



## COVER SHEET

---

**Greisdorf, Howard and Spink, Amanda (2001) Median measure: an approach to IR systems evaluation. *Information Processing and Management* 37(6): pp. 843-857.**

**Copyright 2001 Elsevier.**

Accessed from: <http://eprints.qut.edu.au/archive/00004744>

## MEDIAN MEASURE: AN APPROACH TO IR SYSTEMS EVALUATION

Howard Greisdorf  
School of Library and Information Science  
University of North Texas  
P. O. Box 311068 Denton, Texas 76203-1068  
*E-mail: hfg0001@aol.com*

Amanda Spink\*  
School of Information Sciences and Technology  
The Pennsylvania State University  
511 Rider I Building, 120 S. Burrowes St  
University Park, PA 16801  
Tel: (814) 865-4454 Fax: (814) 865-5604  
*E-mail: spink@ist.psu.edu*

\* To whom all correspondence should be addressed.

## ABSTRACT

In this paper we report results from three studies examining 1295 relevance judgments by 36 IR system end-users. We examined both the region of the relevance judgment, from non-relevant to highly relevant, and motivations or levels of their relevance judgments. Our study has three major findings. First, the frequency distributions of relevance judgments by IR system end-users tend to take on a bi-modal shape with peaks at the extremes (non relevant/relevant) with a flatter middle range. Second, the different type of scale (interval or ordinal) used in each study did not alter the shape of the relevance frequency distributions. And third, on an interval scale, the median point of relevance judgment distributions correlates with the point where relevant and partially relevant items begin to be retrieved. The median point of a relevance judgment distribution may provide a measure of user/IR system interaction to supplement precision/recall measures. The implications of our investigation for relevance theory and IR systems evaluation are discussed.

## INTRODUCTION

Relevance researchers from an information science perspective have sought to identify the qualities, dimensions and attributes of a user possessing an information problem with the qualities, dimensions and attributes of information retrieval (IR) systems, with the potential to help resolve the user's information problem. Current generalization about user interactions with IR systems conceptualize human beings approaching their information problems with an assortment of cognitive contexts, situational perspectives and task orientations that influence their information seeking, searching, retrieving and evaluating behavior. This approach to both IR processes (end-user information seeking and searching behavior) and evaluation (end-user measures of relevance) has led to an abundance of relevance frameworks. Yet there remains no overriding consensus, from a theoretical viewpoint, on the nature of relevance judgments during human interactions with IR systems.

The importance of relevance to IR systems evaluation has been the focus of much attention (Rees, 1966; Tague & Schultz, 1989; Saracevic, 1995; Borland & Ingwersen, 1997; Harter & Hert, 1998). IR system evaluation studies largely use dichotomous relevance judgments and variations of precision and recall measures. The Text Retrieval Evaluation Conferences (TREC) use various precision and recall measures to compare IR system performance (Sparck Jones, 1995, 1999). The limitations of precision and recall have led to calls for the development of new IR evaluation measures. Saracevic (1995) concluded that evaluation was still an integral part of IR, and suggested, "the issue and challenge for any and all IR evaluations are the broadening of approaches and getting out of the isolation and blind spots of single level, narrow evaluations. How can interaction be ignored in IR evaluation at any level?" However, evaluation behaviors by IR system users are still both challenging and elusive.

User-centered relevance studies have led to a better understanding of the user/IR system interaction process. However, few studies have identified relevance effects within a theoretical framework that encompasses pragmatic implications for IR system evaluation. Other than calls for abandoning precision and recall as measures of user/IR system interaction effectiveness, research has not generally looked at relevance judgment data as a whole (Spink & Wilson, 1999). Both parametric and non-parametric statistical approaches to studying relevance judgments have segregated key dependent variables rather than examining the results as an aggregated whole.

Our approach to this research builds on the findings from previous research, particularly by Spink and Greisdorf (1999, in press) and Spink, Greisdorf and Bateman (1998). In this paper we report results from three studies examining the aggregate regions (from non-relevant to highly relevant) and levels or reasons for relevance judgments by 36 IR system end-users.

## RELATED STUDIES

Information scientists have approached human relevance judgments from a variety of perspectives with limited success in unifying the theoretical concepts associated with relevance, including users' cognitive attributes (Ingwersen, 1996). Caudra and Katter (1967) identify relevance as a "black box" and Janes (1994) perpetuates that notion with the concept of relevance as a "big black question mark".

In the area of IR relevance studies, most of the focus has centered on the identification of user attributes and criteria for evaluating items retrieved from an IR system (e.g. Bateman, 1998). User's ability to make relevance judgments has been approached from a variety of directions with limited success in unifying the theoretical concepts associated with relevance, including cognitive modeling (Daniels, 1986; Belkin, 1990; Ellis, 1992; Harter, 1992; Ingwersen, 1996) satisfaction (Gluck, 1996; Thong & Yap, 1996), value (Su, 1998), task (Allen, 1996; Belkin et. al., 1990), utility (Bates, 1996), pertinence (Kemp, 1974; Howard, 1994), and situation (Wilson, 1973).

### Partial Relevance Studies

Spink and Greisdorf (1999) found the middle range of a relevance judgment frequency distribution plays an important role for users in their early stages of seeking information on a particular problem, and in creating changes in the user's information problem or question during the information seeking process. The middle range of relevance judgments was extended by Spink, Greisdorf and Bateman (1998) with a three-dimensional spatial model of relevance that defines the manifestations of this middle region as consisting of partially relevant and partially not relevant judgments. Spink, Bateman and Greisdorf's (1999) successive searching behavior study found that search episodes early in the information seeking process contribute more partially relevant judgments than in later searches on the same information problem. Further work investigating the middle regions of partial relevance identified a taxonomy of end-user descriptions of partially relevant and partially not relevant judgments that identifies the middle regions of relevance as a dimension that consists of combinations of both positive and negative levels of relevance (Greisdorf & Spink, 1999; Spink & Greisdorf, in press), and suggests that the relevance judgment frequency distribution requires more extensive investigation.

### Relevance Judgment Frequency Distributions

Several studies provided data that yielded distributions of IR system users' relevance judgments on an interval scale that displayed bi-modal characteristics. Work by Rees and Schultz (1967), Saracevic (1969), Janes (1991, 1993), and Janes and McKinney (1992) found bi-model distributions with a high number of relevance judgments at each extreme (not relevant/relevant) and a scattering of relevance judgments in lesser numbers in the middle of the distribution. Janes (1993) discussed the characteristics of relevance judgment distributions and concluded it might be a statistical artifact. However, recent studies on the role of partial relevance in IR interaction (Spink, Greisdorf & Bateman, 1998) lead to our further research to clarify the distribution of relevance judgments from non-relevant to highly relevant.

This paper reports results from a study investigating end-users' relevance judgments to explore new directions for IR system evaluation. First, we examine the distribution frequencies of relevance judgments by IR system end-users. Second, we used different scales (interval and ordinal) and examined if the type of scale affects the shape of the relevance judgment frequency distributions. Third, we statistically examined the relevance judgment distribution to derive an evaluation measure for user/IR system interaction that offers a supplement to precision/recall approaches.

## RESEARCH QUESTIONS

In this paper we explored the following research questions:

- (1) What is the nature of relevance judgments frequency distributions?
- (2) Is there any statistical characteristic of relevance frequency distributions that could contribute to the evaluation of IR systems?

## RESEARCH DESIGN

### Data Collection

We collected relevance judgment data during three studies conducted at the University of North Texas from 1998 and 1999, involving 36 end-users and 1295 relevance judgments.

The basic data from each study is shown in Table 1.

Table 1: Summary of Basic Data.

Study	No. of	No. of	No. of	Interval	Ordinal	Database
-------	--------	--------	--------	----------	---------	----------

	End-users	Searches	Items Judged	Scale Used	Scale Used	Resource
1	13	28	655	3-inch line	4 categories	Dialog
2	8	14	370	100mm line	4 categories	Dialog
3	15	15	270	Percentage	4 categories	Inquirus
Total	36	57	1295			

The purpose of three separate studies was to determine if variables such as type of relevance judgment scale, type of population sample, and choice of databases influence the nature of relevance frequency judgment distributions.

### Study 1

The data analyzed in this study was collected from 13 end-user graduate students at the University of North Texas who conducted 28 searches on their own information problems using the DIALOG database. A total of 655 items were retrieved and judged on two scales of relevance judgments. We collected relevance judgments using a relevance judgment worksheet.

### Worksheet

Using a worksheet, end-users indicated for each item retrieved “how” they judged that retrieved item in terms of its relevance to their current information needs within four measurements of their relevance judgments.

[INSERT COPY OF WORKSHEET]

End-users were not restricted in their relevance judgments to only full text or bibliographic records. The worksheet was pre-tested with a small group of end-users before use in the larger group.

The *first measure* on the worksheet was a 3-inch line ranging from not relevant (NR), indicated at the extreme left, to highly relevant (R) indicated at the extreme right. The only instruction given to the end-users for completing this first measure was a request to provide a mark on the line that represented the retrieved items relevance to their current information problem in the range provided (not-relevant to highly relevant). End-users marked the line at a point that they felt represented how relevant the retrieved item was to their current information need. This represented an interval measure of their relevance judgments.

The *second measure* was categorical. Four boxes were provided on the worksheet for end-users to make their judgments for each item retrieved, as either - not relevant, partially not relevant, partially relevant, or relevant. Partially relevant represented a judgment that confirmed some relation by inference existed, but the relationship may be weaker than a relevant relation at the time the judgment was made; and partially not relevant represented that some non-relation existed, but the inference may not be strong enough to totally reject the relation as not relevant at the time the judgment was made.

These definitions are only included to provide the subject with a distinction between positive and negative partially relevant. End-users were not given any definitions for regions of relevance involving this second measure. They were only provided with the four categories within which to make their relevance judgments. This approach was to determine if end-users could make such distinctions and how those distinctions may relate to measures one, three and four.

A *third measure* was used to assist in the identification of “how” the retrieved items were judged based on levels of relevance previously identified and defined by prior research and summarized by Saracevic (1996).

Findings from previous studies by Spink, Greisdorf and Bateman (1998) show that partial relevance plays an important role in defining the nature of retrieved items. The “partial” nature of a relevance judgment impacts the precision ratio used to measure IR system effectiveness. To

assess the impact, end-users were asked to identify the levels of relevance that contributed to their interval measure on the 3-inch line and the ordinal measures of relevant, partially relevant, partially not relevant and not relevant. These levels of relevance were defined as follows:

Systematic Level:	S =	The item retrieved was in a form/format that meets my information need;
	NS =	The item retrieved was NOT in a form/format that meets my information need;
Topical Level:	T =	The item retrieved was on the topic/subject requested;
	NT =	The item retrieved was NOT on the topic/subject requested;
Pertinence Level:	P =	The item retrieved is/will be informative;
	NP =	The item retrieved is NOT/will NOT be informative;
Utility Level:	U =	The item retrieved is/will be useful in resolving my current/future information need;
	NU =	The item retrieved is NOT/will NOT be useful in resolving my current/future information need;
Motivational Level:	M =	The item retrieved will/may cause me to take other action(s) now that I have this information;
	NM =	The item retrieved will/may NOT cause me to take other action(s) now that I have this information.

These systematic, topical, pertinent, utility and motivational levels were used to identify how end-users make their relevance judgments.

A *fourth measure* enabled the end-users to “briefly describe” in their own words “why” they judged the retrieved items as they did.

### Study 2

A second group of 8 end-users conducted 14 DIALOG searches on their own information problems and retrieved 370 items for judgment along two scales identified in Study 1 with one exception - the continuous interval measure in Study 2 was represented by a 100mm line. A different line length was used on the second study to examine how the distribution of users' relevance judgments may vary over different scales.

### Study 3

A third group of 15 end-users searched the Inquirus proprietary Web search engine (Lawrence & Giles, 1998) on their own information problem. A total of 270 items were retrieved and evaluated using both a continuous interval measure and a categorical ordinal measure. The interval measure used in this study was modified from the worksheet used in Study 1 and 2 so that each user could rank each retrieved item as a percentage (0% to 100%) of how relevant the item was to their information problem. The second measure was the same categorical assessment used in Studies 1 and 2.

## RESULTS

This paper extends findings reported in Greisdorf and Spink (2000). The analyzed data provided several interesting findings about how users evaluate items retrieved from an IR system.

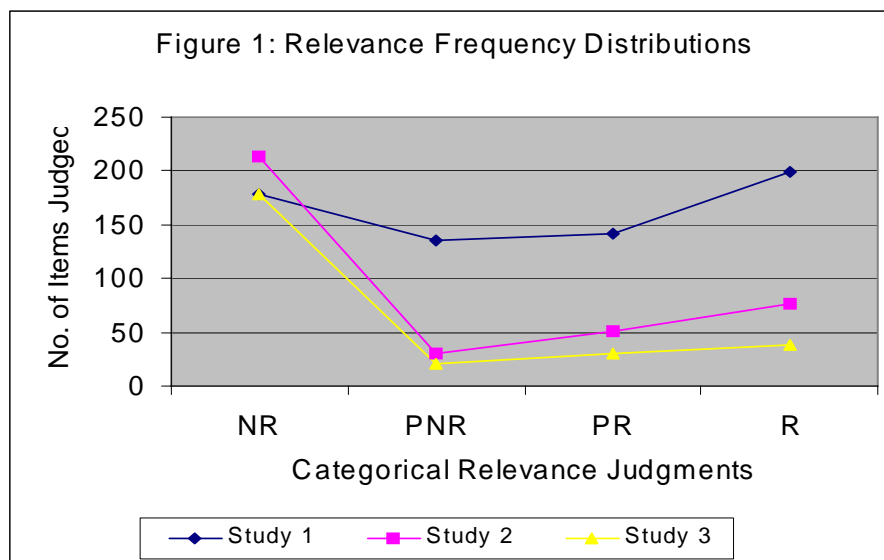
### End-Users Judgments - Ordinal Categorical Scale

In each of the three studies, end-users were presented with an ordinal relevance measure consisting of four categories: not relevant, partially not relevant, partially relevant and relevant.

The relevance judgments for retrieved items on an ordinal scale are shown in Table 1 and Figure 1 below.

**Table 1.** Summary of relevance judgments in the three studies.

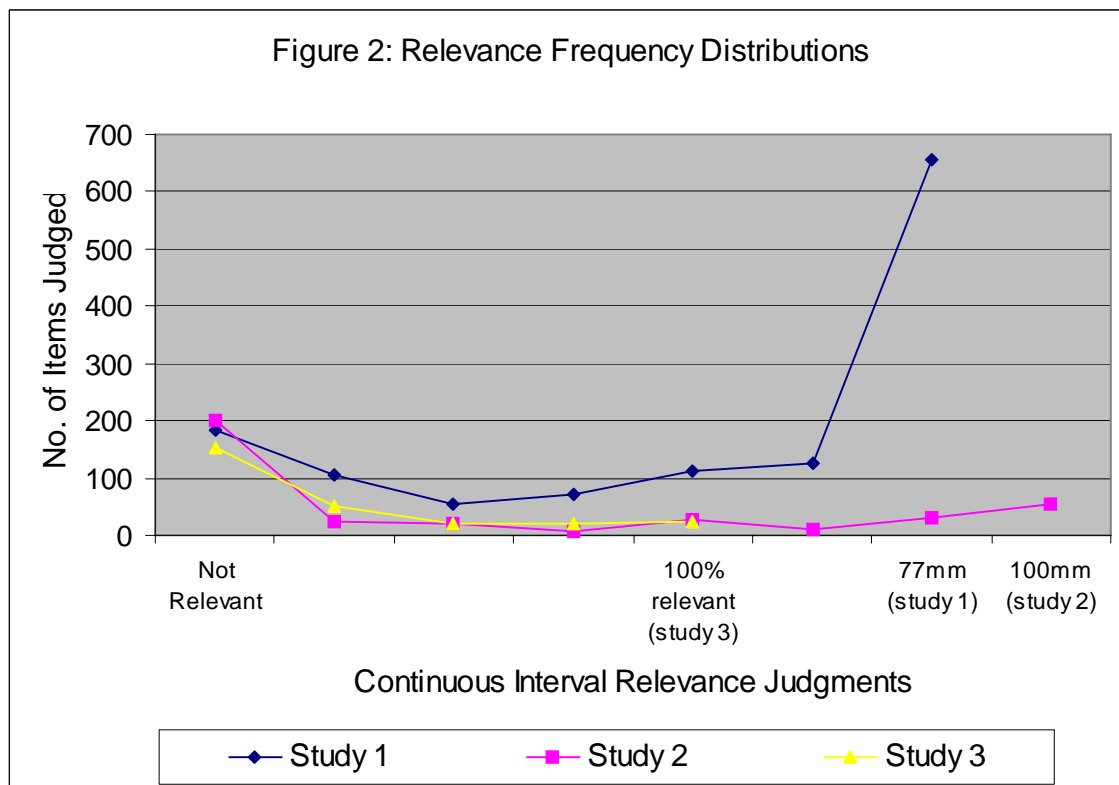
Study Relevance Region	1	2	3	Total
Not Relevant (NR)	178	213	179	570
Partially Not Relevant (PNR)	136	30	21	187
Partially Relevant (PR)	142	51	31	224
Relevant (R)	199	76	39	314
Total	655	370	270	1295



### End-Users Judgments - Interval Scale

In Studies 1 and 2, the 3-inch line and 100mm lines established baseline criteria for judgment at the extremes (not relevant/highly relevant) with end-users marking anywhere on the line to establish the strength of their relevance judgment for each item evaluated. Although this procedure established baselines at the extremes (0% for not relevant and 100% for highly relevant), users had the ability to indicate their own measure of relevance within the limits of the scale.

We next plotted the frequency distributions of the relevance judgments made on the interval scale in each study in Figure 2.



In Study 3 end-users were asked to rank their relevance judgments based on a percentage. Although this procedure also established baselines at the extremes (0% for not relevant and 100% for highly relevant) users had the ability to indicate their own measure of relevance within the limits of the scale. Totals indicating which areas of these various scales were used as a measure of user indications of the strength of their relevance judgments are shown in Tables 2, 3 and 4.

Table 2. End-user relevance judgments on an imposed 3-inch interval scale (77 mm)

Range of marks on scale	Total Marks
0 mm - 10mm	182
11mm - 24mm	105
25mm - 39mm	56
40mm - 53mm	72
54mm - 66mm	113
67mm - 77mm	127
<b>Total</b>	<b>655</b>

Table 3. End-user relevance judgments on an imposed 100mm interval scale

Range of marks on scale	Total Marks
0mm - 4mm	200
5mm - 20mm	24
21mm - 35mm	19

36mm - 50mm	6
51mm - 65mm	27
66mm - 80mm	11
81mm - 95mm	29
96mm - 100mm	54
Total	370

Table 4. End-user relevance judgments on a user-established percentage ranking scale

Percentage Rankings Used	Total
0, 1, 2, 5	152
10,15,20,25	51
30,35,40,50	22
55,60,70,75	20
80,90,99,100	25
Total	270

#### Measures of Central Tendency in Relevance Frequency Distributions

Considering that the shape of each of the relevance judgment frequencies was represented by some form of bi-modal distribution, calculations of a weighted mean, median and skew identified similarities and differences. The results are shown in Table 5.

Table 5. Characteristics of the relevance judgment distributions

Study #	No. of Intervals on the Scale	No. of Items Judged	Weighted Mean of the Distribution occurs at:	Median of the Distribution occurs at:	Skewness of the Distribution (Median minus Mean)
1	78	655	37.5mm	38mm	0.5
2	101	370	31.5mm	4mm	-27.5
3	101	270	20%	5%	-15

#### Positive and Negative Aspects of Interval and Ordinal Distributions

Prior research has identified user judgments of partially not relevant, partially relevant, and relevant items include some positive aspects of the retrieved item (Spink & Greisdorf, in press). As partially not relevant, partially relevant and relevant items included some positive aspect of the retrieved item, categorical data based on mean performance was calculated in order to compare the results with the measures of central tendency calculated for the interval data frequency distributions. This comparison is shown in Table 6.

Table 6. Interaction effectiveness within and between groups.

Study Number/	1	2	3	Total
Total items judged	655	370	270	1295
No. of end-users	13	8	15	36
No. of not relevant items (NR)	178	213	179	570
Mean (NR) items per end-user	13.69	26.63	11.93	
No. of partially not relevant (PNR) items	136	30	21	187

Mean (PNR) items per end-user	10.46	3.75	1.40	
No. of partially relevant (PR) items	142	51	31	224
Mean (PR) items per end-user	10.92	6.38	2.07	
No. of relevant (R) items	199	76	39	314
Mean (R) items per end-user	15.31	9.50	2.60	
Mean net effectiveness (R+PR+PNR)-NR	23.00	-7.00	-5.87	
Precision expressed as: (R+PR+PNR)/Total items judged	73%	42%	34%	
Median mark or rank on the interval scale (Table 2)	38mm	4mm	5%	
Weighted mean (Table 2)	37.5	31.5	20	
Skewness of distribution (Table 2)	0.5	-27.5	-15	

Linear correlation between the average net effectiveness measures of each group (taken as the average relevant, partially relevant and partially not relevant judgments minus the average not relevant judgments), the measures of central tendency on the interval scale distributions, and the precision ratios yielded the results shown in Table 7.

Table 7. Correlation related to average net effectiveness.

Variable	P-Value	Statistically Significant ( $P < 0.05$ )
Weighted Mean of the distribution	0.469	No
Median of the distribution	0.005	Yes
Skew of the distribution	0.273	No
Precision	0.146	No

The results provided several interesting findings about how users evaluate items retrieved from an IR system. While individually different in their information problems and their approach to evaluating retrieved information, some common characteristics emerged across these three studies.

## DISCUSSION

Our research shows three major findings related to relevance judgment distribution.

### Relevance Judgment Frequency Distribution

First, relevance judgment frequency distributions continue to take on the shape of a bi-modal frequency distribution (Rees & Schultz, 1967; Saracevic, 1969; Janes, 1991, 1993; Spink & Greisdorf, 1999; in press). A high left peak at the not relevant end, a flat distribution typically identifies these distributions with smaller peaks in the middle range, and a right peak at the highly relevant end. Although this bi-modal shape was considered a statistical artifact by Janes (1993), continuing research including scales using more than binary assessments of relevance appear to take on this characteristic shape (Greisdorf & Spink, 1999). This study leads to confirmation that

the nature of the scale (3-inch line, 100mm line, percentage ranking, and categorical scaling with 4 regions of relevance) does not alter the nature of the distribution.

#### Relevance Judgment Scales

Various scales were used to identify the degree of relevance represented by items retrieved from an IR system. While many have sought to identify just the right scalar measurement that depicts the full range of individual relevance judgments (Katter, 1968; Eisenberg, 1988; Janes, 1993; Tang, Vevea & Shaw, 1999), it is becoming evident from the research that end-users are capable of using a variety of scales. Any scale that implies some range of utility, value, strength, importance or magnitude appear to suit the end-user in making a relevance assessment of items retrieved from an IR system. Both interval (continuous) scales, as well as ordinal (categorical) scales appear to be acceptable and functional for end-users. Although exceptions may occur at individual levels in terms of scale preference (Tang, Vevea & Shaw, 1999), the results of overall distributions of relevance judgments by end-users indicate that scale choice may not be a key issue in relevance assessments. The choice has generally been a function of the operationalized approach to the study by the researcher. In this study the four scales used, including three interval scales and one categorical scale, all provided frequency distributions that conformed to the same bi-modal shape as indicated in Figures 1 and 2.

#### Median Measure

The results indicated no correlation of measures of central tendency with the precision ratios and a significant positive correlation between the measure of average net effectiveness and the median of the distribution. Recognizing that the median by itself is not comparable across scales of differing total intervals, a conversion is necessary for comparison purposes. That conversion represents the median measure described by the relevance frequency distribution. It provides an approximation of how much more effective one group of searches is from another without actually separating relevant items from not relevant items. As a measure, the median effect is expressed as follows:

$$\text{Median Measure} = \text{Median point on the relevance scale} / \text{No. of points on the scale}$$

Applying that formula to the data in Table 4 yields the following median measure that can be used in conjunction with other IR evaluation measures to compare the effectiveness of IR systems:

$$\text{Study 1 Median Measure} = 38/78 = .487 \quad (\text{Precision} = 73\%)$$

$$\text{Study 2 Median Measure} = 4/101 = .040 \quad (\text{Precision} = 43\%)$$

$$\text{Study 3 Median Measure} = 5/101 = .050 \quad (\text{Precision} = 32\%)$$

At first glance, there appears to be no parsimony in evidence by comparing the median measure to the precision associated with each study. However, considering that the approach to this discussion was from the point of view that relevance judgment frequency distributions exhibit certain characteristics in common with each other it is necessary to complete the analysis by applying a sign (+ or -) to the median effect formula based on the nature of the distribution. The appropriate sign is a function of the skew of the distribution (Median – Mean) provided in Table 4 and when applied to the calculations above yield the following results:

$$\text{Study 1 Median Measure} = 38/78 = +.487 \quad (\text{Precision} = 73\%)$$

$$\text{Study 2 Median Measure} = 4/101 = -.040 \quad (\text{Precision} = 43\%)$$

$$\text{Study 3 Median Measure} = 5/101 = -.050 \quad (\text{Precision} = 32\%)$$

The median measure ranges from –1 to +1 and infers not only how positive were the search results, but how positive they were in relation to all the negative results obtained from the search. The following inferences concerning the median measure could be made:

- The group of end-users in Study 1, as a result of their interaction with an IR system, were able to retrieve and judge more documents with positive aspects surrounding their search for information than negative aspects (identified by the

positive sign attached to the median effect statistic of +0.487).

- The end-users in Studies 2 and 3, however, judged more documents with negative aspects than positive ones as evidenced by median effects of  $-0.040$  and  $-0.050$  respectively. Comparing the median effect of Study 2 with the median effect of Study 3 in terms of the positive versus negative aspects of the retrieved items, the effectiveness of the Study 2 group was greater than that of the Study 3 group. Thus the median measure, without separating the categorical judgments to obtain a precision ratio, could not only approximate how much more effective one search is from another, but provide an indication of whether it was a net positive search interaction or a net negative search interaction.

An overall measure of positive versus negative effectiveness can be achieved using the median measure. However, the numerical values associated with the statistic cannot be used as arithmetic operators on the underlying distribution frequencies since both number of end-users and number of items judged vary in each distribution. The value of the median measure is the provision of an inference of how many more relevant, partially relevant and partially not relevant items have been retrieved and judged in relation to the not relevant items that precision measures do not incorporate.

Therefore, the greater the median measure the greater the IR interaction effectiveness. Unlike the precision measure, the total number of items judged by users does not encumber the median measure. It is more an indicator of how many more positive items were judged than negative ones. A search that retrieves 10 not relevant items and 5 relevant items has a precision of 33% with a search net effectiveness measure of  $-5$  (5 relevant minus 10 not relevant). A search yielding 100 not relevant items and 50 relevant items also has a precision of 33%, yet its search net effectiveness measure is  $-50$ . Is the second search 10 times worse because it yielded 10 times the not relevant items as the first search, or is the second search 10 times better because it yielded 10 times more relevant items than the first? The second search is not the same as the first that precision measures tend to imply. This becomes of even greater importance when items retrieved from an IR system are considered to be only partially relevant or partially not relevant (Spink & Greisdorf, 1999). In those instances the positive and negative aspects of those judgments are not made apparent by the end-users making those relevance decisions or by the precision measures currently in use. Some other statistic needs to be part of the interaction evaluation process. The median measure could represent such a statistic, yet bears further exploration under a variety of conditions.

## CONCLUSIONS AND FURTHER RESEARCH

This investigation provides an avenue for the development of further research evaluating IR systems and with a larger data set of end-users to provide more support for the findings included in this investigation. When end-users query an IR system to help resolve an information need, the evaluation of retrieved items encompasses a variety of individual behaviors, both implicit and explicit, that confound the identification of effects that define relevance evaluation behavior. The findings in this investigation have identified three such effects for further study:

1. The apparent ability of users to make relevance judgments on any type of interval or ordinal scale; and
2. A measure of central tendency from aggregated relevance data that measures search effectiveness (median measure).

Although this investigation provides an avenue for ongoing research, several aspects inherent in its operationalization limit it. First, ordinal and interval scales were both presented to the same end-users that could contribute to possible biased relevance judgments from one scale to the other. Second, no provision was made to account for a domain knowledge-moderating variable that could impact the relevance judging process in the users information problem area.

Third, a greater diversity of users and systems could provide more support for the findings included in this investigation.

In conclusion, research and theoretical perspectives surrounding relevance continue to point in the direction of a process approach. Attempts to uncover, describe or explain single unique variables that contribute to that process appear to confound advances more than they resolve issues that could take relevance theory to a higher plateau.

#### REFERENCES

- Allen, B. L. (1996). *Information tasks: Toward a user-centered approach to information systems*. New York, NY: Academic Press.
- Bateman, J. (1998). Changes in relevance criteria. *Proceedings of the American Society for Information Science, October 26-29, 1998, Pittsburgh, PA*.
- Bates, M. J. (1996). Document familiarity, relevance and Bradford's Law: The Getty Online Searching Project Report No. 5. *Information Processing & Management, 32*(6), 697-707.
- Belkin, N. J. (1990). The cognitive viewpoint in information science. *Journal of Information Science: Principles and Practice, 16*(1), 11-16.
- Belkin, N. J., Chang, S. & Downs, T. (1990). Taking account of user tasks, goals and behavior for the design of online public access catalogs. *Proceedings of the American Society for Information Science, 53<sup>rd</sup> Annual Meeting, vol. 27, November 4-8, Toronto, Canada*.
- Borland, P. & Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation, 53*(3), 225-250.
- Caudra, C. A. & Katter, R. V. (1967). Opening the black box of relevance. *Journal of Documentation, 23*(4), 291-303.
- Daniels, P. J. (1986). Cognitive models in information retrieval: An evaluative review. *Journal of Documentation, 42*(4), 272-304.
- Eisenberg, M. B. (1988). Measuring relevance judgments. *Information Processing & Management, 24*(4), 373-389.
- Ellis, D. (1992). The physical and cognitive paradigms in information retrieval research. *Journal of Documentation, 48*(1), 45-64.
- Gluck, M. (1996). Exploring the relationship between user satisfaction and relevance in information systems. *Information Processing & Management, 32*(1), 89-104.
- Greisdorf, H., & Spink, A. (2000). A new way to evaluate IR systems performance: Median measure. *Proceedings of NOM 2000: National Online Meeting, May 2000, New York* (pp. 137-144).
- Greisdorf, H. & Spink, A. (1999). Regions of relevance: approaches to measurement for enhanced precision. *IRSG 99, 21<sup>st</sup> Colloquium on Information Retrieval, April 19-20, Glasgow, Scotland* (pp. 1-33).
- Harter, S. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science, 43*(9), 602-615.
- Harter, S.P., & Hert, C.A. (1997). Evaluation of information retrieval systems: Approaches, issues and methods. *Annual Review of Information Science and Technology, 32*, 1-94.
- Howard, D. L. (1994). Pertinence as reflected in personal constructs. *Journal of the American Society for Information Science, 45*(3), 172-185.
- Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation, 52*(1), 3-50.
- Janes, J. W. (1991). The binary nature of continuous relevance judgments: A study of user's perceptions. *Journal of the American Society for Information Science, 42*(10), 754-756.
- Janes, J., & McKinney, R. (1992). Relevance judgments of actual users and secondary judges: A

- comparative study. *Library Quarterly*, 62(2), 150-168.
- Janes, J. W. (1993). On the distribution of relevance judgments. *Proceeding of the 56<sup>th</sup> ASIS Annual Meeting*, 30, October 24-28, Columbus, Ohio, 104-114.
- Janes, J. W. (1994). Other people's judgments: A comparison of users' and others' judgments of document relevance, topicality and utility. *Journal of the American Society for Information Science*, 45(3), 160-171.
- Katter, R. V. (1968). The influence of scale form on relevance judgments. *Information Storage & Retrieval*, 4, 1-11.
- Kemp, D. A. (1974). Relevance, pertinence and information system development. *Information Storage & Retrieval*, 10, 37-47.
- Lawrence, S. & Giles, C. L. (1998). Searching the World Wide Web. *Science*, 280(5360), 98-100.
- Rees, A. M. (1966). The relevance of relevance to the testing and evaluation of document retrieval systems. *ASLIB Proceedings*, 18(11), 316-324.
- Rees, A. M. & Schultz, D. G. (1967). *A field experiment approach to the study of relevance assessments in relation to document searching*. 2 vols. Cleveland, OH: Center for Documentation and Communication Research, School of Library Science, Case Western Reserve University.
- Saracevic, T. (1969). Comparative effects of titles, abstracts, and full texts on relevance judgments. *Proceedings of the American Society for Information Science*, 6, 293-299.
- Saracevic, T. (1995). Evaluation of evaluation in information retrieval. *Proceedings of the 18<sup>th</sup> ACM-SIGIR International Conference on Research and Development in Information Retrieval*, Association of Computing Machinery, Seattle, WA, 138-146.
- Sparck Jones, K. (1995). Reflections on TREC. *Information Processing and Management*, 31(3), 291- 314.
- Sparck Jones, K. (1999). Further reflections on TREC. *Information Processing and Management*, 36(1), 37-85.
- Spink, A., Bateman, J., & Greisdorf, H. (1999). Successive searching behavior during information seeking: an exploratory study. *Journal of Information Science*, 25(6), 439-449.
- Spink, A., & Greisdorf, H. (in press). Regions and levels: Mapping and measuring users' relevance judgments. *Journal of the American Society for Information Science*.
- Spink, A. & Greisdorf, H. (1999). How and why end-users make relevance judgments. *Proceedings of the 20<sup>th</sup> National Online Meeting, May 1999, New York*.
- Spink, A., & Greisdorf, H. (1997). Users' partial relevance judgments during online searching. *Online & CD-ROM Review*, 21(5), 271-280.
- Spink, A., Greisdorf, H. & Bateman, J. (1998). From highly relevant to not relevant: Examining different regions of relevance. *Information Processing & Management*, 34(5), 599-622.
- Spink, A., & Wilson, T. D. (1999). Toward a theoretical framework for information retrieval (IR) evaluation in an information seeking context. *Proceedings of MIRA 99: Evaluation Frameworks for Multimedia Information Retrieval Applications, Department of Computing Science, University of Glasgow - Scotland, April 14-16, 1999* (pp. 75-92). [<http://www.ewic.org.uk/ewic/workshop/view.cfm/MIRA-99>]
- Su, L. T. (1998). Value of search results as a whole as the best single measure of information retrieval performance. *Information Processing & Management*, 34(5), 557-579.
- Tague, J. & Schultz, R. (1989). Evaluation of the user interface in an information retrieval system: A model. *Information Processing & Management*, 25(4), 377-389.
- Tang, R., Vevea, J. L. & Shaw, W. M. (1999). Towards the identification of the optimal number

- of relevance categories. *Journal of the American Society for Information Science*, 50(3), 254-264.
- Thong, J. Y. L. & Yap, C. (1996). Information systems effectiveness: A user satisfaction approach. *Information Processing & Management*, 32(5), 601-610.
- Wilson, P. (1973). Situational relevance. *Information Storage & Retrieval*, 9, 457-471.

