Information Processing & Management Vol. 27, No. 6, pp. 629–646, 1991
Printed in Great Britain.

0306-4573/91 $3.00 + .00

RELEVANCE JUDGMENTS AND THE INCREMENTAL PRESENTATION OF DOCUMENT REPRESENTATIONS

JOSEPH W. JANES
School of Information & Library Studies, University of Michigan, 550 E. University, 304 W. Engineering, Ann Arbor, MI 48109-1092, U.S.A.

(*Received* 17 *December* 1990; *accepted in final form* 18 *April* 1991)

Abstract—A new approach to the solicitation and measurement of relevance judgments is presented, which attempts to resolve some of the difficulties inherent in the nature of relevance and human judgment, and which further seeks to examine how users' judgments of document representations change as more information about documents is revealed to them. Subjects (university faculty and doctoral students) viewed three incremental versions of documents, and recorded ratio-level relevance judgments for each version. These judgments were analyzed by a variety of methods, including graphical inspection and examination of the number and degree of changes of judgments as new information is seen. A post questionnaire was also administered to obtain subjects' perceptions of the process and the individual fields of information presented. A consistent pattern of perception and importance of these fields is seen: Abstracts are by far the most important field and have the greatest impact, followed by titles, bibliographic information, and indexing.

I. INTRODUCTION/STATEMENT OF THE PROBLEM

When approaching the study of relevance from an experimental perspective, a paradox quickly emerges. How can you introduce or manipulate any independent variables? Once users have seen and judged a document, they can never again give an unbiased, uncontaminated judgment. They will always be influenced by that initial viewing. It is then impossible to know how they would have judged the document if it had been presented later in the set or in a different format or on a different day.

One solution has been to give the documents to several judges under different conditions. This certainly makes a stab at studying topical similarity of documents to queries, but loses the unique character of real relevance judgments made by real users about their own documents and queries.

This study is an attempt to partially circumvent this problem and find out how users' relevance judgments change as they are provided with more information about documents. Looking at these changes (or lack thereof) may give us some insight into the nature of the relevance judging process.

We introduce this ability to observe changes in users' relevance judgments by presenting retrieved document representations to them *incrementally*, field by field. At each of three presentations, users view and judge a new version of the document, each time adding a field to the record. These judgments can then be analyzed to see what fields of a document representation are more likely to make users change their judgments or not, whether those judgments go up or down (increasing or decreasing relevance), and by how much.

These fine distinctions can be made by using a technique from psychophysics, specifically magnitude estimation, to elicit these relevance judgments. Previous work in using these techniques for this purpose has been fruitful and we have followed that line of methodology.

This study is an exploratory one. No hypotheses were proposed (at least not in the traditional statistical way), and no tests of statistical inference will be made. Rather, the data

This work was supported by a grant from the Office of the Vice President for Research, University of Michigan.

will be examined from a number of perspectives, some traditional, some novel, to try to see what patterns and questions emerge.

The research does arise from a theoretical base, however. Janes, in his dissertation [1], which applied concepts and theorems of search theory to the search for information, did propose that the process of relevance judging could be thought of, in a search theoretic way, as a detection process, "which assumes there are uncertainties inherent in detection and attempts to model them in a probabilistic way." (p. 178). One of the central notions of detection in search theory is that the more time searchers are given and the more information they have about the search, the target, and the environment, the more likely they are to make correct decisions about the nature of potential targets. This study tries to model that process by presenting documents to users incrementally and measuring the changes in judgment.

II. REVIEW OF THE LITERATURE

Relevance

It is not necessary to engage in a lengthy review of relevance-related research here. We do mention two significant general works, however, that are important to this research. Schamber et al. [2] have recently produced an excellent and insightful summary of work in this area, and the classic article by Saracevic [3] remains a touchstone for researchers years after its appearance. Both papers elaborate on the two major traditions in relevance work over the last three decades: relevance to a *subject* (topicality, system relevance, relevance at the source), and relevance to a *user* (utility, pertinence, satisfaction).

Schamber et al. propose a "dynamic, situational definition" of relevance based largely on the work of Dervin and Nilan, and "consider . . . alternatives that allow more room for consideration of internal values stemming from relevance judges themselves." In their conclusion, they present the following "conclusions about the nature of relevance and its role in information behavior":

1. Relevance is a multidimensional cognitive concept whose meaning is largely dependent on users' perceptions of information and of their own information need situations.
2. Relevance is a dynamic concept that depends on users' judgments of the quality of the relationship between information and information need at a certain point in time.
3. Relevance is a complex but systematic and measurable concept if it is approached conceptually and operationally from the user's perspective.

We, in the present study, concur with the above conclusions almost exclusively. This study examines relevance entirely from this user-centric perspective, and thus fits nicely with the alternative research perspective laid out above.

Saracevic [3] undertook to review and synthesize the work done to date on relevance. In so doing, he provided a "framework for thinking on the notion in information science" used to this day. In his Appendix, where he summarizes the results of experimentation, he makes several points salient to this work. Most importantly, under "Documents and Document Representations," he says that documents or representations are the "major variables" in relevance judgments, and that judgments based on titles or abstracts alone may be different from those based on full texts (points 1, 5, 6, p. 340).

Relevance judging

Research on relevance judgment and the measurement of relevance judgments per se has been relatively sparse. Two studies stand out: Eisenberg [4,5] and Eisenberg and Barry [6]. Eisenberg reports on the methodology he developed based on psychophysics research to measure relevance judgments using magnitude estimation techniques, in which subjects "freely estimate in numbers the intensity of a stimulus" (p. 374). Such techniques are rou-

tinely used to measure stimuli such as light, loudness of a sound, or the length of a line. Eisenberg was able to show that magnitude estimation relevance judgments were consistent and exhibited the same characteristics as those of other, physical stimuli, thus validating their use in information retrieval research. He presents, in an appendix, the instructions given to subjects in his study, and guidelines for further use of these techniques. These instructions and guidelines are incorporated, in large part, in the methodology of the present study, with some minor modifications. In particular, four guidelines are identified: (a) the use of a calibration exercise of judgments of lines, (b) randomly selecting four document/stimuli to act as "practice" . . . , (c) taking at least two judgments for each document/stimulus, and (d) adapting the instructions to meet the needs of specific experiments . . . (p. 387). Guideline (a) is specific to the use of number generation in that study; the use of line marking here was not considered to require similar calibration. Guideline (b) was followed: The first two document sets (involving six judgments) were not analyzed and considered to be practice stimuli. Guideline (c) was impractical for this study, given that sets were to be mailed, and subjects were thought unlikely to be willing to make repeated judgments. Guideline (d) was followed.

Eisenberg and Barry [6] examine further the effect on relevance judgment of the order in which documents are presented to a user. Following a similar methodology to the study described above, they found that "where documents are presented to judges in a high to low rank, they will consistently underestimate the significance of documents at the higher end. In a low to high situation, there is overestimation of documents, particularly at the low to middle range" (p. 296). This effect was more pronounced when subjects made judgments using a category (seven-point) scale than when they judged using magnitude estimation. They recommend, as a result, that further studies present documents to judges in random order. The present study follows this recommendation.

Incremental presentation/document components

A number of studies, from the 1960s and early 1970s, attempted to examine the differential impact or effectiveness of various components of documents on relevance judgment. Although their approaches and methods are quite different from those of the present study, some of their techniques and concepts are similar, and they merit note here.

Rath et al. [7] examined the capacity of subjects to identify documents that would answer a list of questions from viewing different parts of the documents. Subjects were given either titles, automatically generated abstracts, the first and last few sentences of the document (pseudo-abstracts), or the full texts. Subjects were volunteers; none were real users, and the experimental setting was highly artificial. They found no significant differences in performance, although subjects who saw titles were best able to determine correctly relevant (useful) documents. The title group also had a high rate of "Type II errors"—incorrectly accepting documents. The group seeing pseudo-abstracts performed the best at rejecting documents.

Resnick and Savage [8] examined, using an SDI paradigm, the consistency of subjects' relevance judgments. Forty-six technical professional people at IBM saw either an original IBM internal document, that document's citation, an abstract (including citation), or index terms (also including citation). They were asked to indicate which ones were relevant to their interests. One month later, the exercise was repeated, and only 10% of the judgments changed.

Thompson [9] examined the role abstracts play in quick screening of documents in a naturalistic setting. Twenty-two researchers at military laboratories recorded how they dealt with documents that came across their desks. On the form, they made an initial relevance judgment on a five-point scale; an interview was conducted later to get a second relevance judgment on some documents. Thompson found that the presence or absence of abstracts in these documents had no effect on the (self-reported) time subjects took to dispose of them, and had no effect on the quality of relevance judgments (measured by whether or not the judgment changed). There are some serious methodological flaws in this paper—there is a severe mortality effect in the subjects, and the operational definition of "good" judgments is questionable at best—but the idea is an interesting one.

The three studies in the literature that most closely resemble the present one are reported by Marcus et al., Hagerty, and Saracevic. Marcus et al. [10], in a report of experiments performed as part of Project Intrex, studied "indicativity," which they define as "[t]he ability of a catalog field to indicate the utility of a document to a searcher for a given problem . . ." (p. 21). The system developed in this project used over 50 fields to represent documents, and it is interesting to note that the fields most often requested and judged highest in utility are the fields used in the present study (pp. 18,19).

Subjects, who submitted actual queries, were presented first with titles, abstracts, index terms, and the index terms that matched those in the query for each of 20 randomly selected documents, and were asked for judgments of usefulness on a three-point scale (highly useful/somewhat useful/not useful). The full texts of the documents were then presented and similarly judged. Indicativity is "the fraction of evaluations made on the basis of the information in [a] field that were the same as those made on the basis of the full text . . ." (p. 21). Their findings were as follows: the title had an indicativity rating of 0.637, matching subjects had 0.672, all subjects had 0.704, and abstracts were the highest at 0.730. These findings suggested a "length hypothesis" to the authors—that longer fields were more indicative than shorter ones—which was subsequently verified by further analysis.

Hagerty [11] explored a similar, but inverse, question to that posed in the present study: "how many questions judged relevant to an article are also judged relevant using different length representations of the article . . .?" (p. 1). Her subjects judged document titles and abstracts of varying lengths (30, 60, and 300 words) against a list of questions generated from the documents. She found that as length of representation increased, so did recall and precision—an increase of about 50% for both measures for 300-word abstracts over titles. She also found a diminishing returns situation, interesting when compared to the Intrex study reported above: "[i]ncrease in recall and precision does not seem to be a function of length of representation" (p. 22).

Saracevic [12] also explored various types and lengths of document representations to determine their effect on judgment. Using 22 real users with 99 real queries, he gave them titles, abstracts, and full texts of documents and asked for relevance judgments on a three-point scale (relevant/partially relevant/not relevant). Users saw all titles first, then all abstracts, then all full texts. Saracevic found that different formats affected judgment, that "[i]t seems to be easier . . . to recognize non-relevance from the shorter formats than relevance . . ." (p. 297). He also found that 15% of judgments changed from title to full text, and that 10% changed from abstract to full text. Over all three formats, 22% changed at some point—a figure he calls significant.

In analyzing the data, he compared performance figures using "partially relevant" (P) judgments as is, and excluding them. Their exclusion raised performance, leading Saracevic to conclude that "Ps have a special, unstable property, wandering the most widely over all of the judgments." He proposes, as a result, that "relevant" and "non-relevant" "can be thought of as the opposite ends of a continuous relevance scale," and that the P documents "can be considered as having various degrees of relevance" (p. 298). All of this foreshadows Eisenberg's findings of 17 years later regarding the behavior of relevance as a psychophysical phenomenon, and leads directly to the aims of the present study.

III. METHODOLOGY

Our methodology largely follows that of Eisenberg, who introduced the techniques of magnitude estimation to the measurement of relevance. The work of Eisenberg and others [4,5,6,13] have shown that magnitude estimation can be used effectively and reliably in this realm, and has given guidelines for further use of these techniques. Some of their language has been adopted for the instruction of subjects, and, as closely as possible, a study of Eisenberg and Hu [13] on binary relevance judgments is being replicated here.

The chief advantage of using magnitude estimation techniques is that they provide the researcher with ratio-level data, which can be used in more sophisticated statistical techniques than ordinal-level data, obtained from categorical relevance judgments (relevant/partially relevant/not relevant, etc.). In addition, magnitude estimation ratings are

free from the contextual biases that category scales are subject to, and have been suggested as a potential measurement tool by the two early large-scale studies of relevance conducted by Rees and Schultz and by Cuadra and Katter [5, p. 374].

Subjects

Subjects were solicited by sending information packets in two groups. The first group included all faculty (including lecturers and people with courtesy appointments) in the School of Education and the Department of Psychology at the University of Michigan, and numbered approximately 215 people. The second group included all doctoral candidates in the Department of Psychology at Michigan, and numbered about 50 people. Faculty were asked to submit one search request, doctoral candidates were asked to submit no more than two.

The information packets included a cover letter describing the project and soliciting subjects' involvement, a consent form, and a search request form. As an incentive to participate, potential subjects were told that they would receive a free copy of the results of the search. The search request form asked subjects to give a brief narrative description of their topic, suggest terms or keywords to use during the search, provide known items or authors, and give specific limitations desired (language, format, year, etc.).

Forty individuals responded, with a total of 48 search requests. Three searches were later excluded from analysis: One subject reported familiarity with all the documents retrieved, and two searches were done in databases with no abstracts. Six sets were not returned or went astray in the mails, so the final number of document sets used was 39.

Design

Subjects were randomly assigned to one of four groups, which determined what fields of the documents they would see, and in what order they would see them. The four orders were:

TAB (Title/Abstract/Bibliographic) $n = 9$
TAI (Title/Abstract/Indexing) $n = 12$
TBA (Title/Bibliographic/Abstract) $n = 10$
TIA (Title/Indexing/Abstract) $n = 8$

where "Bibliographic" refers to the author's name, source (i.e., journal) name, and publication date, and "Indexing" refers to descriptors, identifiers, or other indexing information specific to a given database. Each subject, therefore, saw the title of each document first, the abstract either second or third, and either the bibliographic or indexing information in the other position.

Experimental procedure

The search. When a search request was received, it was assigned to a graduate student with online searching experience. The searcher examined the search request, undertook some preliminary research and analysis, and attempted to identify useful search terms and strategies. Then the searcher telephoned the subject for an interview to refine the searcher's understanding of the request. After the interview, the searcher went online and performed the search. Since the search requests came from faculty and students in education and psychology, the majority of searches were done in the *ERIC* and *PSYCInfo* databases. However, due to the wide variety of search topics (see Appendix 1 for a summary of topics), searches were also conducted in such databases as *ABI/INFORM*, *BIOSIS*, *Sociological Abstracts*, *ENVIROLINE*, *Child Abuse & Neglect*, *Family Resources*, and *PAIS*.

When the searcher felt he or she had retrieved a useful set, that set was typed out and downloaded in several formats. Up to 50 records of that final set were typed in full format. This set was the reward to subjects for participating in the study, but would not be judged. Three other sets were produced, to be used in the package that subjects would judge.

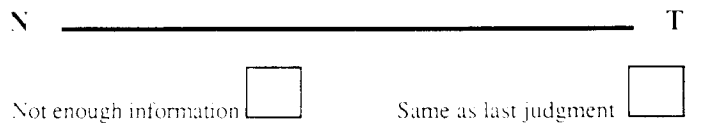

Fig. 1. Judging line.

As an example, if the subject had been assigned to the TAB (Title/Abstract/Bibliographic) group, the three sets produced would have been as follows:

- A set of titles of the first 22 documents.
- A set of titles and abstracts of the first 22 documents.
- A set of titles, abstracts, and bibliographic information of the first 22 documents.

These sets were downloaded to a floppy disk, then laser-printed. The reward set was printed on regular paper; the judging sets were printed on paper with the content of Fig. 1 at the bottom. This line was 100 millimeters long, and was used for subjects to record their judgments of the relevance of document representations to their query.

Experimental packet. The packet sent to subjects after the completion of the search contained the following:

1. A cover letter describing the contents of the package.
2. A sheet giving instructions on how to judge the documents (see Appendix 2), which was attached to a copy of the original search request.
3. An envelope containing the experimental set (3 versions of up to 22 document representations) and a questionnaire (Appendix 3).
4. A sealed envelope containing the 50-item reward set, which the subject was instructed to keep.

If fewer than 22 documents were retrieved, the judging and reward sets would include all documents. This packet was mailed to the subject via campus mail. A return label was attached to the envelope in (3) above, so that subjects could more easily return the experimental set.

Judging instructions. Subjects were instructed to "make a mark on a line corresponding to your impression of the degree of relevance of that document to your query, from none (N) to total (T)." This methodology is after Eisenberg [5]. Subjects were further instructed, "[i]f you do not feel you have enough information to make a decision, or if your judgment is the same as for the previous version of the document, check the appropriate box beneath the line."

No attempt was made to define or describe "relevance." Rather, the primitive notion of Saracevic [3] is relied upon, as in Eisenberg.

Experimental set. The document representations were presented to the subject in random order (to reduce the effects of order of presentation, after Eisenberg and Barry [6]), and the first two were not included in the analysis, but rather were used to give the subjects practice in using this method of recording judgments (following Eisenberg [5]). This left 20 sets of representations for analysis.

For any given document, the subject was presented with three representations. In the TIA group, for example, the subject would see

- the title of the document
- the title and indexing
- the title, indexing, and abstract

making relevance judgments at each step, and then move on to the next document, repeating the process.

Subjects, then, saw more information about each document at each step, in addition to the information already shown. Thus, the relevance judgments obtained show how the user's perceptions of relevance of the documents changed as they saw more information about them. In all, 681 documents were retrieved, and 2043 relevance judgments were recorded.

IV. METHODS OF ANALYSIS

Since this is the first study known to us to generate this kind of data, the analysis techniques undertaken and presented here are largely novel and unique. As this study is exploratory and descriptive in nature, no research or statistical hypotheses were proposed a priori, and no post hoc inferential statistical testing was performed. However, upon consideration of the results of this study, some hypotheses are suggested, and these will be reported in the conclusions section.

When a retrieval set was returned by the subject, the marks on each line were measured to determine the subject's relevance judgment. For example, if the subject made a mark 62 mm from the left side of the line for a given representation of a document, the score would be 62. If a mark was made between 62 and 63 mm, the score would be 62. These scores were recorded, and entered into a computer program for analysis. Three scores, then, were obtained for each document—one for each representation the subject had seen.

Several types of analysis were undertaken in this study, and they are described below, in order of increasing depth of analysis. Results of these methods of analysis are presented in the next section.

Graphical inspection

After the relevance judgments were obtained, in keeping with the exploratory nature of this study, graphical representations of the judgment sets were developed. A sample graph (subject 11) is shown in Fig. 2.

From the very first, these graphs have had strong intuitive appeal. One gets an immediate impression of the simplicity or complexity of a judgment set from inspecting the graph, and can see at a glance whether judgments have risen or fallen (showing an increase or decrease in perceived relevance), stayed flat (showing no change in relevance), and under what conditions (what new information was seen) the changes occurred.

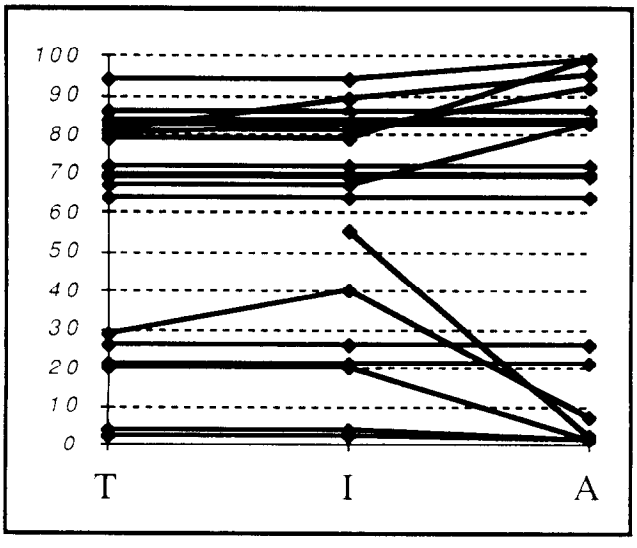

Fig. 2. Sample relevance judgment graph.

Simple changes in judgment

One can get a first-order picture of the effects of adding certain fields of information by simply looking at whether or not a subject's judgment has changed after that field is added. Looking at these proportions, over all subjects, will give a general idea of these fields' impacts. For example, do more judgments change as a result of the addition of abstracts than, say, the addition of indexing information?

The motion index

The two methods presented above (graphical inspection and simple changes) describe, in a variety of ways, the general "look" of the judgment set and how judgments are changing, but in a very simplistic way. We sought to develop a simple statistical measure that could be used to describe individual judgment sets and make comparisons among them.

The measure is called the motion index (*MI*), and is an indicator of the change that occurs in a judgment set. If, for example, the three judgments for a given document were 25, 50, and 60, its motion index would be 35, because, in all, the judgments "moved" by 35. The measure is not sensitive to direction of change: The motion index of 60/50/25 is also 35. The reason for this is that in many sets, roughly equal numbers of judgments go up and down. Preserving direction (and therefore, sign) of these changes would produce an *MI* of near zero, leading to difficulties in interpretation—is *MI* close to zero because almost no changes are made, or because they all cancel out?

Linear change (simple subtraction) was used rather than proportional change (ratios), in the belief that a change of judgment from 50 to 100 is more important than that from 1 to 2, and that the *MI* should reflect that.

Thus, for any given document, the maximum *MI* is 200 (100/0/100 or 0/100/0), and the minimum is 0. For each subject, an overall *MI* is calculated (for all document groups). In addition, other *MI*s are calculated: For the changes from first to second version of document seen (MI_1) and for the changes from second to third version (MI_2). Average *MI*s are also calculated for each of the above. Finally, MI_1 and MI_2 are divided by the overall *MI* to determine the proportion of total motion in a set that is attributable to different types of new information.

MI can be expressed in a formula:

$$MI_1 = \sum_i |RJ_{i2} - RJ_{i1}|$$

$$MI_2 = \sum_i |RJ_{i3} - RJ_{i2}|$$

$$MI = MI_1 + MI_2$$

where RJ_{ij} is the jth relevance judgment made of document i.

For example, for subject 11, in group TIA (the graph of which appears above), the following were derived:

$MI = 189$

$MI_1 = 22$ (indexing); avg $MI_1 = 1.16$, $MI_1/MI = 0.116$

$MI_2 = 167$ (abstract); avg $MI_2 = 8.35$, $MI_2/MI = 0.884$

Thus, over 88% of the total motion in this judgment set is due to the addition of the abstract; only 11.6% is due to the addition of indexing. Using this technique, we may see what effects the revelation of new information has.

To compensate for the loss of directional information in the *MI*, another measure, δ, was developed to show what proportion of movement in a set is positive. If one divides the actual motion in a set by the *MI*:

$$\frac{\sum_i RJ_{i3} - RJ_{i1}}{MI}$$

you get a measure from −1 (all motion negative) to +1 (all motion positive). This measure can be rescaled by performing a linear transformation, multiplying the expression above by 50, and adding 50:

$$\delta = \left(50 \cdot \frac{\sum_i RJ_{i3} - RJ_{i1}}{MI}\right) + 50$$

which produces a measure δ that can take on values from 0 (no positive motion) to 100 (all motion is positive). In fact, the same value could be arrived at for δ by summing all positive changes and dividing by the sum of the absolute value of all changes. For the example above (subject 11),

$$\begin{aligned}\delta &= \left(50 \cdot \frac{-29}{189}\right) + 50 \\ &= (50 \cdot -.153) + 50 \\ &= -7.65 + 50 \\ &= 42.35,\end{aligned}$$

so 42.35% of all motion in that set is positive.

The questionnaire

The final set of data obtained is on the post questionnaire, which subjects filled out after completing the judgments. The first three questions, which asked subjects to describe their perceptions of how more information affected their judgments, and the effects of the different document components they saw, were all scaled on 100 mm lines, so analysis is similar to that of the judgments themselves. Question 4 was open-ended, and asked subjects to specify what other kinds of information would have assisted them in judging relevance. The final question dealt with subjects' perceptions of their "break point" between relevant and nonrelevant documents, again along a 100 mm line. The results of this question, a replication of the work of Eisenberg and Hu [13], will be reported separately.

V. RESULTS

Using the methods of analysis described above, we can now make some statements about users' relevance judgments in general, and about the impact of different fields of information on those judgments in particular.

General observations

Each category of findings reported here was decided upon based on inspection of the graphical representations of users' judgments sets, as described above. The operationalization of the categories was guided by intuition. As such, they are highly subjective, but many of these observations are supported by quantitative data reported below.

Table 1 gives the operational definitions for each of these categories and an overview of the numbers of sets identified in each of the above categories. Please note that the figures in Table 1 will add across but not down. Not all sets fit into a category, and some sets fit into more than one.

In making these observations and attempting to describe the individual users' judgment sets, several features appear frequently and stand out. These include:

Few initial judgments. Many documents received no judgments based on the title alone. In eight sets, fewer than 75% of documents were judged based on title; the smallest

Table 1. Categories of graph characteristics

Category	Operational Definition	TAB	TAI	TBA	TIA	Sum
Few initial judgments	75% or fewer of documents judged based on title	2	2	1	3	8
Large swings under abstract	At least one judgment must have changed by at least 40 after adding abstract	3	8	7	7	25
Small but frequent movement under abstract	30% of the judgments changed by 10-40 after adding abstract	5	5	5	1	16
Stability under bibliographic/indexing	average MI for indexing/bibliographic ≤ 1.00	1	6	1	1	9
Movement under bibliographic/indexing	average MI for indexing/bibliographic ≥ 5.00	2	3	3	5	13
Binary	80% of judgments fall between 0-20 and 80-100	0	2	1	4	7
Ceiling effect	35% of judgments between 97-100	1	0	2	2	4
Floor effect	35% of judgments between 0-3	0	2	1	1	4
Generally increasing	$\delta > 90$	3	2	0	2	7
Generally decreasing	$\delta < 10$	0	2	0	0	2

proportion was 36.8%. The distribution of these sets seems to be even across field orders, and thus this would appear to be an individual trait.

Many of these "non-judgments" are on documents with unclear titles or do not address specific concerns expressed in the user's query. In several of the sets with very few (<50%) initial judgments, there are long strings of documents that are not judged based on title alone, but are judged when more information is given. These chains might be triggered by one ambiguous title, which then puts the user into a rhythm of waiting for more information to make judgments. For example, subject 8, whose query deals with the developmental effects of traumatic brain injury on children, failed to make a judgment on a document entitled "Brain disorder as a cause of behaviour change," perhaps because it did not explicitly address the traumatic nature of the disorder. This document eventually received a judgment of 12. The user then did not make initial judgments on the next three documents. Some other subjects in this category had similar long strings (#31 a string of four, #4 a string of eight, and #46 a string of eleven beginning with a very ambiguous title, "Masochism in a new key"), while some (subjects 19 and 21) did not. These findings could have interesting and serious implications for systems (e.g., INFOTRAC) that present only titles first.

Large swings under abstract. This indicates a substantial change in estimation of relevance based on adding the abstract. Of the 39 sets, 25 had a judgment change by at least 40 after adding the abstract. Subjects who saw abstracts last were slightly more likely (14 of 18) to make such swings than those who saw abstracts directly after titles (11 of 21). In all, there were 54 swings of 40 or more—27 were increases, 27 were decreases. The largest jump was 94, seen twice (2 to 96, 5 to 99).

"Small but frequent" movement under abstract. Several users ($n = 16$) frequently changed their judgments after seeing abstracts, but those changes were not the dramatic ones described above. Sixteen judgment sets fell into this category, only one of which was in the TIA group. The large number of sets in these two categories points out the extensive use of abstracts by subjects, a finding that will be corroborated later.

Stability under bibliographic or indexing information. Many subjects ($n = 9$) exhibited little or no movement of judgment when presented with this new information, producing flat judgment lines. Nine judgment sets showed very little motion after the addition of this information, five of which had no motion at all. Six of those nine were in the TAI group. This could be partially explained by the low use of indexing in general (discussed later), but it could also be that subjects got into a pattern, seeing title, abstract, then indexing, and deciding early that indexing would be of no help in judging, and learning to ignore it.

Considerable movement under bibliographic or indexing. Other subjects ($n = 13$) appeared to use this information more often, and thus changed their judgments more frequently when this was added. Given the results reported elsewhere in this report that show relatively little use of these fields, this result may be somewhat surprising—13 sets fall into this category. Eight of these are sets in the TBA and TIA groups, who saw this information after the title, followed by the abstract. A situation similar to the one described above, but in reverse, may have arisen. Without having seen the abstract, a user may find bibliographic or indexing information more important or useful in making relevance judgments. Subjects who saw abstracts last are more likely to change judgments based on bibliographic or indexing information, are more likely to make big changes based on abstracts, but are *less* likely to make frequent small changes based on abstracts. This could be yet further evidence of the power of abstracts in users' relevance judging behavior.

"Binary" sets. Several users ($n = 7$) gave final judgments that clustered at the top and bottom of the scale. A few of these appeared to be *ceiling effects* or *floor effects*, artifacts of the restricted range of response available to the user (the line was only 100 mm long; some subjects seemed to "run out of room" at times, wanting to go beyond the line). Other subjects simply divided their final judgments rather dramatically. This may be another individual trait.

General trends: increasing or decreasing. Some sets, when viewed graphically, seemed to have a general upward or downward trend, indicating that the subject's judgments generally increased or decreased as more information was viewed.

These five factors (binary, ceiling and floor effects, and general trends) are relatively uncommon. Binary sets could be an artifact of some subjects' attempts to make binary (relevant/nonrelevant) or binary-like decisions, or could be caused by the nature of the documents retrieved in those sets. Only eight subjects exhibited potential ceiling or floor effects, a small number, but a concern in the context of the method used for capturing judgments. A modification of the judging line might be useful in eliminating these effects. One of the initial questions behind this study was whether or not judgments would generally rise or fall, regardless of order of presentation or field order. Only nine sets showed this characteristic, seven of which are generally increasing. Again, this could be an individual trait, or could be due to the documents involved.

Title

The title was not "additional" information in the same sense that the other fields were, so we have no *MI* scores for the title. We do, however, have two sources of information about subjects' use of and feelings about the title: how often the initial judgment (based on the title) was maintained throughout, and ratings of titles' importance on the final questionnaire.

In all, 681 judgment sets were recorded. Of those, 165, or 24.23%, consisted of an initial judgment based on the title, which did not change as more information was presented. There are two possible explanations for a situation such as this. The further information may have either reinforced a subject's opinion about a document, or simply not changed the subject's mind. This question was not asked of subjects, but would be an interesting question for further study. The distribution of these stable judging sets (the position of the stable judgment on the 100 mm line) is interesting:

Range	# of sets	% of sets
100	10	6.1
90–99	23	14.0
80–89	13	7.9
70–79	15	9.1
60–69	10	6.1
50–59	6	3.7
40–49	3	1.8
30–39	6	3.7
20–29	6	3.7
10–19	19	11.6
0–9	53	32.3

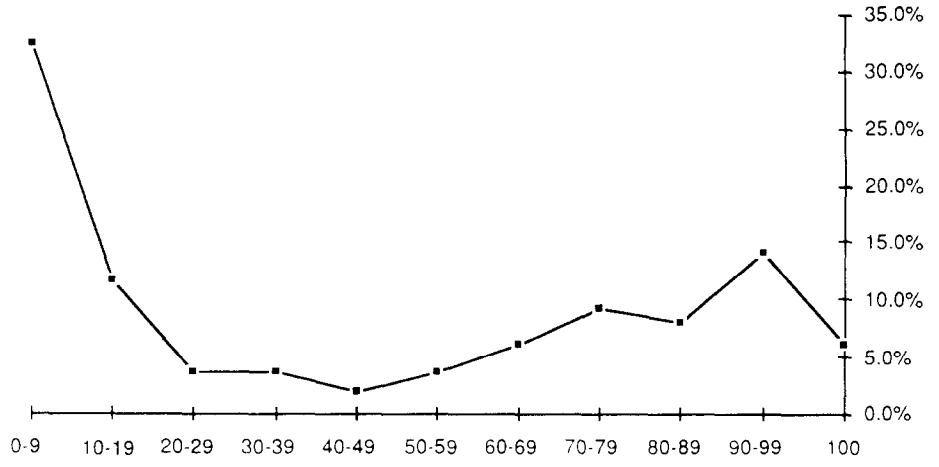

Fig. 3. Distribution of stable title judgments.

These judgments are concentrated at the two ends of the scale (72% are within 20 mm of each end of the line), with the greatest proportion at the low-relevance end. These results are interesting when compared with those of Saracevic regarding the partially relevant (P) judgments, described above.

In contrast, in 107 of the 681 judgment sets, the title was not seen as bearing enough information to permit a judgment (15.7% of all sets).

The users' opinions about the titles were recorded in the questionnaire administered after judgments had been made. The title was seen as quite important by most users, with users' marks on the 100 mm line provided averaging 67.59, with a median of 70, and a standard deviation of 19.26. However, there is great dispersion among those rankings. Eleven subjects (out of 39) rated titles at 50 or lower, with the lowest rating falling at 25. Only seven subjects clearly rated title as being the most important field, while 21 rated it second, and three rated it at least important. The remaining subjects had tied ordinal rankings. See Table 2 for a summary of these results.

Indexing

We do have "additional" information for the other fields (i.e., the effects seen when this information is added to what the user has already seen). In addition to the questionnaire data, we can see whether or not a judgment changed after the subject saw the new field, and by how much (measured by *MI*).

The indexing information was not perceived as very important by users. Its rating on the questionnaire (mean of marks on 100 mm line) was 21.15 ($SD = 19.15$), or about a third the rating of the title. These scores also form an interesting distribution. Most ratings (11

Table 2. Subjects' rating of importance of fields (from questionnaire)

Questionnaire rating of importance of:	average	median	range	sd
Title	67.59	70	25-100	19.26
Abstract	85.44	91	52-100	12.92
Indexing	21.15	13.5	0-67	19.15
Bibliographic	33.58	32	1-68	19.44

Table 3. Simple changes in judgments

when you add:	
Bibliographic	94 out of 324 judgments changed (29%)
Indexing	101 out of 357 judgments changed (28.3%)
Abstract	469 out of 681 judgments changed (68.9%)

out of 20) fall between 0 and 20, and there is a group of seven scores between 30–45 with an outlier at 67. All 20 subjects rated indexing as the least important field they saw; one of these rated it the same as title.

The actual impact of indexing information corroborates these opinions. The addition of indexing changed users' judgments only 101 out of 357 times (28.3%), and those changes were relatively small. The average *MI*s for indexing were 2.48 (in the TAI group) and 6.22 (in the TIA group), and the average proportion of motion due to indexing in those groups varied from 14.6% of all motion in TAI to 28.0% of all motion in TIA. Most users' (13 out of 20) judgments moved less than 15% as a result of indexing, but there were some marked outliers (50%, 55% and 77%). This information is summarized in Tables 3 and 4.

Bibliographic information

The situation is very similar for bibliographic information, with some subtle differences. Users rated bibliographic information lower than titles in importance on the questionnaire, but higher than indexing. The average importance rating for bibliographic was 33.58 ($SD = 19.44$), about half the rating of the title, but 50% higher than indexing. This distribution is quite flat, with few peaks or outliers, but a wide range (1–68). Most subjects (15 out of 19) rated bibliographic information as their least important field, three rated it as second most important, and one rated it tied with title as second.

Of the 324 judgment sets that included bibliographic information, 94 (29%) changed when it was added, slightly greater than the figure for indexing. The magnitude of these changes is also small, with average *MI*s of 3.40 (for TAB) and 5.50 (for TBA). The proportion of motion accounted for is quite consistent: 26.0% on average in TAB, and 26.7% on average in TBA. The majority of scores (12 out of 19) ranged from 10% to 30%, but there was a small cluster from 35% to 50%, and an outlier at 65%.

For both indexing and bibliographic information, an interesting pattern emerges. In the TIA and TBA groups (where abstract came last), the average *MI*s for bibliographic or indexing is higher than the reversed orders (TAI or TAB), and the standard deviation is

Table 4. Motion index averages and fractions

		Motion Index		Fraction of all motion in a set due to this field	
		average	sd	average	sd
for Abstracts	TAB	11.80	4.99	74.0%	20.2
	TAI	14.43	6.46	85.4%	19.5
	TBA	13.91	9.12	73.3%	13.6
	TIA	15.57	9.59	72.0%	25.6
for Indexing	TAI	2.48	3.17	14.6%	19.5
	TIA	6.22	5.39	28.0%	25.6
for Bibliographic	TAB	3.40	2.81	26.0%	20.2
	TBA	5.50	5.60	26.7%	13.6

higher as well. This could be another instance of the effect of abstracts on judgment—perhaps seeing the abstracts before this more auxiliary information inhibits their impact in some way.

Abstract

Abstracts were by far the most highly rated field, and produced the greatest number and magnitude of changes in users' judgments. The average rating on the questionnaire was 85.44 ($SD = 12.92$), and 29 of 39 subjects rated abstracts at 80 or higher. There is a small negative tail, and 2 possible outliers at 52. Of the 39 subjects, 26 rated abstract as clearly the most important field they saw, 7 rated it as second most important, and 6 rated it as tied with title for most important.

The addition of abstracts produced changes in 469 of the 681 judgment sets (68.9%), and the average changes (by user) ranged from 11.80 (TAB) to 15.57 (TIA). The proportion of motion due to abstracts was also high, ranging from 72.0% (TIA) to 85.4% (TAI), and averaging 76.9% overall. Most proportion scores (31 out of 39) were 60% or above (including 5 at 100%—all motion due to abstract), with a long tail extending to outliers at 35% and 23%.

Other questionnaire results

Three other questions were posed on the post questionnaire. The first two asked whether more information had helped users in their judgments or confused them. No subject gave the second a higher rating than the first, and the averages are quite different. The average rating, on a scale of 100, for helpfulness was 72.46 (median = 76, range = 13–100, $SD = 21.34$), and the average rating for confusion was 4.49 (median = 3, range = 0–34, $SD = 6.16$).

The last question asked subjects to list information about the documents that was not presented but would have assisted them in making their decisions. The most often requested pieces of information were name of journal or source ($n = 10$) and author's name ($n = 10$), all from subjects who did not see bibliographic information. It is interesting to note that no subjects who did not see indexing requested it. Other requested information included author's affiliation ($n = 3$), date of publication ($n = 2$), and document type ($n = 2$). See Table 5 for a full summary of these results.

VI. CONCLUSIONS

Confirming previous studies

The findings reported here corroborate with the theoretical conclusions of Schamber et al., the empirical results of Saracevic, and the methodological results of Eisenberg.

Table 5. Other information requested (from questionnaire)

Information Requested	TAB	TAI	TBA	TIA	Total
Journal/source name	0	4	0	6	10
Author's name	0	8	0	2	10
Author's affiliation	0	1	1	1	3
Date of publication	0	1	0	1	2
Document type	0	1	0	1	2
Number of references	1	0	0	0	1
Length of article	0	1	0	0	1
"Citation index" [sic]	0	0	1	0	1
"'How good'...journals are"	0	0	1	0	1
"Type of journal (peer-reviewed v. popular press"	0	0	1	0	1
Whether article is "research, theory, or experience-based"	0	1	0	0	1
Totals	1	17	4	11	33

Schamber's conclusions on relevance (that it is a measurable, dynamic concept dependent on users' perceptions) are certainly in line with these results. Saracevic stated that using different presentation formats affected judgments, that judgments of partial relevance were unstable, and that relevance is a continuous variable. The present findings coincide here, as well. Eisenberg claimed that magnitude estimation techniques could be used to measure relevance, and this study has replicated and lent support to his claims.

Differing from previous studies

On the other hand, these findings differ from those reported by Marcus et al. regarding the relative usefulness of the various fields. They report far higher indicativity for "subjects" (here, indexing) than one might expect based on the present results, even higher than for titles. Recall, though, that indicativity is not a measure of use; rather, it measures how often judgments based on a single field agree with those based on full text. In this context, the distinction from the present results is clear: Users do not use subject/indexing information very often when judging relevance, but when they do, those judgments tend to be borne out by more complete information. Marcus et al.'s length hypothesis, a correlation between length of a field and its indicativity, is not present in our results—the abstract is the longest, most often used and most extensively used field, but the title is used far more than either indexing or bibliographic information, which are often longer.

New findings

This study is the first to be able to make claims about the behavior of relevance judgments as information is incrementally presented to users. Relevance judgments do change as information is added, and the degree of this change can be measured by calculating simple changes in judgment and the motion index.

The use of the 100 mm line is validated, although some modifications to reduce the likelihood of ceiling or floor effects in judgments is indicated. Subjects reported no difficulty in using the line to record judgments, and the data received as a result is of high quality and at a ratio level. Therefore, the use of this methodology can be recommended to other researchers, under the circumstances described in the methods section above (random presentation of documents, use of practice stimuli).

Most importantly, virtually every measure examined in this study has shown the same pattern of importance and use of the four fields studied. Clearly, the abstract is the most important and most used single piece of information in relevance judging. Titles are important, but less so than abstracts. There is then a considerable drop to bibliographic information, and finally, indexing. Each field was used, and the combination of many types of information allows the best judgments, but this pattern persists.

VII. FURTHER RESEARCH

Some further research building on these findings has already been undertaken, and will be reported separately. Other people have been asked to make judgments on some of the document sets retrieved as part of the present study. This second study (conducted with Reneé McKinney) seeks to examine how these secondary judges' judgments differ from those of the original users.

Other questions have arisen during the course of this research that would be interesting to examine in more detail. When a user reports that the relevance judgment has not changed as a result of seeing more information, is that because the previous judgment has been confirmed, or because the user's mind simply has not changed?

The subjects in this study were faculty and doctoral students at a major research university. How do users in other settings (public library users, undergraduates, faculty in other disciplines, children) make relevance judgments? Will there be similar patterns? Will new pictures emerge?

Some more "statistical" questions have also been generated as a result of these findings and methods. These results suggest that abstracts have a bigger impact on users' judgments the later they are seen; this suggestion could be more completely and rigorously tested. Is there a correlation between users' perceptions of the importance of fields to their

judgments (as per the questionnaire) and actual use? Is there a correlation between the average amount of movement in a user's judgment (the motion index) and final relevance judgments? Both of these showed weak positive correlations here, but a structured test of statistical significance would aid in our understanding.

In a further exploratory vein, asking users to generate categories or brief descriptions of their relevance judgments and correlating them with judgments made based on this methodology would be extremely interesting.

The methodology used here could also be used in other research not explicitly studying relevance, but that uses relevance as a measure of searcher or retrieval system performance. Taken with the work of Eisenberg, Schamber, and their colleagues, the results presented here make a case for re-examining the role of relevance and its measurement in information retrieval.

Finally, what is the effect of non-relevant documents? If one defines "non-relevant" documents as those whose judgments here fell to the left of the mark the subjects made for the final question on the questionnaire, 301 of the 681 documents judged here were non-relevant (about 44%). Do subjects just ignore these documents, or do they have an effect on later (or even previous) judgments? The author is indebted to Jeffrey Katzer, for planting this idea some years ago; the question is still an interesting one.

Acknowledgements—The author would like to thank the members of the project staff: Catherine Allen, Paul Crandall, Heather Farnan, Matilda Flores, Sandra Goldstein, Jane Ploughman, Jeffery Ring, Daniella Williams, and Leona Williams for their invaluable assistance in carrying out the research reported here.

REFERENCES

1. Janes, J.W. Towards a search theory of information. Doctoral dissertation, Syracuse University, Syracuse, NY; 1989.
2. Schamber, L., Eisenberg, M.B., Nilan, M.S. A re-examination of relevance: Toward a dynamic, situational definition. Information Processing and Management, 26:755–776; 1990.
3. Saracevic, T. Relevance: A review of and a framework for the thinking on the notion in information science. Journal of the American Society for Information Science, 26(6):321–343; 1975.
4. Eisenberg, M.B. Magnitude estimation and the measurement of relevance. Doctoral dissertation, Syracuse University, Syracuse, NY; 1986.
5. Eisenberg, M.B. Measuring relevance judgments. Information Processing and Management, 24(4):373–389; 1988.
6. Eisenberg, M., Barry, C. Order effects: A study of the possible influence of presentation order on user judgments of document relevance. Journal of the American Society for Information Science, 39(5):293–300; 1988.
7. Rath, G.J., Resnick, A., Savage, T.R. Comparisons of four types of lexical indicators of content. American Documentation, 12:126–130; 1961.
8. Resnick, A., Savage, T.R. The consistence of human judgments of relevance. American Documentation, 15:93–95; 1964.
9. Thompson, C.W.N. The functions of abstracts in the initial screening of technical documents by the user. Journal of the American Society for Information Science, 24:270–276; 1973.
10. Marcus, R.S., Kugel, P., Benenfeld, A.R. Catalog information and text as indicators of relevance. Journal of the American Society for Information Science, 29:15–30; 1978.
11. Hagerty, K. Abstracts as a basis for relevance judgment. Master's thesis, Graduate Library School, University of Chicago, Chicago, IL; 1967.
12. Saracevic, T. Comparative effects of titles, abstracts, and full texts on relevance judgments. Proceedings of American Society for Information Science, Washington DC, 6:293–299; 1969.
13. Eisenberg, M., Hu, X. Dichotomous relevance judgments and the evaluation of information systems. Proceedings of the American Society for Information Science 50th Annual Meeting, 1987, Medford, NJ: Learned Information Inc.; 1987.

APPENDIX A: SEARCH TOPICS

1 undergraduate-initiated research
2 employee suggestion programs, especially in universities
3 empowerment in schools, hospitals, professional relationships
4 how animals extract information from environment
6 influence of person's life philosophy on psychological health
7 circadian and seasonal rhythms of lizards and snakes
8 psychosocial and behavioral effects of traumatic brain injury in children
9 influence of weather variables on mood; seasonal affective disorder

11 how people get ahead in status and dominance hierarchies
12 children's naive mechanics; gears and movement understanding
13 economics of child care
14 selection of fellowship recipients
15 age and gender differences in self-esteem
16 source reduction; reducing waste at point of production
19 Alzheimer's Disease, aging, and driving
20 marginal groups; commonalities between how they manipulate their environment
21 latch-key children and homeless people in libraries; how do organization and people cope
22 relationship: age and achievement; is there an "over the hill"
23 recruitment and retraining of girls and women in science
24 treating incest survivors: transferrence, group therapy, feminist analyses
25 memory and problem solving in psychology
26 history and evaluation of case study method
27 young children's understanding of mental states and processes
28 operational definition of efficiency in education
29 effect of premarital pregnancy/birth on mental well-being of black and white newlywed couples
31 expectations about relationships between variables
32 psychological effects of sexual child abuse of boys
34 effects of alcohol consumption on memory
36 processes by which males and females derive and maintain global self-esteem
37 period of development in which people form their identity
38 ways people think about goals along dimension of concreteness v. abstractness, as relates to role of affect
39 experience and personal meaning of parenthood for black men and women, especially fathers
40 effects of parenting patterns, childhood discipline, and early experiences on adult career/vocational choice
42 motor production in singing
43 auditory information processing and its neurophysiological correlates
45 age differences in social/emotional/psychological development among people w/leukemia or lymphoma
46 female sexual masochism
47 relationship: parenting and borderline personality disorder
48 adult friendships and social support

APPENDIX B: JUDGING INSTRUCTIONS

You have expressed a need for information, which is attached.

A set of descriptions of documents has been compiled in response to this information need.

In this experiment, we would like to find out how relevant various document descriptions appear to you in relation to the stated information need. For this purpose, you will be asked to look at a series of document descriptions one at a time. Your task will be to make a mark on a line corresponding to your impression of the degree of relevance of that document to your query, from none (N) to total (T).

Here is a sample line:

N _____ T

As a preliminary exercise, can you imagine a document that would be highly relevant? Can you imagine a document that you would judge to be low in relevance? Can you imagine one that would be medium in relevance?

You will now see a series of document descriptions. These document descriptions have been marked to assist you in reading them. Tags have been placed at the left-hand margin to tell you what information you are seeing. The tags are:

AB abstract of the document
AU author of the document
JO journal in which the original document appeared
PY year in which the document was published
TI title of the document

You will see **three** versions of each document. For each, you will first see the title of the document alone. Your task will be to make a mark on a line corresponding to your impression of how relevant the document is. Next, you will see the title and the bibliographic information about the document. Again, make a mark on the line corresponding to your judgment of the relevance of the document. Finally, you will see the title, the bibliographic information about the document, and the

abstract, and you will make a mark on the line. **Do not look ahead.** Please make your judgments about the relevance of the documents based **only** on the information which you see at a given time.

If you do not feel you have enough information to make a decision, or if your judgment is the same as for the previous version of the document, check the appropriate box beneath the line.

Feel free to take as much time as you like in making your decisions.

When you have made judgments on all documents in this package, please place the entire package in the envelope, attach the enclosed mailing label, and return it through campus mail. The envelope marked #2 contains an unprocessed copy of the retrieval set we obtained for your query, and may include documents not in the experimental set. It is yours to keep with our compliments.

Thank you for your assistance in this study.

TBA

APPENDIX C: QUESTIONNAIRE

Please answer these few questions, and return this questionnaire along with the document set and your relevance judgments.

1. Did more information about these documents help you in making decisions about their relevance to your query?

   None of the time _____ 100% of the time

2. Did more information about these documents confuse you in making decisions about their relevance to your query?

   None of the time _____ 100% of the time

3. Please rate your perception of the importance of these parts of documents in making decisions about their relevance:

	Not at all Important	Totally Important
Abstract	_____	
Title	_____	
Bibliographic Information	_____	

4. Was there any other information about the documents you did **not** have which would have assisted you in making decisions? Please list here:

5. Assume that the following line represents a continuum of relevance from 0 relevance (non relevance) to complete relevance. If your only choice was to state that a document of citation is NR or R, where would you draw the break point? That is, how much 'relevance' does a document have to have before you consider it relevant?

_____