis, and task analysis) I attempt to understand the population, the environment, and usual tasks. I use Sensemaking to guide data collection and analysis in order to bridge the gap between the user/environment/task and the problems/reasons a user might come to the site.

Sensemaking begins with a jolt. During the course of activity, the jolt enacts Sensemaking behaviors. I try to clearly explicate the frame of reference of the user when they come to the site. What is their problem? What is their need? What do they want? How will the site support their Sensemaking activities?

Once the data on the users and needs has been collected, I use Sensemaking to guide analysis. I use several poster-sized sheets topped with Weick’s seven attributes of Sensemaking (see above) and the question, “Who Am I?” As data is analyzed, I enter specific statements of needs, purposes, problems, and/or desires, always attempting to be clear and concrete.

When all the data has been entered, I look for clusters – similar types of information needs. I have done this with three very different target populations now and in each case most of the types of information needs lend themselves to clustering. There will often be some outliers, and a decision must be made (dependent on the specific case) as to whether these outliers should be imperfectly classed in another group, discarded, or remain as independent categories. Each category then becomes a tag. In appropriate
categories, the specific summaries within the category become the controlled vocabulary for the tag. Specific examples will be provided in the results section.

Methods

Interviews were conducted in two phases; the first phase consisted of 13 on-site visits, combining observation of the work setting as well as an interview. The second phase was comprised of nine telephone interviews. The researchers used the model to both gather and analyze data and developed a set of unique tags. The tag set will be discussed in the results section.

Phase One Interviews

In phase one, the client organization wanted to better understand teachers’ needs and problems, as well as finding ways to both increase site usage and make search/retrieval easier through unique meta-tags. We proposed conducting in-person interviews in order to identify 1) why the users might come to the site, 2) problems and challenges encountered in their job for which the site might hold an answer 3) current uses for the web 4) other sites that were useful. We felt it important to have a series of face-to-face interviews not only to gain a feeling for the environment, but also to help the developers identify more closely with the daily work of the users. We decided to interview between nine and twelve teachers, two university researchers and one or two administrators. Further, it was decided attempt to recruit respondents largely from poor, rural schools, as the educational lab had a vested interest in developing ways to bridge the gap between rich, urban districts and poor, rural districts. Because of time and cost considerations, respondents were recruited only from schools within driving distance of the researcher’s location (a convenience sample). Respondents were recruited through
email solicitations by researchers and their colleagues to known teachers and computer coordinators.

Thirteen interviews were conducted with teachers, administrators, tech coordinators, a pre service teacher, and a researcher. Of that thirteen, two were administrators and seven respondents were in rural schools. All interviews were conducted face-to-face. All respondents were computer users and were comfortable with technology. Questions included brief demographics and usage, identification of current uses of computers, identification of problems in their work and possible solutions. In addition, in order to provide a richer understanding of needs, rather than asking, “What kinds of services would you like?”, a list of potential services were presented and respondents were asked to identify services they would value.

Phase Two Interviews

After the first phase of interviews, the client organization decided that it was not in a position to consider new services, nor was it ready to prioritize outreach to a new user population. They wanted to focus on improving the user experience and ease of search/retrieval for existing users. This brought the project to exclusively concentrate on the development of unique meta-tags. The interviews for current users were focused on how they used the site, what problems they were trying to solve when they used the site, search techniques, aspects of other sites they liked, and which aspects of the site they liked and did not like. Nine interviews were conducted by telephone. The subjects were identified by the client organization as administrators or trainers and current users of the site. Two telephone interviews with client managers were conducted and detailed
interviews and observation of one manager, who was the main contact for the project, extended over several sessions.

Results of Tag Development Model

Phase One Interviews

In the first phase of interviews, the needs expressed were far reaching. However, most respondents asked for information objects that they could “get and go”. Some wanted online tools such as grade books, rubric makers, and test generators. Nearly all wanted lesson plans, but written by teachers with annotations by teachers of what worked and what did not.

I look for lesson plans by other teachers, notes from other teachers for presenting a topic, feedback from other teachers, and stories from other teachers to see if they are going through the same things. I found the journal of one guy who is teaching in Japan. (Bill. First year science teacher.)

Another need expressed was the sharing of information both between teachers and from “above” on such hot topics as assessment and standards. Again, the desire was for very practical application. Said another teacher about assessment,

That’s in the back of my head all the time…where are we going and will I be able to do it. How are they assessing how we are meeting those standards? I would like to talk to other teachers. Are we the only ones going through this? I don’t think we are. (Jean. Third grade teacher.)

Another common request was for guides or answers to applying technology in the classroom, again with a regular preference stated that these guides and experiences be
from actual classroom teachers, not experts. Suggestions included, “How I used a color printer to help my teaching”; “How to use hand held computers and probes in teaching science”; and “How to use specific technology to teach specific standards”.

First year teachers had extra concerns. What of all the “required” paperwork was important? How long does a lesson take to present and how should one proceed if it takes longer or shorter than expected? Discipline was a concern for all pre-service and new teachers interviewed. A first year teacher echoed the concerns of other new teachers not knowing what was important to test and really wanting the guidance that doesn’t come from teacher training, but from experience:

I’d like to hear about how to go about evaluating tests. Do I make good tests? I hate writing multiple choice tests. (Joyce. First year teacher.)

A high school science teacher who had just been assigned a new class was overwhelmed with trying to find useful information for the class and said:

Right now, I need anything that will make a lesson relevant. I don’t like the text book, but they have to have some kind of resources.. written resources... web sites they can go to. (Hannah. High school science teacher.)

A math teacher/coach who complained about the time he spent grading, not teaching also wanted more specific lessons to share.

Most of the sites I see are very general. I would like to see a site that says, “Here’s an idea.. now use it this way.” Show me how to do it. I need new ideas. I can come up with some, but it comes more natural for other people. (John. Junior high math teacher.)
Finally, a common need expressed was help in sorting through the plethora of information available. Lesson plan sites abound, however, getting a good lesson plan that can be effective with little additional work is still a rarity.

The lesson plans sites are getting hard to get through. You click on 6th grade, then social studies and you get 150 lessons. Some are complete, some don’t work…anyone can submit the lessons.. so they aren’t consistently listed. (Adele. Tech coordinator.)

In order to develop a richer understanding of user needs, respondents were asked to value a list of services. They were asked to choose five, rank them, and to note if there were any on the list that would either definitely never use or might consider using.

Researchers assigned a value of 1-6 (1 for no response) to each response, then calculated mean and mode. Table 1 shows the list of services and rankings. Most notable were the items ranked as by mode by mode (5 services tied for highest ranking) that were very also low when ranked by mean. Discussion boards (ranked # 11 by mean 11, yet #1 by mode) and grant resources (ranked #12 by mean, yet also #1 by mode) both rated as services teachers often said they wanted, but never rated as high need. However, the low placement in ranking by mean coupled with the relatively high cost of participation (both activities are time consuming) makes this researcher believe that, at least for this group, discussion boards and grant resources would be relatively unused in practice. It is also interesting to note that the top four items when ranked by mean scores (lesson plans, articles, pictures, guidelines for standards) are all discrete objects, things that a teacher can find, evaluate and download in a relatively short time. While suggested functions
were not adopted by the client organization, the information gleaned from this exercise was very useful in understanding users and user needs.

Table 1. User rating of proposed functions

<table>
<thead>
<tr>
<th>Function</th>
<th>Mean rank</th>
<th>Mode rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discussion boards to talk with other teachers/administrators/tech coordinators</td>
<td>11</td>
<td>1*</td>
</tr>
<tr>
<td>Web casts with experts – educational, or from the department of education</td>
<td>8*</td>
<td>5*</td>
</tr>
<tr>
<td>Research reports on the latest teaching techniques (Inquiry, multiple intelligences, etc.)</td>
<td>4</td>
<td>1*</td>
</tr>
<tr>
<td>Lesson plans</td>
<td>1</td>
<td>1*</td>
</tr>
<tr>
<td>Pictures for download</td>
<td>2*</td>
<td>5*</td>
</tr>
<tr>
<td>Video for download</td>
<td>10*</td>
<td>4*</td>
</tr>
<tr>
<td>Tutorials on software and hardware</td>
<td>6*</td>
<td>4*</td>
</tr>
<tr>
<td>Reports on what happened at educational conferences</td>
<td>7</td>
<td>6*</td>
</tr>
<tr>
<td>&quot;Ask an Expert&quot; columns</td>
<td>5</td>
<td>5*</td>
</tr>
<tr>
<td>Online encyclopedias</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>Articles on using technology in support of teaching</td>
<td>2*</td>
<td>2</td>
</tr>
<tr>
<td>Guidelines on applying standards</td>
<td>3</td>
<td>1*</td>
</tr>
<tr>
<td>Ask the Board of Education column</td>
<td>10*</td>
<td>6*</td>
</tr>
<tr>
<td>Discussions on discipline</td>
<td>6*</td>
<td>6*</td>
</tr>
<tr>
<td>Resources for getting and writing grants</td>
<td>12</td>
<td>1*</td>
</tr>
<tr>
<td>Links to other web sites</td>
<td>8*</td>
<td>5*</td>
</tr>
</tbody>
</table>

Phase Two Interviews

The interviews completed during the second phase were shorter and more directly linked to user needs during search and retrieval. The researcher tried to understand if this user group had different needs, came with different problems, and had different expectations for making sense of the results than the teachers interviewed in phase one. The questions tried to expose the reasons the respondents came to the site. Questions
included: “Consider a recent time you went to the site. What were you looking for? Why did you need that information? How did you look (what search words)? Did you find it?”

The client manager interviews were an effort to make sure that our work was in line with the perceived needs of the client organization both in the present and in the future.

The most common reflection voiced by these users was that the information available was fine, but getting to the information was a challenge. The site has a number of different project sites cobbled together. Most useful information was not found without knowing beforehand that it was there. Illustrative comments include:

There is no indication of what is new. I DO know there's a lot that is old.

The site has enough information, but it's getting at it. When I go to (the site), I usually know what I want. (Gene, District Technology Coordinator)

Another common suggestion was having information accessible by the purpose or topic sought by the visitor to the site. Administrators often come with a specific purpose and would like to be able to request answers for a specific need. Common needs mentioned were guides, facts, models, and diagrams to explain or justify new directions in technology or pedagogy.

I think you have to think about how you will present all this to the community and to parents. Why are we doing this? Why will teachers and science do it different. We have to make it so they understand. They may come home with different homework. Its not that they aren't learning. They do more with hands on.
Parents HAVE to be considered. It's one group we forget about. They will
back you if you show that you want to show them. (Ellen. Director of
technology training for a large museum.)

Jane (consultant/trainer for several districts): Things should be accessible
by topic or purpose.

Interviewer: What’s an example of purpose?

Jane: I’m a principal and I need to explain how technology can be used to
help us meet standards. I want research, models, applications.

There were many positive comments about the site and the utility of the
information it contained. Users noted the wealth of research, the links to other sites
(though many were out dated) and the reliability of the information. Information on the
site has been screened, impartially screened and copyright cleared. This was mentioned
as very important.

The APM (a library of copyright free graphic images) is the only graphics
site that our screening software will allow teachers and students to access.
We installed the screening software and suddenly teachers had no access
to pictures. No other site screens for improper pictures. (Jim. Tech
coordinator for 25 schools)

The best thing about --- is that they are not commercialized. So their stuff
has more validity, for me and people I'm showing it to. If you show stuff
from other sites, even sites that are not product sites, but that do accept
ads, they will say it’s sponsored. ---- has no special interest. (Jill. Training coordinator for science research center.)

Summary of Phase Two Interviews

As with phase one, the interviews conducted during phase two provided information about the types of problems users sought to solve through a search. Questions asking about the last search completed were designed to move the participant from hypothetical interactions to responses rooted in practice. These users tended to seek more general information about a topic instead of the very direct – “how do I…” type questions common in the first group.

Applying the Model

Having used HPT and Sensemaking to guide data gathering, I now analyzed the data using a Sensemaking lens. I posted Weick’s seven properties: 1. Grounded in identity construction, 2. Retrospective, 3. Enactive of sensible environments, 4. Social, 5. Ongoing, 6. Focused on and by extracted cues, 7. Driven by plausibility rather than accuracy (Weick, 1995), on a poster-sized paper. Atop several more pages, we wrote, “Who Am I?” Going through the data, I listed explicit reasons users would come – what problems they faced and what needs they had. Table 2 shows the list generated.
Table 2. “Who am I?” derived from educator interviews

<table>
<thead>
<tr>
<th>Who Am I?</th>
</tr>
</thead>
<tbody>
<tr>
<td>I am an administrator or principal.</td>
</tr>
<tr>
<td>I need to justify a decision to use technology. I need numbers, quote</td>
</tr>
<tr>
<td>I'm a trainer doing professional development on how to use</td>
</tr>
<tr>
<td>I want to know links for other information and research.</td>
</tr>
<tr>
<td>I'm trying to find people.</td>
</tr>
<tr>
<td>I'm trying to find references.</td>
</tr>
<tr>
<td>I write grants and to fund technology and need statistics.</td>
</tr>
<tr>
<td>I defend technology to the school board.</td>
</tr>
<tr>
<td>I am BUSY !!!</td>
</tr>
<tr>
<td>How do I keep up with technology?</td>
</tr>
<tr>
<td>How does technology help me meet standards?</td>
</tr>
<tr>
<td>How does technology help me do assessment?</td>
</tr>
<tr>
<td>I need to do a PTA presentation.</td>
</tr>
<tr>
<td>I want to use technology to help literacy.</td>
</tr>
<tr>
<td>Can I use technology for assessments?</td>
</tr>
<tr>
<td>I'm a principal who wants to set preparation guidelines for teachers</td>
</tr>
<tr>
<td>What have other districts tried? Best practices - What works -</td>
</tr>
<tr>
<td>I need to draw up a technology plan for my school... district.</td>
</tr>
<tr>
<td>I need a fast bit of information - where is. who is. what?</td>
</tr>
<tr>
<td>I am developing a literacy program.</td>
</tr>
<tr>
<td>I am evaluating a literacy program.</td>
</tr>
<tr>
<td>I have to show an administrator what I'm talking about.. I need</td>
</tr>
<tr>
<td>I have to teach a worship on engaged learning.</td>
</tr>
<tr>
<td>I want to explain engaged learning.</td>
</tr>
</tbody>
</table>

I then began to cluster the list – removing duplicates and listing similar needs together. I had always assumed that adding unique tags to existing standard tag sets (Dublin Core or IMS) would significantly increase the number of meta-tags necessary to label an object. Indeed, it had been a worry for future development. I wondered how could we hope to have users bring objects into the system themselves if doing so meant an arduous process of adding many tags, both standard and unique? However, as we looked at what tags were necessary to answer the needs of someone coming to the site with a problem, the result was significantly fewer tags, not significantly more. Indeed, in
our beta version of the meta-tag schema, there were a few general “housekeeping” tags
(location, short name, description, etc.) which mapped to both Dublin Core and IMS tags,
but to get from “I have this problem.” to “Here are a list of probably useful objects.” took
only five tags plus a text field. These were refined during a daylong meeting with the
client, but the number remained the same. None of the tags is required for either
searching or adding an object into the system, but we believe that using even one can
make a search more effective. The five unique tags developed were:

1. Who am I (what role, such as teacher, administrator, researcher).
2. Content area (if they are searching for an object which applies to math, science,
   literacy, etc.).
3. Type (What kind of a thing are you looking for? A model, an article, a video, a
tool for analysis?)
4. Usage (What are you going to do with it? I need to plan a professional
development day. I need to present to the PTA.)
5. Grade level (Some objects, such as lesson plans, are grade specific).
6. Since completing this research, based on user tests and feedback, I have added
two additional tags.
7. Quality (a subjective rating 1-5 of the quality of the object).
8. Completeness (a 1-5 rating, as many web resources lack significant parts or
   contain bad links).

The simplicity of this schema is an extremely important result. If this schema
proves useful for search and retrieval (which must be explored in a future study), it means
that it is reasonable to consider that a group of users would regularly add their own objects to a performance system.

The combination of HPT and Sensemaking provided direction throughout the data collection and analysis. HPT encouraged us to look to the environment to see how I could most easily affect performance. Gilbert talks of the importance of looking at accomplishment over behavior (Gilbert, 1996). This orientation was reflected in the question asked of respondents, “What do you want to do (with the system and within your work)?” Sensemaking helped guide the development of the tags further, asking “What is your problem? Why have you come here? What are you missing?” In combination, I gathered information about people, not objects. I sought to see what their needs and problems were, then I considered the actions they would take to make sense of situations, the type of searches they would like to undertake.

We see the importance of keeping focused on the end user – their needs, problems, and perceptions. Losing sight of the end user is too easy and too common. The end user must drive the design and development throughout the process of analysis, design, development, implementation, and evolution. Keeping this in mind helped to guide us to and through the completely unexpected finding that fewer rather than more tags were needed for the unique schema.
Chapter 5: Results

Description of Participants

Recruiting took place over an eight-week time from December 2003 through February 2004. Over 700 emails soliciting participation were sent out, resulting in 63 educators agreeing to take part. Participants were able to complete the questionnaire at their own speed and convenience, stopping and returning whenever their schedule permitted. Data was collected from the beginning of January 2004 through the end of February 2004. Of the original 63 participants:

Table 3. Number of Participants

<table>
<thead>
<tr>
<th>Participation</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Never started</td>
<td>11</td>
</tr>
<tr>
<td>Completed: lacking data for 2 searches</td>
<td>1</td>
</tr>
<tr>
<td>Completed: lacking data for 1 search</td>
<td>7</td>
</tr>
<tr>
<td>Completed: data for all searches</td>
<td>27</td>
</tr>
</tbody>
</table>

The following demographics were collected on this group.

Table 4. Demographics of Participants

<table>
<thead>
<tr>
<th>Jobs</th>
<th>Number</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teachers</td>
<td>24</td>
<td>70.6 %</td>
</tr>
<tr>
<td>Administrators</td>
<td>3</td>
<td>8.8 %</td>
</tr>
<tr>
<td>Librarian/Media</td>
<td>4</td>
<td>11.8 %</td>
</tr>
<tr>
<td>Other</td>
<td>3</td>
<td>8.8 %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grade level taught by participants</th>
<th>Number</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-3</td>
<td>3</td>
<td>8.8 %</td>
</tr>
<tr>
<td>4-6</td>
<td>8</td>
<td>23.5 %</td>
</tr>
<tr>
<td>7-9</td>
<td>2</td>
<td>5.9 %</td>
</tr>
<tr>
<td>10-12</td>
<td>11</td>
<td>34 %</td>
</tr>
<tr>
<td>Multiple</td>
<td>6</td>
<td>17.6 %</td>
</tr>
<tr>
<td>Did not answer</td>
<td>4</td>
<td>11.8 %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Years Teaching</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - 5 years</td>
<td>4</td>
<td>11.8 %</td>
</tr>
<tr>
<td>Number of Documents Selected</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>While any number of documents (from 0 – 20) could be returned for any search, and respondents were allowed multiple searches, respondents were constrained to selecting a maximum of 7 documents and a minimum of one document per search to be evaluated. When searching, a short description of the document was displayed. Based on that information, respondents decided whether or not to select the document. A majority of respondents did not select 7 documents for searches. Fifteen respondents selected only one document for at least one search. In total, 22 searches resulted in 1 document being selected. The average number of documents per search was 3.27. This trend was more apparent in the tag-based searches with 17 searches resulting in one document being selected. There were only five searches of one document using the text-based search tool. Overall, the average number of documents selected was 2.94 per tag-based search and 3.54 per text-based search.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The number of documents selected particularly impacts the precision and recall measures. These measures were calculated based on the documents selected, not on the total number of documents returned. As stated earlier, precision is the percentage of relevant documents in the search results (in this case, the documents selected) and recall is the percentage of total relevant documents in the document set that are retrieved in a
search. A smaller number of documents selected will, of course, create larger differences in these measures. For example, if the participant only selects one document and it is relevant, then the precision rating is 100%. If it is not relevant, the precision rating is zero. If two are selected and one of them is relevant, the precision score is 50%. If seven are selected with the same one relevant document, the precision score is 14%.

The researcher did not collect data that detailed reasons for the number of documents selected. However, some respondents’ comments at the end of the search (in response to the question – “What did you like or dislike about the search tool?”) indicated either an inability to find many documents or a dislike of the documents returned (indicating that a number of documents were returned, but the respondent did not select them). This anecdotal evidence was limited to ten respondents and was evenly divided between the two reasons stated.

This limited number of documents selected reflects the personal, task-based nature of the search. Each document was displayed in the search with a short description, so respondents were able to select documents they felt were worthwhile to them to accomplish their particular task.

Table 5. **Number of Documents Searched**

<table>
<thead>
<tr>
<th></th>
<th>Number</th>
<th>Text</th>
<th>Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Searches selecting one document</td>
<td>22</td>
<td>5</td>
<td>17</td>
</tr>
<tr>
<td>Number of People with one document searches</td>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average number of documents per search</td>
<td>3.3</td>
<td>3.5</td>
<td>2.9</td>
</tr>
<tr>
<td>Average documents viewed for Utility</td>
<td>2.8</td>
<td>3.1</td>
<td>2.5</td>
</tr>
</tbody>
</table>
A feature of the utility measure is that respondents can indicate when they would stop reviewing documents, reflecting a more realistic searching behavior, where one stops looking at documents after finding an adequate answer. In this research, respondents were asked to indicate after reviewing each document whether they would like to continue reviewing documents, or if they had enough information to stop searching. While they were then asked to continue reviewing all documents that they had selected, the utility score reported is based on documents up to the point they declared that they would stop given the choice. There was a difference in number of documents analyzed. Overall, while searches resulted in an average of 3.3 documents per search, the utility measure was based on an average of 2.81 documents per search, or .5 a document less using when using utility measures.

There was also a difference between text and tag based searches in number of documents reviewed (before the respondent indicated that they would stop) for utility measures. While respondents selected an average of 2.95 documents per each tag search, they indicated for the utility measure that they would have actually looked at an average of 2.5 documents a search. Respondents selected an average of 3.59 documents per search using the text system, but identified 3.1 documents that they actually would have reviewed for the utility measure for text-based systems.

*Explanation of Descriptive Statistics*

Table six below gives an overview of the different measures, including the traditional measures of Precision and Recall, reported as a percentage, as well as the new measures derived from multiplying the construct measures P/R, Utility, and Information Need by the construct measure Satisfaction.
Recall/ Precision reported as a percentage: The traditional measure

The first four lines in Table 3 are the traditional measures – a paired measure for each system reported in percentage. This is the current, ubiquitous measure used to compare systems. So, for this research, the average precision (the percentage of retrieved documents that are relevant – a measure of how “noisy” the results are) for the tag-based system is 51.13% and for the text-based system is 43.85%. By this measure, tag-based systems are more accurate than text systems.

However, the differences between systems in this study are not statistically significant. An independent samples t-test was conducted to compare the precision scores
for text and tag systems. There was no significant difference in the precision scores for tag systems (M=51.128, SD=32.99), and text systems (M=43.85, SD=28, t(66)=.981, Sig=.33). The magnitude of the differences in the means was very small (eta squared=.014).

For recall, the measure of what percentage of total significant documents was retrieved, the systems are even closer, with text-based being slightly better (37.11%) than tag-based systems (35.48%). Again, the differences are not significant between tag systems (M=35.48, SD=23.1), and text systems (M=37.11, SD=21.8, t(66)=-.299, Sig=.766). The magnitude is very small (eta squared-.001).

This data is included in this report, in this way (using t tests instead of analysis of variance) because it is the information that would be presented in the traditional measure of a system. In addition, as there are only two measures, a t test is the most appropriate statistic.

Table 7. No significant difference between system means using Recall measure

<table>
<thead>
<tr>
<th></th>
<th>Levene's Test for Equality of Variances</th>
<th>t-test for Equality of Means for Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>Sig.</td>
</tr>
<tr>
<td>Recall as a percentage</td>
<td>Equal var. assumed</td>
<td>.002</td>
</tr>
</tbody>
</table>

Table 8. No significant difference between system means using Precision measure

<table>
<thead>
<tr>
<th></th>
<th>Levene's Test</th>
<th>t-test for Equality of Means for Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>Sig.</td>
</tr>
<tr>
<td>Precision as a percentage</td>
<td>Equal var. assumed</td>
<td>1.163</td>
</tr>
</tbody>
</table>
Development of New Constructs for Evaluation

As detailed in Chapter 3, this research examines new constructs for evaluation, specifically Information Need and Utility. The measures of these two constructs produce a score on a scale of 1-7, with 1 being best. To more easily compare the traditional paired percentages reported for Precision and Recall, a new, combined measure (P/R) using a 1-7 scale was developed.

The measures of these three constructs were used to compare the two systems (text and tag) using a two way between groups analysis of variance (ANOVA). While the means did show a slight difference in scores between systems, the difference in means were not statistically significant, just as the difference between the two systems using traditional measures reported in percentages were not significant.

Table 9. ANOVA of P/R, UT, and IN showing no significant difference

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of $^2$</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig</th>
<th>Eta $^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected Model</td>
<td>10.96(b)</td>
<td>5</td>
<td>2.19</td>
<td>1.42</td>
<td>.22</td>
<td>.035</td>
</tr>
<tr>
<td>Intercept</td>
<td>5335.76</td>
<td>1</td>
<td>5335.76</td>
<td>3457.13</td>
<td>.00</td>
<td>.95</td>
</tr>
<tr>
<td>Measure</td>
<td>10.55</td>
<td>2</td>
<td>5.28</td>
<td>3.42</td>
<td>.035</td>
<td>.033</td>
</tr>
<tr>
<td>System</td>
<td>.01</td>
<td>1</td>
<td>.01</td>
<td>.01</td>
<td>.95</td>
<td>.000</td>
</tr>
<tr>
<td>Measure * System</td>
<td>.41</td>
<td>2</td>
<td>.20</td>
<td>.13</td>
<td>.88</td>
<td>.001</td>
</tr>
<tr>
<td>Error</td>
<td>305.59</td>
<td>198</td>
<td>1.54</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>5652.32</td>
<td>204</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>316.55</td>
<td>203</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a Computed using alpha = .05
b R Squared = .035 (Adjusted R Squared = .010)
Table 10. Descriptive Statistics for 3 measures comparing systems

<table>
<thead>
<tr>
<th>Measure</th>
<th>System</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>P/R</td>
<td>Tag</td>
<td>4.81</td>
<td>1.45</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>Text</td>
<td>4.94</td>
<td>1.12</td>
<td>34</td>
</tr>
<tr>
<td>Info Need</td>
<td>Tag</td>
<td>5.09</td>
<td>1.166</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>Text</td>
<td>5.01</td>
<td>1.316</td>
<td>34</td>
</tr>
<tr>
<td>Utility</td>
<td>Tag</td>
<td>5.43</td>
<td>1.211</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>Text</td>
<td>5.41</td>
<td>1.152</td>
<td>34</td>
</tr>
</tbody>
</table>

It is not possible, using these new constructs to ascertain a statistically significant difference between these two systems. However, there was a statistically significant difference indicated between the measures (sig = .035). This is unfortunate, as the researcher hopes support the claim that the new constructs are easier and “cheaper” to use than traditional measures, but are comparable, so researchers could reasonably consider turning to these new constructs for evaluation. One must note that the effect size is rather small (.03). A Tukey posthoc test indicates that the significant difference is between the P/R measure and the Utility measure.

Table 11. Difference between mean of measures – UT, IN, P/R

<table>
<thead>
<tr>
<th>(I) Measure</th>
<th>(J) Measure</th>
<th>Mean Difference (I-J)</th>
<th>Std. Error</th>
<th>Sig.</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P/R</td>
<td>Info Need</td>
<td>-.1728</td>
<td>.21306</td>
<td>-.6759 -.3303</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Utility</td>
<td>-.5450(*)</td>
<td>.21306</td>
<td>-1.0481 -.0419</td>
</tr>
<tr>
<td>Info Need</td>
<td>P/R</td>
<td>-.1728</td>
<td>.21306</td>
<td>.697</td>
<td>-.3303 .6759</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Utility</td>
<td>-.3722</td>
<td>.21306</td>
<td>-.8753 .1309</td>
</tr>
<tr>
<td>Utility</td>
<td>P/R</td>
<td>.5450(*)</td>
<td>.21306</td>
<td>.030</td>
<td>-.1309 1.0481</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Info Need</td>
<td>.3722</td>
<td>.21306</td>
<td>-.1309 .8753</td>
</tr>
</tbody>
</table>

Based on observed means.
* The mean difference is significant at the .05 level.
Upon consideration, this researcher posits two possible explanations for the lack of distinction between the two systems using these measures. The first is that both systems in this research were searching the same document set which was relatively small (500 documents) when compared to the 4 billion pages that Google has indexed. Searching such a document set might decrease the difference between what documents the two systems would retrieve. This suggestion is given more credence from the finding that the traditional measures of precision and recall also did not indicate a significant difference between the systems. However, developing a significantly larger document set was outside the scope of this research. As part of the methods for this study, each document was individually examined, tagged, and judged relevant or not for each query. Gathering the document set took over eight weeks of work. Research that uses larger document sets uses other, automatic methods for deciding relevance. These methods, such as searching the document set with several search engines and judging as relevant either all documents returned by any system or only those documents returned by every system, present their own issues. In fact, this difficulty with document sets is one of the main advantages of non-relevance-based measures. It was the requirement of having relevance judgments that forced the need for a document set. There was not a tenable solution to this problem.

The second cause suggested for the lack of variance of the measures was that the numbers from all measures tended to be rather closely clustered. The variances observed were too modest to allow a better differentiation in mean comparisons. This consideration does suggest a solution.
New Measures

After studying the data, the researcher developed a set of new constructs by considering the measure of satisfaction reported by respondents. As discussed in the methods chapter, a measure of satisfaction was derived from two questions – one asking about satisfaction with the system and the other asking about satisfaction with the documents retrieved. While I felt that satisfaction was too thin to be a unique measure, it did provide important information as well as an indication of how much a system would be used. A respondent who likes a system is more likely to use that system, experimenting and trying to understand it in order to get the most out of the system. Conversely, if the respondent does not like a system, they tend to give up searching more quickly, often to the detriment of the final set of documents retrieved. So, I believe that satisfaction is an important indicator of system success. Instead of using it as a unique measure, I created three new measures by multiplying a participant’s satisfaction score for a system with each of the existing three scores for that system. These new measures were called P/R*SAT, IN*SAT, and UT*SAT and were calculated for both the tag-based and the text-based systems. These new measures, which resulted in scores ranging from 2.75 to 49 gave a range of data that allowed a richer view of the measures and of evaluation of the systems. These are the measures whose descriptive statistics are displayed in Table 12 below.

Two-way between groups ANOVA

A two-way between-groups analysis of variance was conducted to explore the scores generated for each of the evaluation constructs (P/R, IN, and UT) for each system.
This analysis showed a difference between the text-based and the tag-based systems. The plot (Figure 1) clearly shows the difference between the two systems.

Figure 1. Plot of marginal means comparison between systems using SAT*measures

![Plot of marginal means comparison between systems using SAT*measures](image)

Note that the text system scored better (lower numbers) than the tag-based system. This is the opposite of original expectations, a matter discussed in more detail in subsequent chapters. However, it is important to remember that the main focus of this research is the effectiveness of these new measures. From this plot, the difference between the systems using the measures can be clearly seen.

There was a statistically significant main effect for system F (1, 198) = 8.43, SIG = .004), however the effect size was small (eta squared = .041). This means that a difference between text and tag systems is indicated by this data analysis. In Table 9, the grayed row is the difference between the tag and text system. By looking in the Significance column, one can see that there is a significant difference. Note that this
number does not indicate magnitude of difference, just whether or not it is a significant
different. The next column, eta squared is an indication of magnitude and shows a small
effect (.06 is a moderate effect and .14 a large effect).

The three measures (UT*SAT, IN*SAT, and P/R*SAT) do not account for a
significant difference in scores – F(2, 198) = 1.14, SIG = .32. This can be seen in the row
labeled Measure. This is an indication that each of the three measures is similar in
measuring the difference between the two systems. If that is the case, we may state that in
this case the two new measures (Utility and Information Need) seem to be comparable to
the existent, traditional measures as evaluation measures. All measures show a better
score (closer to 1) for the text-based system. Next, we shall examine a correlation of the
measures to see how strong the relationship between the measures is.

Table 12. ANOVA effects of differences-systems and measures using SAT* measures

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
<th>Partial Eta Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected Model</td>
<td>1320.372(a)</td>
<td>5</td>
<td>264.074</td>
<td>2.224</td>
<td>.053</td>
<td>.053</td>
</tr>
<tr>
<td>Intercept</td>
<td>71021.407</td>
<td>1</td>
<td>71021.407</td>
<td>598.020</td>
<td>.000</td>
<td>.751</td>
</tr>
<tr>
<td>System</td>
<td>1001.573</td>
<td>1</td>
<td>1001.573</td>
<td>8.434</td>
<td>.004</td>
<td>.041</td>
</tr>
<tr>
<td>Measure</td>
<td>271.189</td>
<td>2</td>
<td>135.594</td>
<td>1.142</td>
<td>.321</td>
<td>.011</td>
</tr>
<tr>
<td>System * Measure</td>
<td>47.610</td>
<td>2</td>
<td>23.805</td>
<td>.200</td>
<td>.819</td>
<td>.002</td>
</tr>
<tr>
<td>Error</td>
<td>23514.674</td>
<td>198</td>
<td>118.761</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>95856.452</td>
<td>204</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>24835.046</td>
<td>203</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a R Squared = .053 (Adjusted R Squared = .029)

Correlation

To see how closely related the three measures were, a bivariate correlation was
run. All three measures were correlated within each system, significant at the .01 level.
For tag systems P/R was correlated to IN at .704 (79%) and UT at .806 (80%). These are
both very strong correlations. For text systems P/R was correlated to IN at .862 (80%) and UT at .878 (88%), again strong correlations. So, the three measures of constructs (P/R, UT, and IN) are strongly correlated within both systems.

Table 13. Correlations between measures P/R*SAT, IN*SAT, and UT*SAT

<table>
<thead>
<tr>
<th></th>
<th>P/R * SAT - Tag</th>
<th>P/R * SAT - Text</th>
<th>IN * Sat - Tag</th>
<th>IN * Sat - text</th>
<th>UT * Sat - Tag</th>
<th>UT * SAT - Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>P/R * SAT - Tag</td>
<td>Pearson</td>
<td>1</td>
<td>.195</td>
<td>.704(**)</td>
<td>.286</td>
<td>.806(**)</td>
</tr>
<tr>
<td></td>
<td>Correlation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P/R * SAT - Text</td>
<td>Pearson</td>
<td>.195</td>
<td>1</td>
<td>.311</td>
<td>.862(**)</td>
<td>.282</td>
</tr>
<tr>
<td></td>
<td>Correlation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.268</td>
<td>.074</td>
<td>.000</td>
<td>.101</td>
<td>.000</td>
</tr>
<tr>
<td>Info Need * Sat - Tag</td>
<td>Pearson</td>
<td>.704(**)</td>
<td>.311</td>
<td>1</td>
<td>.420(*)</td>
<td>.936(**)</td>
</tr>
<tr>
<td></td>
<td>Correlation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.000</td>
<td>.074</td>
<td>.013</td>
<td>.013</td>
<td>.000</td>
</tr>
<tr>
<td>Info Need * Sat - text</td>
<td>Pearson</td>
<td>.286</td>
<td>.862(**)</td>
<td>.420(*)</td>
<td>1</td>
<td>.380(*)</td>
</tr>
<tr>
<td></td>
<td>Correlation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.101</td>
<td>.000</td>
<td>.013</td>
<td>.013</td>
<td>.027</td>
</tr>
<tr>
<td>Utility * Sat - Tag</td>
<td>Pearson</td>
<td>.806(**)</td>
<td>.282</td>
<td>.936(**)</td>
<td>.380(*)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Correlation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.000</td>
<td>.106</td>
<td>.000</td>
<td>.027</td>
<td>.000</td>
</tr>
<tr>
<td>Utility * SAT - Text</td>
<td>Pearson</td>
<td>.317</td>
<td>.878(**)</td>
<td>.395(*)</td>
<td>.894(**)</td>
<td>.397(*)</td>
</tr>
<tr>
<td></td>
<td>Correlation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.067</td>
<td>.000</td>
<td>.021</td>
<td>.000</td>
<td>.020</td>
</tr>
</tbody>
</table>

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

There were also some statistically significant, but much less strong correlations between some tag-based measures and some text-based measures. All these were at the .05 level and ranged from .39 - .42. The IN measure in tag systems was correlated with the IN measure in text systems (.42 at .05 level) as well as with the UT measure in text systems (.40). The UT measures in tag-based and text-based systems were also correlated (.40). With this correlation of measures across systems that is rather small (less than 10%), but is statistically significant, it would appear that we are dealing with the same phenomena, but that there is a difference between the two systems, which is the goal of system evaluation.
Chapter 6: Conclusions

This research has reviewed the process of evaluation of two cases – two systems that use very different methods to retrieve information. An attempt was made to emulate an authentic situation, with the information system taking the part of a performance support tool. To begin to consider what has been learned, let us return to the original research questions.

- Can evaluation of information retrieval be improved through the use of multiple constructs?
- Can user-centric constructs and measures provide time and cost effective comparative evaluation of two information retrieval systems?
- Are the non-relevance based constructs of information need and utility useful for evaluation of information retrieval systems?

In both cases, multiple measures clearly provided a richer view into the evaluation of the systems. While analyzing the data, limitations or confusion around each measure was apparent in respondent’s answers. While this researcher would not like to be limited to any of these measures as a single measure, using some in conjunction with another clearly offers a richer view.

The limitations of the traditional constructs of precision and recall were quickly apparent. During relevance judgments, with only two judges, there was a high (over 25%) rate of disagreement. This meant that some tasks had only one or two relevant documents. It was also apparent while reviewing the cases that it was common for respondents to rate a document as useful that had not been classed as relevant. The question – is this document relevant to whom, when, doing what was obvious and
frustrating. In addition, as mentioned previously, the measure of relevance when using
the recall and precision constructs compels the evaluation to take place in an
experimental setting. To calculate these measures, knowledge of the number of relevant
documents is required. That is not possible in an authentic setting. This burden of this
limitation was made clear during the research.

This research has convinced me of the value of in situ evaluation, using authentic
tasks with the needs, drives, limitations, and time constraints of a real user. The
difference between evaluation in these circumstances and in an experimental setting, even
a setting such as this one, which attempted to evoke authenticity, puts one in mind of car
commercials in which a car spins and dances around obstacles with ease while small text
at the bottom of the screen states, “Professional driver on closed track. Do not attempt to
replicate.” A real advantage of the two new measures is that they can be used in authentic
settings.

The Information Need construct provided a good window into the systems. There
was some confusion about these measures evidenced with some participant’s remarks and
scores. In a small number of searches, scores indicated confusion. In some, during the
pretest the users said they had no information need. Either they were confused or they
picked a task to search on that they knew all about. In some, participants rated their
information need significantly higher in the posttest. Unless the process of searching
drove up their information need, this indicates confusion. In previous studies that looked
at information need, the research was conducted with a researcher present to aid the
participant in assessing information need. As I choose to attempt to implement an online
evaluation tool, a researcher was not there to provide that clarification. While a toll-free
number for support was provided, less than 5 calls and less than 10 emails asking for support or clarification were received. Only one of these asked for clarification about this measure. With more refinement of the assessment tool, its instructions, help functions, and interface, this problem may be reduced or negated.

The Utility construct was certainly effective. I particularly liked the explicit question after each document – “Do you have enough information now? Would you stop or continue?” While the Information Need measure also does not require a review of all documents retrieved, Utility explicitly encourages the participant to behave in an authentic manner. The measure seemed clear to respondents. Several did say they would stop after one or two documents. While not quantifiable, this researcher felt a quality to the response this measure provided. The scores seemed more closely tied to the actual practice of searching and the individual documents. It seemed to be more based in respondents’ actions. It was not based on someone else’s view of relevance, like precision and recall. It did not ask the respondent to try to assess and quantify their information need – a relatively fuzzy concept. It asked the respondent to look at each document and say whether or not it was useful and then to try to quantify how useful it was.

Each of the constructs provided a remarkably similar view of the two systems. Indeed, with the correlation so high between constructs within each system (49% - 82% correlation in the tag system and 74% - 80% for text systems) it seems that these are not new constructs. These are new measures for the same constructs. So, instead if instead of identifying the constructs as traditional, information need, and utility, we identify the construct of system efficacy, then this research indicates that, at least in these cases, there
is evidence that traditional, information need, and utility are three measures of the same construct. So, the first research question – “Can evaluation of information retrieval be improved through the use of multiple constructs?” must be answered in the negative. Multiple constructs did not provide a richer view in this research. Instead, it appears that multiple measures were used, and those did give a richer evaluation.

The next two questions focus on application and the two specific constructs used in this study. Of what use is developing new constructs if they cannot be operationalized in a manner that is time and cost effective? Whether strength, foible, or failing, practitioners in instructional systems like to build things and study the building of those things. So, an important part of this research was the development of three new online tools. The case studies examined not the development of the tools, but the evaluation of systems. However, it is appropriate here to discuss the technical achievements.

For tagging and searching for objects, the findings are conflicting. This is the first time an online tool for tagging web-based objects was built, tested, and used extensively. The tagging tool worked extremely well. A screen capture of the tagging tool is in Appendix E. With the implementation of this tool, a number of technical decisions were made, including data structure and interface design. A fundamental question of whether to locate the objects in a database or provide pointers to the URL of the object was decided in favor of providing pointers. So, some very fundamental decisions were made and an important first step was taken.

Using the tool to tag over 500 objects was instructive. On a positive note, it was certainly easy and fast to tag an object. This is a basic question that cuts to the heart of the eventual utility of tagging systems. If it takes too long to tag objects, people will not
tag them. A surprising question that arose during the process was one of uniformity. In all tagging systems, there is a problem with having tagging judgments made with some degree of uniformity, particularly when an untrained group of users is allowed to tag objects. I have suggested the use of controlled vocabulary as a means to try to enforce uniformity. However, during this project, I realized that even with controlled vocabulary, tagging would not be uniform. Even with one tagger, over the eight weeks of collecting and tagging objects, my tagging was not uniform. Such questions as, “If the lesson plan meets a standard, should it be tagged in usage as meeting standards, or should that category be reserved for exercises that explicitly talk about how to ensure a teacher is meeting standards?” were not uniformly answered, even with only one tagger. On one hand, this points to the near impossibility of achieving any kind of uniformity. On the other, it shows the importance of explicit, thoughtful, controlled vocabulary as a much stronger method for encouraging uniformity than text fields with no restrictions. In addition, I doubt that perfection in tagging is either possible or necessary. Even if the same tags went on, people searching would not be uniform in their search behavior. Over time, we will see how damaging this lack of uniformity is to the usefulness of tag-based systems. I began this project an advocate of tag-based systems. I end with less enthusiasm, but am not ready to dismiss them. What is certainly necessary to fine tune tagging systems are constructs for evaluation that allow in situ evaluation.

The second tool built was the tag based search tool. Again, this was the first time that such a tool was implemented and used extensively by many people. Many technical and design considerations were answered, at least for this implementation. While there
are areas that need improvement, including interface design and testing, this is an excellent first step. A screen capture of the tag-based search tool is in Appendix B.

Finally, the design, implementation, use of the online questionnaire tool is a significant achievement. Again, with this new tool, there were some problems. For some respondents, the third search (of four) was not recorded. Some respondents tried to use the back arrow on the browser to go back and select a different task. This resulted in double entries in the data table. Some respondents were never able to get the tool to work, but these were very few. Overall, this tool is an important advance. Because of this tool, it was possible in this research to draw on a very wide population. It is also beneficial to allow respondents to answer questions in the same milieu as they are searching. This tool will allow researchers to have participants search using any search engine or retrieval system from within the system, then have the questionnaire right on the screen at the same time, allowing respondents to engage in authentic information seeking behavior with a very minimal intervention or influence by researchers. Coupling this power with the richer understanding of evaluation and the success in this case study of the two measures that did not require relevance judgments opens the possibilities of being able to study information seeking in situ and, perhaps, eventually evaluation theory being able to be extended to allow “tuning” of retrieval tools for different audiences. These are very exciting possibilities and provide an emphatic affirmative within these cases to the second question – “Can user-centric constructs provide time and cost effective comparative evaluation of two systems?”

The final question addresses the two new constructs – information need and utility. Harter, in 1996, wrote, “Our approaches to evaluation must reflect the real world
of real users”(Harter, 1996). Cooper suggested his naïve method in 1973. Why have no new ubiquitous measures come to the fore? I believe it is because there have been alternatives that are easy, effective, and inexpensive. Non-relevance based evaluation constructs usually involve measures that require researcher observation and interpretation. Precision and Recall have also required “expensive” measures – requiring relevance judgments. If, as discussed above, we view information need and utility as new measures of the construct System Efficacy, then we have powerful new measures, easily operationalized with the online questionnaire tool, to evaluate this new construct. These new measures certainly do provide useful insight into comparative evaluation within the context of these cases, with the advantage of freeing the researcher from the fundamentally flawed relevance decisions and the need to evaluate in an experimental rather than an authentic setting.

The results, particularly the strong correlations between measures within systems and the weaker correlations across systems, certainly indicate that more research is called for. It is not within the scope of this research to generalize outside of these two cases. The methodology does not support any assertion of application outside of this particular case study. However, the findings of this research do begin to build a theoretical foundation for eventual research that may be able to be more sweeping in its conclusions. What this research shows us is that there is certainly reason for further study. There is reason to continue to examine the application and use of these measures in other settings and other systems.

One thing that this research did not do was answer the question that brought me to this research. I believe the jury is still out on tag-based systems. Remember, a tag-based
system is more expensive to implement than a text-based system. It must not be as good, it must be significantly better. While these measures in this case study showed that the text system was better, I believe that further research and development is called for. Based on participant comments, it became clear that several participants had a significant problem understanding and using the tag-based search tool. As this was the first implementation of a tag-based system with a large enough document set of tagged objects to allow extensive search and retrieval, it was not until I was able to use the system extensively and talk with participants and read their comments that I began to understand that tag-based searching is fundamentally different from text based searching. It is a very responsive system where a small change (clicking one or two items) makes a huge difference. Often participants clicked at least one item in every tag. The result was often no documents. Even with explanations and, eventually, an online movie with screen captures of searching using the system, it is not clear that it was the system, not the interface or the instructions that was being tested. More study is indicated.
Chapter 7: Implications

In the weeks since completing this research, I have had many opportunities to explain what I have been working on for the past year. One of the key phrases Weick uses to represent sensemaking is “How can I know what I think until I see what I say?” (Weick, 1995). This has certainly proven to be the case for me. When one is in the midst of a research project, it is impossible to have the ironic distance necessary to judge its significance. One must focus on all the details of getting pieces built and working, recruiting participants, getting approvals, collecting data, and evaluating results. As I have explained and explored my research, I have come to believe that there is a potential for significant advances in the field of information retrieval, as well as important foundational work in the study of tagged based systems as a result of this research. In this chapter, I will first cover some limitations of this research, and then will briefly outline some possible future research. Finally, I will discuss the importance of this research, pointing to what has been learned, trying to place it in context, and suggesting further work that this research suggests.

**Limitations**

As with any research, there are limitations to this study. Some have been alluded to previously. The first is the relatively small document set. Traditional methodology usually uses document sets that contain tens of thousands of documents. This study uses only 500 documents. The reason for this is two-fold. The tag-based system used in this study required that all documents be studied and tagged. To increase the validity of these tags, the tags for each document were examined by a second researcher who is a former teacher. So, each document in the set had to be examined and considered by two different
judges. This made even a document set of 500 documents a significant undertaking. Collecting and tagging the documents took over eight weeks. The second reason was the method used to judge relevance. Researchers using large document sets do not examine each document and make a relevance determination. Different methods are used to automatically assign a relevance judgment including text matching or retrieval overlap with the same query using multiple search engines. These methods introduce their own set of problems and questions about the efficacy of the judging. This researcher wanted to be able to examine each document and make a relevance judgment for each query. In that way, it was possible to ensure that no relevant (topically relevant) documents were “missed”, as has been suggested by researchers looking at the Cranfield tests (which used an automated method of assigning relevance). Again, within the scope of a research project completed by a single researcher with no funding, a document set of 500 was at the upper limit of available resources and was large enough to reasonably consider the results as tenable. The size of the document set may have contributed to the small difference between the two systems.

The second limitation stems from the research method. Because of the convenience method of sampling, the findings of this study cannot be generalized to the population. It is not possible to conclude based on this research that one system is better at returning relevant or useful documents returned than another. It is not possible, based on this research, to claim that the new measures of information need and utility correlate with the traditional measures in any situation except in this case. The intent of this research is to support generalizing to theory, to provide a richer theoretical basis for
evaluation. We cannot generalize, but the results certainly allow us to say that the results are promising and further research would be time well spent.

In addition, the tasks and the situation was not authentic. This was not research completed in situ, with the pressures and drives that would normally drive information seeking behavior. The researcher made an attempt to make the tasks as authentic as possible by basing them on teacher interviews and having teachers and former teachers review the tasks. The research protocol also asked the respondents to rate the authenticity of the task and most tasks selected were rated as authentic to the respondent. However, studying search behavior, particularly when the purpose of the search is performance support – answering a question in the midst of performance - not merely information seeking, in this setting where the respondent does not have the time pressures and the real need for the information that an authentic setting would provide limit the value and reliability of this research. This was a tradeoff between richer data and a larger set of respondents. To study authentic searching would require a much more massive system, with a much larger document set and addressing the problem of data collection without researcher contamination. With current technology, such a study would require researchers to observe respondents over a long period of time, noting and evaluating search behavior as it arose. Perhaps in the future such a study will be possible, particularly with further development of the online questionnaire tool developed for this study.

Finally, as has been mentioned, there were many technical problems, which resulted in frustration for some respondents, loss of some data, and loss of some respondents who were unable or unwilling to complete the research. As this research
required the development of several new technologies and while I am proud of those accomplishments and feel that much was learned, implementing these new tools did come at a cost in frustration and some lost data. Additionally, the fundamental difference between tag-based searching and established practice should have been addressed more clearly. New interface design and testing of instruction sets is called for.

**Future Studies**

The findings of this research do hold great promise for the evaluation of information systems. Enough has been uncovered to make further research worthwhile. With the conclusion that, in this case, the measures of Utility and Information Need are effective in evaluating systems, an important next step is to add to this foundational work by using these measures in evaluating other systems, attempting to replicate the results. It would be interesting to see the measures used with existent large document sets that have been evaluated using precision and recall, again comparing the new measures with the existing standards. Another extension would be to compare these measures to measures of normalized precision and recall to attempt to draw a comparison. Finally, using these measures in situ with different audiences in order to test efficacy in different populations will extend the value and application of these new measures.

Of particular interest for this researcher is the extension of both the measures and the online tool for application in authentic situations. Evaluating search behavior and results over an extended period of time within the actual settings for information seeking would be invaluable to an understanding of systems and of methods of evaluations. I feel that this holds great promise, not only adding to the theoretical foundation, but also in the creation of tools that will allow other researchers to explore evaluation.
Finally, I expect to continue research into tag-based systems. In following sections, I will outline some of the research that this work suggests, attempting cull out independent variables to ascertain a more certain view of the efficacy of tag-based search tools. I see three areas of research: tag development and use; search tool design, development, testing, and use; and research into users - different settings, with different needs, different user groups, and different support/training functions. As tagging is an integral part of the semantic web, questions of how tags are developed and used will continue to be an important area of research. This work, rather than putting questions to bed, has stirred up more questions, more considerations, and more possibilities. Can a researcher wish for anything better?

*A New Era of Evaluation*

The finding that information need and utility seem to measure the same thing that the traditional measures of precision and recall do (a construct I call system efficacy) is very exciting. If future research substantiates the findings that these new measures do provide good valuation of system efficacy, a new generation of search tools is in the offing. Prior to this time, any testing had to be done in experimental settings – with pre-defined queries and a pre-defined and judged document set. An experimental setting is so far removed from the experience of seeking information in situ as to be nearly unrecognizable. When one is seeking information on the web, the searcher brings knowledge, expectations, time-constraints, and personal preferences to the search. None of these can be adequately reflected in an experimental setting. So, the evaluation of information systems has had very little to do with users (as recall and precision really do
not need users, they may be developed from batch testing) and less to do with the reality under which the retrieval systems will be used.

As a result, the tools are really the best guesses of different groups and one relies on the marketplace of users to sort out which is “better”. This set of evaluative tools is gross and clumsy, putting one in mind of stone axes being used for separations. Very fine distinctions are just not possible. However, with the new measures, we may trade in the stone ax for a scalpel.

Both information need and utility rely on user evaluation of systems. They allow in situ evaluation, so the people may evaluate with real questions, searching the web or whatever document set they would usually search, taking into account all the personal expectations, constraints, pressures, and motivations present in an actual search. Indeed, these measures are appended to an actual search – the only difference is that, in addition to searching, the respondent answers a question before and after the search for information need and answers one to three questions for each document for determining utility.

The result of this in situ research is that search engines can be fine-tuned for specific user groups. Decisions within the search engine that include word combinations, ranking, and weighting are no longer just based on best guesses of search engine designers. Instead, by testing the decisions on groups of intended users, the internal decisions can be adjusted to develop engines that are specifically accurate for particular groups. Instead of “one size fits all” search tools, we can have tailored search tools. As in suits, when buying off the rack, these systems fit everyone poorly and no one well. I
believe we now have the potential to throw away our Sears suits and wear nothing but custom tailored garments.

These new measures change the underlying assumptions of what makes a retrieval system “better”. Remember that the assumption underlying precision and recall is that the system that retrieved more documents is the better, particularly if the documents seemed to relate to the query. The assumption underlying information need is that a better system will reduce your need for information from when you start searching to when you stop searching. This intuitively makes more sense as a means to rate the efficacy of a system. Utility has as it’s assumption that a better system returns documents that you find more useful, rating only the documents you look at. So, information need looks at the end result of having used the tool and utility looks at the documents that one examines. Both methods seem important and useful assumptions about what makes a retrieval system “good”.

Evaluation Tool

The potential effect of the new measures of evaluation is significant. In addition, the online evaluation tool that was designed, developed, and used for this research holds great promise. Several measures have been proposed in the literature over the years. However, few offered easy paths to operationalization. To make the leap from the journals of the Academy to actual application, changing the ways information retrieval systems are evaluated, allowing for the fine-tuned search tools described above, it is essential that an easy way of using these new measures is provided. While working on my dissertation, a friend told me that she had completed her dissertation a decade earlier and had put a $20 bill with a note that read, “Keep Me” inside the bound copy in the
library. Recently, she visited the library and found the $20 bill still there. I sincerely hope that this dissertation has more readers. The online tool offers a simple, customizable way to test systems. With it, a participant can do a search and, in an open, second window on their computer, enter their information need and utility ratings. It will be possible to test systems with a minimum of intrusion. With such a tool, testing and retesting a system will be an easy matter. Testing should also be less expensive than using the traditional methods, as there will be no need to gather a document set or to make relevance judgments. Steve Jobs has said that in order for a new technology to be adopted, it must be ten times better than what precedes it. We may be approaching that threshold with new measures and a tool that allows their easy, rapid, and inexpensive use. It is important to note here that these tools can be used to compare any systems. While this research compared text and tag systems, these tools are system independent – one can be comparing different text-based systems, different ranking algorithms within the same system, or different tag sets for the same system, so any systems can be compared and any system can be improved.

_Considerations of Tag-Based Search_

While the results in this study showed the text-based systems to be better, this research was directed at the development of new measures, not real comparisons of the systems. Much was learned about tag-based systems and, because of this study, we can see clear next steps in research that can use these tools to ascertain when and where tag-based search tools may prove useful.

It is important to remember that tag-based search is significantly more expensive than text-based search. For tag-based searching, a document must have been analyzed
and tagged. Text-based searching is automatic. In order for tag-based searching to be
useful, it must not just be better than text-based search, it must be significantly better.
While efforts to build the semantic web rely on tags of one kind or another, it is doubtful
that tagging will ever become ubiquitous. This is not envisioned as a replacement for
current methods of searching the web. The application of tag-based searching, in my
opinion, is for specific groups that share homogeneous information needs. The power is
within a performance support system that allows users to tag objects as they add them to
the performance system, making it fast, simple, and effective to retrieve targeted
performance objects. I can envision an application that allows individuals to add tags to
their own documents on their own computers or networks, making it faster and easier to
find documents. Indeed, I have constructed a prototype web-based system that I used
with my programmer during the development of this dissertation.

The accuracy of tag-based searching did not shine in this research. Indeed, the
setting tended to limit the efficacy of tag-based searching. Text-based systems have
problems with graphics, sounds, animations and other multimedia elements – as they do
not contain words that can be matched. Tag-based systems do not look at the object, but
at the tags, so are stronger when searching for these objects. In this research, I avoided
multimedia objects. In addition, I hypothesize that in larger document sets with more
documents that have similar, but irrelevant terms (When I search for rock, am I interested
in music, geology, motion, chairs, wrestlers, Alcatraz… or something else?) tag-based
searches, that look at the tags, not the text, will show themselves to be more effective.
However, in this relatively small document set, where most documents had something to
do with at least one of the queries, this distinction did not come out.
I did not anticipate how difficult it would be for people to understand a new way of searching. In spite of extensive instructions, there was anecdotal evidence that the tag-based search tool was confusing to respondents. A common action was for respondents to select one or more vocabulary options (presented as check boxes) in each of the tags (for example, a check in the grade level tag of 6th and 7th grades, a check in the role tag of teacher, a check in the type tag of web site and graphic, etc.) The problem with this type of approach is that tag-based searching is very accurate. If one checks an option or two in each of the tags, then the only result will be a document that has ALL of those tags. In a document set of 500, this often resulted in the tag-based search tool returning no documents, which was confusing and frustrating for the respondents. Even with the addition of a screen movie that showed selecting only a couple of options at first, then adding more to continue to filter the results, respondents tended to ignore instructions and use the tool incorrectly. This is certainly a common occurrence for developers. The fault is not in the users, but in the design. Within the scope of this research, it was not possible to go through several iterations of interface testing and redesign. A single cycle of development, interface testing, and redesign was the all that was done. This is a critical question, for there have been many interesting advances in search tools over the years, in particular those that show visual links to other information (such as webbrain and grokker), that are very interesting, but of little impact, because they are not easy to use. Traditional measures, while fundamentally flawed, have been used for nearly 50 years because they are relatively easy to understand and use. Given a text box on a search page, users pretty intuitively know what to do. This is not the case with tag-based searching. A great deal of time and research is required to improve the interface of a tag-based system.
so that we may compare the effectiveness of the information retrieved, rather than the ease of use of the interface. A possible interim step would be research that includes a personal training on tag-based searching to ensure that respondents actually understood this new search method.

It is very probable that the background of the users included searching using some sort of text-based search tool. It is also very probable that users were able to use that past experience to inform their search behaviors when using the text-based search tool. The satisfaction ratings were higher for text-based searching. This may well be because users tend to prefer the familiar. The text-based search tool was nothing new. Users did not have similar experience with the tag-based search tool and were less comfortable, less successful, and less satisfied with it.

This research examined measures of evaluation. In future research to examine tag systems, it is important to try to reduce the number of independent variables that may point to the reason for the differences between systems. So, while tag-based searching did not rate well in this research, possible reasons that bear examination in future research include: the quality of the tags, the granularity of tags, the method of tagging (who, when, and how). In addition, variables concerning use and user may be examined, including: interface, function set, help/support issues (already discussed), the environment of the user when searching, the technological environment, time taken and time available during search, and user background and experience.

Applications to the Semantic Web

In the discussion of the history of information retrieval, the third era was identified as searching for information on the web. However, the web is a rapidly
evolving technology. Much of the consideration of future web development concerns the semantic web. In an article in Scientific American in May, 2001, Tim Berners-Lee and his colleagues provided a view of a web where the information is found and used by software agents, delivering only the end results to users, just as we now consume food, with no regard to the process of planting, growing, harvesting, or, in many cases, preparation. “The Semantic Web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users” (Berners-Lee, Hendler, & Lassilia, 2001).

Agents will read information about the pages that help them understand and use those pages. What is this information that is crucial to the understanding of these documents, and so, to the future development of the Semantic Web? What else could this be but metadata? “…implementing and harvesting metadata is fundamental to the success of the Semantic Web” (Greenberg, Sutton, & Campbell, 2003). Tagging, or metadata, is the means by which the semantic web will function. It is the metadata that will tell the agents what sense to make of the documents. “The Semantic Web aims to provide that guidance in the form of encoded metadata that provides a context for Web-based data…Machine-readable does not mean machine-understandable. Metadata is the proposed solution…” (Bonner, 2002).

So, my underlying question of whether or not objects will/should be tagged is moot. Objects will be tagged, increasingly so. How they are tagged and what tags are used will continue to evolve. My particular interest is in developing tag sets and systems that allow users to tag and add objects into a system. These unique, relatively small document sets would, at least in the near term, be searched by users, not agents.
However, it is of no matter who or what is searching. The tags must result in search/retrievals that are useful to the end user. Having tools with which to test the efficacy of search tools and tag sets become more important as these tools develop. The scenarios used to explain the Semantic Web implicitly return more carefully targeted searches. There are no searches that return several million options. The searches return a few options and all of them are correct matches. This approach is using the web in support of performance, where an answer is often better than a plethora of possible answers. Fine-tuning search engines without evaluation tools that can illuminate how well a search tool (or agent) meets the needs of a user, using the “stone ax” of the traditional methods will make implementation of the Semantic Web difficult, if not impossible.
References


### Appendix A Queries

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Investigate ways to finance computers for your school.</td>
</tr>
<tr>
<td>2</td>
<td>You have been appointed to head the committee making your school’s technology plan.</td>
</tr>
<tr>
<td>3</td>
<td>You must decide whether or not to use filtering software.</td>
</tr>
<tr>
<td>4</td>
<td>Conduct a workshop for teachers on using technology to support instruction.</td>
</tr>
<tr>
<td>5</td>
<td>Teach a class in using search engines.</td>
</tr>
<tr>
<td>6</td>
<td>You want to find information to improve your abilities in writing good tests.</td>
</tr>
<tr>
<td>7</td>
<td>Develop a plan that shows how your teaching (or teachers) are assessing and meeting standards.</td>
</tr>
<tr>
<td>8</td>
<td>Develop a plan to improve my classroom discipline.</td>
</tr>
<tr>
<td>9</td>
<td>Make a web site for yourself.</td>
</tr>
<tr>
<td>10</td>
<td>Choose and use a computer based grade book.</td>
</tr>
<tr>
<td>11</td>
<td>Develop a plan to use handheld computers for science projects.</td>
</tr>
<tr>
<td>12</td>
<td>Develop a plan to use computers to support the writing process.</td>
</tr>
<tr>
<td>13</td>
<td>Develop a plan to use inquiry or discovery science methods.</td>
</tr>
</tbody>
</table>

#### Grade 1-3

| 14 | Teach a lesson that meets the standard - 3.1.8 Describe how discarded products contribute to the problem of waste disposal and that recycling can help solve this problem. |
| 15 | Teach a lesson that meets the standard - Observe and describe that the moon looks a little different every day, but looks the same again about every four weeks. |
| 16 | Teach a lesson that meets the standard - Investigate and describe how moving air and water can be used to run machines like windmills and waterwheels. |
| 17 | Teach a lesson that meets the standard - (5.3.6) Demonstrate that things on or near Earth are pulled toward it by Earth’s gravity*. |
| 18 | Teach a lesson that meets the standard - Describe how waves, wind, water, and glacial ice shape and reshape Earth’s land surface by the erosion* of rock and soil in some areas and depositing them in other areas. |
| 19 | Teach a lesson that meets the standard - (5.3.11) Investigate and describe that changes in speed* or direction of motion of an object are caused by forces*. Understand that the greater the force, the greater the change in motion and the more massive an object, the less effect a given force will have. |

#### Grade 6-8

| 20 | Plan a lab that meets the standard - Find the mean* and median* of a set of data. |
| 21 | Plan a lab that meets the standard - Use technology, such as calculators or computer spreadsheets, in analysis of data. |
| 22 | Teach a lesson that meets the standard - Use models or drawings to explain that Earth has different seasons and weather patterns because it turns daily on an axis that is tilted relative to the plane of Earth’s yearly orbit around the sun. Know that because of this, sunlight falls more intensely on different parts of Earth during the year (the accompanying greater length of days also has an effect) and the difference in heating produces seasons and weather patterns. |

#### Grades 9-10
<table>
<thead>
<tr>
<th>Page</th>
<th>Lesson Plan</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>Teach a lesson that meets the standard - Describe Hubble's law. Identify and understand that the &quot;Big Bang&quot; theory is the most widely accepted theory explaining the formation of the universe.</td>
</tr>
<tr>
<td>24</td>
<td>Teach a lesson that meets the standard - Describe the characteristics and motions of the various kinds of objects in our solar system, including planets, satellites, comets, and asteroids. Explain that Kepler's laws determine the orbits of the planets.</td>
</tr>
<tr>
<td>25</td>
<td>Teach a lesson that meets the standard - (ES.1.24) Understand and discuss continental drift, sea-floor spreading, and plate tectonics. Include evidence that supports the movement of the plates, such as magnetic stripes on the ocean floor, fossil evidence on separate continents, and the continuity of geological features.</td>
</tr>
<tr>
<td></td>
<td><strong>Grade 10-12</strong></td>
</tr>
<tr>
<td>26</td>
<td>Plan a lab that meets the standard - Measure or determine the physical quantities including mass, charge, pressure, volume, temperature, and density of an object or unknown sample.</td>
</tr>
<tr>
<td>27</td>
<td>Teach a lesson that meets the standard - (CP.1.6) - Understand and explain how an atom can acquire an unbalanced electrical charge by gaining or losing electrons.</td>
</tr>
<tr>
<td>28</td>
<td>Plan a lab that meets the standard - Explain how the size and rate of growth of the human population in any location is affected by economic, political, religious, technological, and environmental factors, some of which are influenced by the size and rate of growth of the population.</td>
</tr>
<tr>
<td>29</td>
<td>Teach a lesson that meets the standard - (ES.1.2) Differentiate between the different types of stars found on the Hertzsprung-Russell Diagram.</td>
</tr>
<tr>
<td>30</td>
<td>Teach a lesson that meets the standard - Compare and contrast the evolution of stars of different masses.</td>
</tr>
<tr>
<td>31</td>
<td>Teach a lesson that meets the standard - (B.1.3) Know and describe that within the cell are specialized parts for the transport of materials, energy capture and release, protein building, waste disposal, information feedback, and movement. In addition to these basic cellular functions common to all cells, understand that most cells in multicellular organisms perform some special functions that others do not.</td>
</tr>
<tr>
<td>32</td>
<td>Plan a lab that meets the standard - (B.1.38) Understand and explain the significance of the introduction of species, such as zebra mussels, into American waterways, and describe the consequent harm to native species and the environment in general.</td>
</tr>
<tr>
<td>33</td>
<td>Plan a lesson that answers the standard - (C.1.12) Demonstrate the principle of conservation of mass through laboratory investigations.</td>
</tr>
<tr>
<td>34</td>
<td>Plan a lab that meets the standard (C.1.8) - Use formulas and laboratory investigations to classify substances as metal or nonmetal, ionic or molecular, acid or base, and organic or inorganic.</td>
</tr>
</tbody>
</table>
Appendix B Search tools and sample retrieval

Figure B1. Tag Based Search Tool
Figure B2. **Text Based Search Tool**

A Boolean "and" between words is assumed; that is, documents will be retrieved that have **all** the specified words. If you wish to find all the documents that have **any** word, then use "or" between your search words.

To truncate a word, use an asterisk (*) at the end. For example, the search "comput*" would retrieve documents that have the words "computer", "computing", "computation", etc.
Figure B3. Search Results
Appendix C Walk through of questionnaire

Figure C1. *Optional Contact Information*
Figure C2. Demographics

PLEASE - Do not use your browser’s Back or Reload Buttons - This will confuse the data collection - PLEASE

Today you will be using a search tool to find information to complete a task. This research is looking at ways of evaluating search tools, so you will be asked a series of questions during the process. You will need to have two browser windows open at the same time - one with the search tool and this one with the questions.

If at any time you have questions, call 800-314-1493 or email Steve Schatz at sschatz.edu.

The process may take up to two hours. However, you may save your work at any time and return to it at a later time. We ask that you try to complete the interview within 7 days.

Background questions

These questions are to give a general feeling of your background. If you don't fit exactly into a category, aim for the closest fit possible. Don't concern yourself too much here, it is just trying to put you in a general ballpark.

1. Position. What is your current primary occupation? Teacher

2. If a teacher, administrator, or media specialist, what grade level do you currently work with? Other

3. How many years have you been working in the educational field? 24

4. Have you ever put on a teacher training workshop?  Yes  No

5. About how much time do you spend online a day - on average? 4 hour(s)

6. What search engine(s) do you usually use? Yahoo, Dogpile

7. Do you ever use the advanced search functions?  Yes  No

8. How often do you use the advanced search functions?  Never  Sometimes  Often
Figure C3. **Overview and test search**

**Overview**

For this research, you will

1. Complete a test search to make sure you are comfortable using the system.
2. Select a task from the list provided.
3. Search for documents that will help you meet that task.
4. When you are finished, evaluate the documents you retrieved.

You will be using a standard test based search tool. This search tool will search the documents we have gathered to see if there is a match between the words you type in and the documents. When you type in two or more words, it assumes you want ALL those words to appear in the document. For example, if you type in star, black hole, and telescope, you will ONLY get documents that have all 3 terms. If you want to see documents that have any of these terms, use OR - for example star OR black hole OR telescope.

If you get confused, please call Steve (303-314-1493) or email sechatt@indiana.edu.

Now, try a test search.

**Test Search**

To get used to searching, try to find a document that tells 4-5 graders how to make hand shadows.

Click this button to open the search tool you will be using. [Begin Test Search]

Leave this window open.

You can switch back and forth between the two windows by holding down the alt key and hitting the tab key.

If, at any time, you have questions or problems, call Steve Schatt 303-314-1493 or write sechatt@indiana.edu.
Figure C4. Click to open second window of tasks. Click button beside task to select.
Figure C5. First Search. Respondents randomly assigned to begin with text or tag based search.
Figure C6. Text based search tool

A Boolean "and" between words is assumed; that is, documents will be retrieved that have **all** the specified words. If you wish to find all the documents that have **any** word, then use "or" between your search words.

To truncate a word, use an asterisk (*) at the end. For example, the search "comput*" would retrieve documents that have the words "computer", "computing", "computation", etc.
Figure C7. First search results

<table>
<thead>
<tr>
<th>Title</th>
<th>Description</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>How Search Engines Work</td>
<td>Page with explanation of how search engines work. Somewhat hard to read, but a lot of good information. No lesson plan.</td>
<td>100.0%</td>
</tr>
<tr>
<td>Thinking Critically about World Wide Web Resources</td>
<td>Short (1 page) list of things to consider when looking at web content, including: Content &amp; Evaluation, Source &amp; Date, and Structure.</td>
<td>92.6%</td>
</tr>
<tr>
<td>Are You Square? a lesson plan for grades 9–12 comp</td>
<td>Very complete lesson plan using spreadsheet and calculator. In this lesson, students will find area and perimeter of polygons and circles. The students will also use technology to find maximum and minimum values for area and perimeter problems. Contains pre worksheet, worksheet and assessment test as well as links. Maps to NC standards.</td>
<td>64.6%</td>
</tr>
<tr>
<td>Best list of classroom management links</td>
<td>Many good links for classroom management.</td>
<td>51.1%</td>
</tr>
<tr>
<td>Running a Successful Technology Campaign: A Politi</td>
<td>Detailed article on getting funding for technology. Links</td>
<td>46.0%</td>
</tr>
</tbody>
</table>

[Image of the text-based search results screen]
Figure C8. Second search results. Upper two documents saved from first search.
Figure C9. System automatically enters document numbers and names into questionnaire.
Figure C10. **Reviewing first document. Click opens the document in second window.**

User evaluates, then closes. Note that directly under cursor, the task selected is repeated for the convenience of respondent.
Figure C11. After closing document, respondent enters evaluation of document. If
document was not useful or did not leave respondent better off, the next document is
reviewed. If the document was useful, the utility of the document is quantified.
Figure C12. **Quantifying the utility of document.**

- **Question 1:**
  - How much time would you be willing to spend searching in order to find this document?
  - Options: 1. Over an hour, 2. 31-60 minutes, 3. 16-30 minutes, 4. 11-15 minutes, 5. 5-10 minutes, 6. 1-5 minutes, 7. about a minute or less.

- **Question 2:**
  - How much time do you think finding this document has saved you toward completing your task?
  - Options: 1. More than two hours, 2. 1-2 hours, 3. 30 minutes - an hour, 4. 15-30 minutes, 5. 5-15 minutes, 6. up to 5 minutes, 7. a minute or two.

- **Question 3:**
  - Do you think you have enough information to complete your task now? If you were at work, trying to complete the task, would you stop looking now, or would you continue to look at information?
  - Options: Stop, Continue.

*Note: Even if you would have stopped, we ask that you continue for the benefit of the research.*
Figure C13. Summary. After all documents are reviewed, respondent completes post information need test.

Summary Questions

THAT'S IT! You have evaluated all of the documents you selected.

Thank you for your work so far. We have a few final questions for this search.

1. Now that you have searched and looked at all the documents, think about gathering materials, just considering your information needs, how ready are you to complete your task? (Make a web site for yourself)?

   1. I have no information need. I'm ready to complete my task right now.
   2. 
   3. 

2. I have some need. I have a general idea, but I need to fill in gaps in my knowledge or get some supporting information.

   4. I have some need. I have a general idea, but I need to fill in gaps in my knowledge or get some supporting information.

3. Large information need. I have little knowledge of this and very little, or no supporting information.

4. Give a rough estimate or how much more time you think you would need before you had all the information you needed to complete the task.

   [ ] hours

The next two questions look at your overall satisfaction with this search/retrieval system.
Next respondent evaluates satisfaction and adds comments.

3. How satisfied are you with the results of the system
   - 1. I am very satisfied with the results of the system. Overall, I got great information.
   - 2
   - 3
   - 4. The results I got from the system were OK for my needs.
   - 5
   - 6
   - 7. I'm not at all satisfied with the results of this system. Overall, I did not get good results.

4. How much did you like using this system?
   - 1. Overall, I really enjoyed using the system.
   - 2
   - 3
   - 4. The system was fine.
   - 5
   - 6
   - 7. I did not enjoy using this system.

5. What in particular did you like or not like about the system?
   - Liked the colors.
Figure C15. **Respondent searches twice using the first system, then is introduced to second system. Tag system instructions shown here.**
Figure C16. Second page of tag instructions.

About the text box

The text box searches ONLY the Name, Author, and Description fields. It does NOT search the document like the text searches in common web browsers. For example, if you want lesson plans about stars, you would check the lesson plan box and type in stars. This would return just the documents that have been tagged as

If you want to search with multiple terms (like star, black hole, telescope), just separate the terms with a comma. Notice that black hole is two words, but one term... searching for black hole is different than searching for black hole (which could return a document about digging post holes). The search tool looks at the text terms as OR searches. For example, if you check lesson plan and enter stars, black hole, telescope, all the objects must be tagged as lesson plans AND have stars OR black hole OR telescope in the title, author or description.

This is a different way of searching. If you get confused, please call Steve (800-314-1493) or email (sschatz@indiana.edu). If we walk through it once, you will get the hang of it.

Now, try a test search.

Test Search

Suppose you want to find a document that tells 4-5 graders how to make hand shadows.

1. Open the search tool.
2. The rule doesn’t really matter, so leave it blank.
3. The content area doesn’t fit in any of the three areas, so leave it blank.
4. Grade level - that makes a difference. If you leave this blank, it will return things tagged for all grades, and you want for 4-5 graders. So, check 4-5
5. What kind of information do you need? It’s an activity or how-to... so check both.
6. It may be necessary to limit to presentation/explanation. Try it with and without.
7. You may want to only have high quality. However, if nothing returns, don’t select a quality or completeness. If too many return, you can select a quality or completeness to cut down the number of documents.
8. In the text field, enter hand shadow... which will search title, description, and author.

Now... try it yourself. See what happens when

you select different tags. Remember, no tags will return everything. Every check is a filter to reduce and focus the documents returned.

Click this button to open the search tool you will be using.

Begin Test Search

Leave this window open.
Figure C17. Third page of tag instructions.
Figure C18. Selecting task for tag-based searches. No difference from first system selection.
Figure C19. Tag based search tool.
Figure C20. Tag-based search results. Note that these have quality and completeness ratings not included on text system and do not show percentage of match.

Description and title are the same.

Evaluation of documents and summary questions are the same with both systems.
Appendix E – Tagging Tool