A Study of Information Seeking and Retrieving. II. Users, Questions, and Effectiveness*

Tefko Saracevic
School of Communication, Information and Library Studies, Rutgers, The State University of New Jersey, 4 Huntington St., New Brunswick, N.J. 08903

Paul Kantor
Tantalus Inc. and Department of Operations Research, Weatherhead School of Management, Case Western Reserve University, Cleveland, Ohio 44106

The objectives of the study were to conduct a series of observations and experiments under as real-life a situation as possible related to: (1) user context of questions in information retrieval; (2) the structure and classification of questions; (3) cognitive traits and decision making of searchers; and (4) different searches of the same question. The study is presented in three parts: Part I presents the background of the study and describes the models, measures, methods, procedures and statistical analyses used. Part II is devoted to results related to users, questions and effectiveness measures, and Part III to results related to searchers, searches and overlap studies. A concluding summary of all results is presented in Part III.

Summary of the Study

This is a second article in a series of three, reporting on a study of information seeking and retrieving. The first dealt with the methodological aspects describing the aim, objectives and approach, related works, and models, measures and procedures used, including references appropriate for the study as a whole [1]. This second part concentrates on results connected with users, questions, and effectiveness measures. The third part concentrates on results connected with searchers, searches, and overlap studies. A Final Report together with appendices was deposited with ERIC and NTIS [2]; it contains the details of the study with emphasis on procedures and presentation of "raw" data. A summary of methods used in the study is provided here, so that a reader interested in results only could proceed without having to read "Part I: Background and Methodology."

The aim of the study was to contribute to the formal, scientific characterization of the elements involved in information seeking and retrieving, particularly in relation to the cognitive decisions and human interactions involved. The objectives were to conduct experiments and observations under as real-life conditions as possible related to: (1) user *context* of questions in information retrieval; (2) the structure and classification of *questions;* (3) cognitive traits and decision-making of *searchers;* and (4) different *searches* of the same question. Models and measures were developed to reflect the nature and effects of these four classes of variables.

Forty users each posed one written questions related to their ongoing research or work. In addition, during an interview each user supplied a tape recorded statement on the underlying problem and intent of his or her research. Each user also marked measures on the context of the question dealing with problem definition, intent, internal knowledge, and public knowledge estimate. Thirty nine searchers were assembled: 36 so called "outside" searchers who were paid to search five or six questions[1] based on the user's written question ("outside searches") and three "project" (full time staff) searchers who conducted four different types of searches ("project searches"). The project searches were based on:

(1) The user's tape recorded problem statement only;
(2) the taped problem statement *plus* the written question;
(3) terms from the written question only *without* elaboration, and
(4) terms from the written question *plus* elaboration by thesaurus.

*Work done under the NSF grant IST85-05411 and a DIALOG grant for search time.

Received July 9, 1987; accepted December 11, 1987.

[1]There were five outside searches per question, but since there were 36 searchers and 40 questions, some searchers did six searches.

Each searcher was tested on three cognitive tests: Remote Associates Test (RAT) designed to test ability in making word associations; Symbolic Reasoning Test (SRT) designed to test the ability in making deductive inferences from symbolic inequalities; and Learning Style Inventory (LSI) designed to characterize an individual according to preferred style of learning. Searchers also indicated their frequency of DIALOG use. Searching was done on DIALOG. A single DIALOG database was searched for each question; altogether 21 different databases were used.

A model for question structure and a classification scheme for questions were developed and then tested in a separate experiment. Twenty-one judges (separate from searchers) were assembled to judge the structure and classification of the questions used in the experiment proper. Each one judged 20 questions. Consistency of judgments was compared. Different classes of questions as assigned by the judges were used for correlation with retrieval results.

Each of the 40 questions was searched by five outside searches and four project searches. The output from these nine searches was merged into a union, i.e., duplicates were eliminated. The union was then sent to the user for evaluation. If the union exceeded 150 retrieved items, only the most recent 150 items were sent to the user. (For three questions that slightly exceeded the 150 limit we sent all items retrieved). This was done to avoid user overload and to ensure return. Users indicated whether each item was relevant (R), partially relevant (pR), or not relevant (N). Additionally, users scored five utility measures: worth of their involvement in comparison to time it took, amount of time spent in evaluation, estimated dollar value to them, contribution to problem resolution, and overall satisfaction. The user evaluations of relevance were used as benchmarks for figuring relevance odds, and precision and comparative recall of the searches and for study of the other variables.

Statistical analyses involved study of distributions, analyses of variance, regression analysis, and logarithmic cross product ratio analysis. The last is a powerful technique describing the relation between a given variable and the odds or chances that a retrieved item be relevant as opposed to not relevant. It also relates a given variable to the odds that precision or recall be above the mean. Analyses were done on two levels in search of variables which could provide significant explanations of the observed processes: micro or item-wise analysis and macro or search-wise analysis. On the item-wise level, impact of variables on the odds that retrived items be relevant or partially relevant (as opposed to not relevant) was considered. On the search-wise level, impact of variables on precision and comparative recall was considered. Conclusions are grouped according to main classes of variables involved in the study, describing users, questions, searchers, and searches.

Overall Retrieval

How Many Items Were Retrieved?

Each of the 40 questions used in the study was searched nine times (five outside searches and four project searches); thus there was a total of 360 searches (200 outside and 160 project searches). The items are summed in two ways:

(1) Sum of all items retrieved over all searches *including* duplicates, i.e. number of items retrieved by each search summed over all searches for 40 questions where duplicates of identical items retrieved for the same question were NOT eliminated.
(2) Sum of unique items retrieved over all questions *excluding* duplicates, i.e. number of items retrieved after duplicates for the same question were eliminated.

The difference between the two sums is the same as the difference between tokens and types in word counts. One is a sum, the other a union.

The following were figures on retrieval:

- The 360 searches for 40 questions retrieved all together 17,708 items; of these, 11,804 items were unique and 5,904 items were retrieved by more than one search.
- Of the 17,708 items retrieved by all searches, 8956 (or 51%) were evaluated, the rest were not. The not evaluated items belong to the set of items exceeding 150 per question; the evaluated items were used in various analyses and the not evaluated items were not used in any way.
- Of the 11,804 unique items retrieved, 5411 (or 46%) were evaluated by users.
- To recapitulate, the sum of items retrieved *and* evaluated (counting search by search for 360 searches) before elimination of duplicates consisted of 8956 items; after elimination of duplicates, the total number of unique items was 5411 (counting question by question the items sent to users).

How Were they Judged as to Relevance?

The data on relevance judgments by users is also summed in two ways:

(1) Items retrieved by all searches, *including* duplicates, as presented in Section I of Table 1. This data served as benchmark for analysis of variables related to searches and searchers.
(2) Unique items retrieved by all searches, *excluding* duplicate, as presented in Section II of Table 1. This data served as benchmark for analysis of variables related to users and questions.

As can be seen:

- Of 8,956 items retrieved by all searches, some 59% were judged relevant or partially relevant, and 41% were judged not relevant.
- Of the 5411 unique items retrieved, some 52% were judged relevant or partially relevant, and 48% were judged not relevant.

Table 2 provides means, standard deviation, and range of items retrieved per question and per search. Calculations per question were done on the basis of the union of all nine searches for each question (where duplicates were elimi-

TABLE 1. User relevance judgement on all items retrieved and evaluated (including duplicates) and on unique items retrieved and evaluated (excluding duplicates). (N questions = 40; N all searches = 360; N outside searches = 200; N project searches = 160; N all items retrieved = 8956; N unique items retrieved = 5411. Note: Section I of the table on all items retrieved refers to the sum of retrieved items over all searches for the 40 questions, and Section II on unique items retrieved refers to the union of items retrieved for each question summed over 40 questions).

User Judgement	Retrieved by outside searches No.	%	Retrieved by project searches No.	%	All searches No.	%
I. All items retrieved (including duplicates):						
Relevant	1378	28.4	1371	33.3	2749	30.7
Partially relevant	1326	27.5	1212	29.5	2538	28.3
R + pR	2704	55.9	2583	62.8	5287	59.0
Not relevant	2137	44.1	1532	37.2	3669	41.0
Total	4841	100.0	4115	100.0	8956	100.0
II. *Unique items retrieved (excluding duplicates):*						
Relevant	924	25.1	830	28.4	1343	24.8
Partially relevant	973	26.3	861	29.5	1448	26.7
R + pR	1897	51.4	1691	58.0	2791	51.5
Not relevant	1794	48.6	1229	42.1	1620	48.5
Total unique items	3691	100%	2920	100%	5411	100%

TABLE 2. Items retrieved per question and per search. (N questions = 40; N all searches = 360; N unique items retrieved i.e. excluding duplicates (used for calculation per question) = 5411; N all items retrieved i.e. including duplicates (used for calculation per search) = 8956).

Number of	Mean	Standard Deviation	Min.	Max.
Items per question:				
Relevant	33.6	25.3	1	113
Partially relevant	36.2	24.4	4	135
Not relevant	65.5	37.3	0	156
Total	135.3	36.3	35	229
Items per search:				
Relevant	7.6	10.7	0	86
Partially relevant	7.1	10.9	0	113
Not relevant	10.2	13.9	0	89
Total	24.9	24.5	0	125

nated) while the calculations per search were done on the basis of actual output for each of the searches (where duplicates were *not* eliminated).

The Appendix contains a summary of the text for each question, together with information on the DIALOG file searched and the number of items judged relevant, partially relevant and not relevant. The summary of per question data in Table 2 comes from there. Altogether 21 DIALOG files were searched; the effect of different files of retrieval was not investigated here, because it was not in the objectives of the study to investigate systems variables. The precision and recall of searches are treated later in the article.

As can be seen, the mean number of items judged relevant or partially relevant per question (i.e. $R + pR$) approximately equaled the number of items judged not relevant, however, the range (minimum and maximum numbers) was quite wide. The same is not true on the per search basis: the mean number of relevant or partially relevant items per search was significantly larger than not relevant ones and the range is also large. However, a caveat is in order regarding the interpretation of these and all other means and standard deviations in this study: the distributions observed are *not* normal (bell shaped); some distributions are skewed toward an end value, others have several peaks and valleys. Thus, the means are not typical as in a normal distribution (i.e. most cases do not have the mean value indicated) and the standard deviations do not represent the width of a central peak. Results should be interpreted accordingly.

In general, the mean numbers of items per question and mean number of items per search differed to a great extent. This suggests that different searches for the same question retrieved substantially differing sets of items, and as a consequence, when all searches for the same question were put together, the total number of unique items had to be much larger than for any single search. Overlap studies reported in Part III explore this further.

Users

Who were the Users and How Did their Intended Application Affect Precision?

The study involved 40 users with one question each: 19 (48%) faculty, 15 (37%) graduate students, and 6 (15%) from industry. As to the application, the users indicated the following: 17 (43%) of the questions were related to faculty research; 14 (35%) to graduate study; 5 (13%) to industrial research; and 4 (10%) to general work.

Table 3 provides the overall precision for questions in each application category:[2]

As can be seen, the mean overall precision for the three larger categories [graduate study (.50), faculty research (.49), and industrial research (.49)] are almost identical. The analysis of variance showed no significant difference in the overall precision of questions with different application. However, while the means were about .50, the range for precision of individual questions in each category is very wide.

[2]As mentioned, overall precision for a question as a whole is calculated on the basis of the relevant or partially relevant items in the union of all (nine) searches. While the overall precision for a question can be calculated in such a way, the overall recall cannot, because we do not know how many relevant items for a question were left unretrieved from a given DIALOG file. Thus, we cannot give overall recall. However, comparative recall for the searches that contributed to the union was calculated, using the relevant items in the union as denominator.

TABLE 3. Overall precision for questions according to types of application for which information was requested by users (N users = 40; N questions = 40).

Application	No. of Questions	Mean Precision	Stand. Deviation	Minimum	Maximum
Faculty Research	17	0.50	0.24	0.14	0.91
Graduate Study	14	0.49	0.24	0.12	0.79
Industrial Research	5	0.49	0.25	0.11	0.79
General	4	0.71	0.38	0.25	1.00

For other variables we will report relevance odds and precision and recall odds. Unfortunately we could not do that for application variables, because the method requires that a variable be divided into two classes only and for applications we had four classes. This shows a limitation of an otherwise powerful method.

How Was the Context Assessed by the Users and by Searchers?

Each user assigned a value for the context measures pertaining to his or her perceptions of the following.

(1) How well was their *problems* defined? [From 1 (poorly defined) to 5 (well defined)];
(2) How well was their *intent* formulated? [From 1 (open to many avenues) to 5 (well defined)];
(3) What was their estimate of the probability that *public knowledge* existed on the subject of their question? [From 1 (low level—highly improbable that it exists) to 5 (high level—it exists)];
(4) What was their *internal knowledge* on the problem giving rise to the question? [From 1 (little personal knowledge) to 5 (considerable personal knowledge)].

Project and outside searchers were also asked to assign context scores for questions they searched so that degree of agreement between users and searchers may be observed. While the project searchers (having access to both taped problem statements and written questions) could score on all four context variables, the outside searchers (having access only to written questions) could score only on the last two. Clearly, the searchers' score on Internal Knowledge reflects their own and not the users'.

The results are presented in Table 4 subdivided in three sections to reflect separately assignments by users, project searchers, and outside searchers. The table provides the cumulative number of scores for each of the five values available for a context characteristic over 40 questions. For example, it shows that for their questions, 10 users indicated 5 (well defined) on Problem Definition and 9 project searchers did the same. For outside searchers the data represents not the cumulative but the average number of questions (rounded to the nearest number) with a given score (the average over 5 searchers for a question). For instance, for 9 questions, the average score by outside searchers on Public Knowledge was 5 (highly probable that it exists).

TABLE 4. Summary of scores in context characteristics as assigned by users, by project searchers and by outside searchers (N questions = 40; N users = 40; N project searchers = 3; N outside searchers = 36). Note: The sum of scores in each context variables (each row) equals 40, the number of questions.

Context Variable	No. of Questions Assigned the Score 1	2	3	4	5	Not Assigned
I. Users						
Problem definition	1	6	10	13	10	
Intent	6	12	7	9	6	
Public knowledge	1	9	6	12	12	
Internal knowledge	0	4	18	15	3	
II. Project searchers						
Problem definition	6	7	7	11	9	
Intent	9	10	7	7	7	
Public knowledge	10	10	8	7	5	
Internal knowledge	24	10	4	0	2	
III. Outside searchers						
Public knowledge	1	2	12	11	9	5
Internal knowledge	14	10	7	6	1	2

The results indicate that:

- On Problem Definition, 58% of users considered their problem well defined (top two scores, Section I, Table 4); in contrast, 50% of project searchers (Section II) considered it to be in the same range;
- On Intent, 45% of users thought that their intended use could be "open to many avenues" (lowest two scores); in contrast, 50% of project searchers indicated the same;
- On Public Knowledge 60% of users indicated that there was close to certainty that information requested exists (top two scores); in contrast, 30% of project searchers and 58% of outside searchers (of those who scored, Section III) believed that that is the case for the questions they searched;
- On Internal Knowledge, 45% of users considered themselves quite knowledgeable about the problem at hand (top two scores); as expected, the project and outside searchers indicated the opposite: the project searchers indicated for 85% of questions and the outside searchers for 60% of the questions as having quite low knowledge about the problem at hand (lowest two scores).

In general, users and searchers had a significant agreement on two context variables: Problem Definition and Intent. There was lesser agreement on the estimate of Public Knowledge. The expected disparity on Internal Knowledge did materialize. This suggests that to quite a large extent searchers can approximate or predict users' estimates of the context variables.

What was the Relationship between Context Variables and Relevance Odds?

The following question was asked: What were the odds that retrieved items be relevant or partially relevant (as opposed to not relevant) in the questions for which a given context characteristic was assessed as high (above mean) as opposed to questions for which it was assessed low?

A method called cross product ratio analysis (described in detail in Part I) was used to answer this question, i.e., to

study the relation between relevance odds and context characteristics (and other independent variables as well). This is done by first constructing a 2 × 2 contingency table in which the independent variable is broken into a number of retrieved items by searches or for questions of high value (i.e. above a cut point or mean) and low value (below a cut point), and each of these is further divided into items judged not relevant, and relevant or partially relevant. The cross product ratio taken next reflects the increase in odds that an item be relevant or partially relevant due to moving from a low value of the variable (or below mean) to a high value of the variable (or above mean). A logarithm of the cross product ratio is taken to obtain a symmetrical distribution, allowing a *t*-test of the statistical significance of the observed effect. A *t*-value above 2 indicates statistical significance at 95%. (For an example of calculation see Part I of this article.)

Table 5 presents the end results or summary of calculations on the relation between the four context characteristics of questions as assessed by users and the odds that a retrieved item be relevant or partially relevant as opposed to not relevant. Table 5a presents the 2 × 2 contingency tables containing the values used for cross product ratio calculations of relevance odds in Table 5. The analysis includes the unique items retrieved by all (outside plus project) searches, that is, the set of items related to questions as a whole, rather than individual searches. The context of a question is the same for both outside and project searches, thus the inclusion of combined output from both. The effects of project searches alone are treated in Part III in the section "Project Searches."

We give a detailed description of the organization and contents of these two tables. This description is applicable mutatis mutandis to six additional tables organized the same way and report relevance odds of other variables (Tables 7, 10, 15, and 17 in this part, and Tables 20 and 23 in Part III), allowing for a briefer description of these other tables. However, for these other tables, in order to save space we are *not* providing the related 2 × 2 contingency tables similar to the ones presented in Table 5a. We are providing and describing the contingency tables in Table 5a to demonstrate for interested readers the basis for calculation of summary results on any and all relevance odds.[3]

TABLE 5. Summary of the relation between context characteristics assessed by users and the odds that a retrieved item be relevant or partially relevant (*N* users = 40; *N* questions = 40; *N* all searches = 360; all characteristics indicated on a scale from 1 to 5; statistical significance at 95%).

Context characteristic	Cut point (Mean)	Odds ratio	Log odds	Stand. error +/−	*t*-Value	Stat. Signif.
Problem definition	3.67	1.21	0.19	0.05	3.43	Yes
Intent	2.93	0.92	−0.08	0.05	−1.48	No
Public knowledge	3.63	1.67	0.51	0.06	9.05	Yes
Internal knowledge	3.48	0.94	−0.06	0.05	−1.16	No

TABLE 5a. Values for the calculation of relevance odds for context characteristics assessed by users (*N* unique (*R* + *pR* + *N*) items retrieved by all searches = 5411; *N* relevant or partially relevant (*R* + *pR*) = 2620; *N* not relevant = 2791; values in boxes = no. of items retrieved).

		Problem Definition		
		Below Mean	Above Mean	Total
Items retrieved	NREL	1217	1403	2620 48.4%
	R + *pR*	1167	1624	2791 51.6%
	Total	2384 44.1%	3027 55.9%	5411 100%
		Intent		
		Below Mean	Above Mean	Total
Items retrieved	NREL	1178	1442	2620 48.4%
	R + *pR*	1311	1480	2791 51.6%
	Total	2489 46%	2922 54%	5411 100%
		Public Knowledge		
		Below Mean	Above Mean	Total
Items retrieved	NREL	1143	1477	2620 48.4%
	R + *pR*	884	1907	2791 51.6%
	Total	2027 37.5%	3384 62.5%	5411 100%
		Internal Knowledge		
		Below Mean	Above Mean	Total
Items retrieved	NREL	1199	1421	2620 48.4%
	R + *pR*	1321	1470	2791 51.6%
	Total	2520 46.6%	2891 53.4%	5411 100%

The columns in Table 5 provide the following data for each variable:

Column 1: The mean of the particular variable serving as the cut point, used to divide items into those retrieved for questions connected with an above mean (or high) values of the variable and those with below mean (or low) values. In Table 5, this is the mean of assignments by 40 users on a scale from 1 to 5 about the context of their question.

Column 2: The cross product ratio or the relevance odds. This is calculated from the 2 × 2 contingency tables presented in Table 5a. If the values in the first row of a 2 × 2 table are labeled as A and C, and in the second row as D and B, then the cross product ratio is (A/D)/(C/B) = AB/CD. For example: for Problem Definition, the cross product calculation (taken from Table 5a) is (1217 × 1624)/(1403 × 1167) = 1.21. When the *t*-value is above 2, this represents a statistically sig-

[3]For those interested, a tape containing all the data and calculations in the project (discussed in Part III), also contains all contingency tables.

nificant relation at 95%. It says that for questions where Problem Definition was assessed by users as above mean the odds that a retrieved item be relevant or partially relevant as opposed to not relevant is increased by a factor of 1.21 (or 21%) over the questions with below mean Problem Definition.

Column 3: The logarithm of the corresponding odds ratio taken to allow for calculation of the *t*-value.

Column 4: The standard error assuming the given value of logarithm of odds ratio.

Column 5: The *t*-value (that is, the measured value of log odds ratio divided by its standard deviation under the hypothesis of no affect) allowing for a test of statistical significance.

Column 6: Indication if the relation is statistically significant at 95% which occurs when the *t*-value is above 2.

Table 5a contains four 2 × 2 contingency tables used for calculation of relevance odds for four context characteristics in Table 5. We show here the values for both significant and not significant relations. The values in boxes indicate the number of retrieved items with given characteristics. For instance: for Problem Definition it shows (box A) that there were 1,217 items (out of 5,411 unique items retrieved by 360 searchers) that were judged not relevant by users *and* at the same time were in questions that had below mean (low) assessment of Problem Definition; there were 1,624 items (box B) that were judged relevant or partially relevant *and* were in questions with above mean high assessment of Problem Definition. As to totals, there were 2,384 (44.1%) items retrieved by searches in questions with below mean assessment on Problem Definition and 3,027 (55.1%) with above mean assessment; 2,791 (51.6%) were judged not relevant and 1,897 (51.4%) were judged relevant or partially relevant.

Note that the affect of each factor is reported independently, thus when two or more factors affect the odds they cannot be multiplied. We may assume when two factors are present the odds will increase by a combination of the odds ratios.

The results indicate that when the Problem Definition for a question (as assessed by a user) was high (above mean), the odds that a retrieved item be judged relevant increased by a factor of 1.21. In other words, with a well defined problem we may expect a slight increase in relevance of retrieved items.

Variable Intent had no significant impact. Be they well or ill defined the relevance odds remained more or less the same. However, if we included in the analysis only the items retrieved by the 200 outside searches (as we did in the Final Report [2]), and *not* the output for all 360 searches (as we did here), the variable Intent showed a slight negative relation with relevance odds, i.e., the odds slightly increased when Intent was ill-defined.

Estimates of Public Knowledge had the largest impact. An item retrieved in response to a question for which the user estimated that there is substantial public knowledge (above 3.63 on a scale of 1 to 5) was more likely to be judged relevant by a factor of 1.67 than one retrieved in response to a question on which the public knowledge was judged to lie below this cut point. In other words, higher estimates of public knowledge most significantly increased odds of relevance by 67%.

Finally, Internal Knowledge had no significant effect. Be it low or high the odds on relevance remained more or less the same.

In general, well defined problem and high estimates of existence of public knowledge increased relevance odds, while specificity of intent and the degree of internal knowledge made no difference.

What Was the Relationship between Context Variables and Precision and Recall Odds?

The relation between relevance odds and context variables as presented above involves an item-wise, or micro, level of analysis, while the relation between precision and recall odds and context variable as presented here involves a search-wise, or macro, level of analysis. At the end of Part III, we summarize both micro and macro levels of analysis together for all variables to which they were applied.

The same method of cross product ratio analysis used for calculation of relevance odds was used for calculation of precision and recall odds. We first construct a 2 × 2 contingency table in which the independent variable is broken into a number of searches of high values (i.e., above a cut point or mean) and low values (i.e., below a cut point). Each of these is further divided into searches with above mean precision (recall) and below mean precision (recall). The cross product ratio then reflects an increase in odds due to moving from a low value of the variable (or below mean) to a high value of the variable (or above mean) in relation to searches of high (above mean) precision (recall). A logarithm of the cross product ratio is taken to allow for a *t*-test of significance.

The following question was asked in this analysis: What were the odds that precision or recall for searches be above average (as opposed to below average) in questions for which a context variable was assessed as high (above mean) as opposed to questions for which it was assessed as low (below mean)?

Table 6 presents the answers, that is, it presents a summary of calculations on the relation between context characteristics of questions (as assessed by users) and the odds that precision and recall be above average. Table 6a presents the 2 × 2 contingency tables containing the values used for cross product ratio calculation of precision and recall odds in Table 6.

As for the preceding table and for the same reasons, we provide here a detailed description of Tables 6 and 6a. This description is appropriate for six other tables on precision and recall odds as related to given variables (Tables 8, 11, 16, and 18 in this Part, and Tables 21 and 24 in Part III), allowing for a briefer description of these other tables. However, for these other tables we are *not* providing the related 2 × 2 contingency tables similar to ones presented in Table 6a. We are providing the contingency tables in Table 6a just to demonstrate how the summary data in Table 6 (and all other related tables) is calculated. Table 6 is divided into two parts—one for precision odds and the other for recall odds. (Note that recall odds, wherever used, pertain only to the comparative recall of the nine searches

TABLE 6. Summary of the relation between context characteristics assessed by users and the odds that precision and recall be above average (N users = 40; N questions = 40; N all searches = 360; statistical significance at 95%; mean precision for all searches = 0.57; mean recall = 0.22).

Context characteristic	Cut point (Mean)	Odds ratio	Log odds	Stand. error +/−	t-Value	Stat. Signif.
Precision						
Problem definition	3.67	1.23	0.20	0.22	0.95	No
Intent	2.93	0.87	−0.14	0.21	−0.65	No
Public knowledge	3.63	1.87	0.62	0.22	2.88	Yes
Internal knowledge	3.48	0.82	−0.19	0.21	−0.90	No
Recall						
Problem definition	3.67	1.11	0.11	0.22	0.48	No
Intent	2.93	0.80	−0.22	0.22	−1.01	No
Public knowledge	3.63	0.76	−0.28	0.22	−1.27	No
Internal knowledge	3.48	0.99	−0.01	0.22	−0.02	No

TABLE 6a. Values for calculation of odds for the relation between context characteristics assessed by users and precision and recall. (N all searches = 360; values in boxes = no. of searches; calculation shown for the significant and one not significant relation).

		Public Knowledge Below Mean	Public Knowledge Above Mean	Total
Precision	Below Mean	81	88	169 46.9%
	Above Mean	63	128	191 53.1%
	Total	144 40%	216 60%	360 100%
		Public Knowledge Below Mean	Public Knowledge Above Mean	Total
Recall	Below Mean	81	136	217 60.3%
	Above Mean	63	80	143 39.7%
	Total	144 40%	216 60%	360 100%

for a question as related to the union of output for these nine searches; they do not relate to the absolute recall for a question.)

The columns provide the following data for each variable:

Column 1: The mean for the particular variable serves as a cut point to divide searches into those associated with above mean (or high) values of the variable and those associated with below mean (or low) values. These means are the same ones that are used as cut points for relevance odds.

Column 2: The cross product ratio of the precision and recall odds. This is calculated from the 2 × 2 contingency tables. The first row in Table 6a contains the number of searches with below mean precision (recall) and the second row the number of searches with above mean precision (recall). If the boxes in the first row are labeled A and C, and the second row D and B, then the cross product ratio is AB/CD. When the t-value is above 2 this is a significant relation at 95%. It says, for instance that for questions where existence of Public Knowledge was assessed by users as above mean, the odds that a precision of a search be high increased by a factor of 1.87 or 87%.

Columns 3, 4, and 5: These represent the corresponding logarithm of the odds ratio, standard error under the null hypothesis, and t-value.

Column 6: This says whether the relation was significant or not.

Table 6a contains the 2 × 2 contingency tables used for calculation of precision and recall odds presented in Table 6. For demonstration, we are providing only one contingency table that relates to a statistically significant relation and another that relates to a relation that was not significant. The values in boxes indicate the number of searches with given characteristics. For instance, for Public Knowledge it shows (box A) that there were 81 searches (out of 360 searches) that were associated with questions of below mean value of Public Knowledge assessment *and* at the same time had below mean precision and 128 searches (box B) in questions with above mean value of Public Knowledge and above mean precision and so on. The precision odds are then calculated as: $(81 \times 128)/(88 \times 63) = 1.87$.

Results show that the only context variable that had a significant relation to precision odds was Public Knowledge. In questions where the users estimated the existence of public knowledge as high (i.e., above mean), the odds that searches had high precision increased by a factor of 1.87 or 87%. No other variable had a significant impact on precision odds and none had an impact on recall. Macro or search-wise level of analysis was less powerful in identifying significant relations than micro or item-wise level. We shall return to a discussion of this point in the conclusions.

In general, in questions for which the users estimated that the existence of public knowledge was high, the odds increased that the precision of searches be high.

What Was the Relationship between User Constraints on Searches and Overall Precision for Questions?

Users were given a choice to indicate several constraints or restrictions to be placed on the search for their question. The following constraints have been placed on indicated number of questions (N questions = 40):

Type of search requested	broad:	2 (5%)
	precise:	36 (90%)
	not specified:	2 (5%)
Language:	English only:	25 (62%)
	Any language:	15 (38%)
Time limit for searching:	Up to last 5 years:	4 (10%)
	Up to last 15 years:	11 (28%)
	No limit on years:	25 (62%)

The "type of search" variable could not be studied because the sample in one of the categories to be compared was too small for a meaningful comparison, (i.e., the number of questions for which users requested a broad search was 5% and for which users requested a precise search was

90%). In itself, this is a comment on users: an overwhelming majority desired a precise search.

As to the language constraint, the overall precision for questions requesting items in English only was 0.56 and for those requesting any language was 0.42—a statistically significant difference. As to the time limit, the overall precision for questions restricting retrieved items to last 5 years was 0.58 and to last 15 years was 0.56; precision for questions on which no time limit was posed was 0.49—again a significant difference.

In general, users requested precise searches by a very large margin. Questions restricting answers to English had a significantly higher precision, as did questions on which users placed time limits.

What Was the Relationship between User Constraints on Searches and Relevance Odds?

The following question was asked: What were the odds that retrieved items be relevant or partially relevant in questions for which there was no restriction on searches as to language and time of publication of answers? On language the choice for users was to indicate if they desire answers reported in English or any language, and for time limit to indicate whether there should be a specified time limit on answers (e.g., published up to the last five years) or no time limit. "Any language" and "Unrestricted time limit" represent high (or above mean) values, while "English" and "Restricted" represent low values of the variables. For 25 questions (out of 40), users indicated that they desire answers in English only, and for 25 questions they had no constraints on time limit, however, these were not necessarily the same 25 questions for both constraints.

Table 7 contains the summary of relations between the two search constraints and relevance odds. Descriptions presented with Table 5 are appropriate here. The analysis involves 5,411 unique items retrieved for all 360 searches. Inclusion of all unique items in this analysis (regardless of whether they came from outside or project searches) is warranted because a constraint is associated with a question as whole, i.e., the validity is the same for all searches.

The results show that a lack of restriction on either the language or time limit had a negative relation with relevance odds. When the language was not restricted to English, relevance odds decreased by 37% (1–0.63), and when time limit was unrestricted, they decreased by 29% (1–0.71).

TABLE 7. Summary of the relation between constraints on searches placed by users and the odds that a retrieved item be relevant or partially relevant (N users = 40; N questions = 40; N all searches = 360).

Search Constraints	Cut Point	Odds Ratio	Log Odds	Std. Error +/−	t-Value
Language	Engl./ any	0.63	−0.46	0.06	−8.18
Time Limit	Rest./ unrestr.	0.71	−0.35	0.06	−6.13

Or in opposite terms: when language of answers was restricted to English, relevance odds increased by a factor of 1.59 (1/0.63). When publication was restricted to a given time period, relevance odds increased by a factor of 1.41 (1/0.71).

In general, restrictions to English and restrictions on time of publications increased the chances that a retrieved item be relevant or partially relevant, and lack of restriction decreased them.

What Was the Relationship between User Constraints on Searches and Precision and Recall Odds?

The following question was asked: What were the odds that precision and recall be above mean in questions for which there was no restriction on searches as to language and publication time of answers? The answers are presented in Table 8 containing the summary of the relations.

The results show that lack of restriction on searches either as to language or time limit had a negative effect on precision and no significant effect on recall. Odds that precision be above average declined 43% (1–0.57) for questions for which answers in any language were applicable and 50% (1–0.50) for questions for which there were no limit on time of publication. To put it in opposite terms: odds that precision be above average improved by a factor of 1.75 (1/0.57) for questions for which answers were restricted to publications in English; they doubled (1/0.50) for questions for which there was a time limit on publication of answers.

In general, restrictions of searches by users to English only and to more recent literature had a significant positive impact on retrieval of relevant items and precision of searches and no impact on comparative recall. This suggests that unrestricted searches are less effective than restricted ones.

Questions

A summary text of the 40 questions is provided in the Appendix. The full texts of written questions as submitted by users are assembled in Appendix A of ref. 2 (*Final Report*). The questions were used in two ways: (i) for searching and (ii) in a separate classification experiment.

This experiment involved 21 judges (separate from searchers) doing two things: (i) assessing the structure of the questions and (ii) categorizing them (or describing their characteristics) according to a classification scheme developed in the project. Each judge judged 20 questions, or in other words, each of the 40 questions was judged by about 10 different judges.

As to structure, the question was postulated to have three parts: lead-in, query, and subject. The subject is the main concept(s) in the question and the query represents the question(s) asked about the subject. The lead-in, while not directly searchable, may give clues as to presuppositions. Consistency of assessment of question structure is being analyzed by using linguistic methods and the results are not reported here. Results on consistency of classification be-

TABLE 8. Summary of the relation between constraints on searches placed by users and the odds that precision and recall be above average (N questions = 40; N all searches = 360; statistical significance at 95%; mean precision for all searches = 0.57; mean recall = 0.22)

Search Constraint	Cut Point	Odds Ratio	Log Odds	Stand. Error +/−	t-Value	Stat. Signif.
Precision						
Language	Engl./ any	0.57	−0.55	0.22	−2.53	Yes
Time limit	Restr./ Unrestr.	0.50	−0.70	0.22	−3.12	Yes
Recall						
Language	Engl./ any	0.97	−0.03	0.22	−0.14	No
Time limit	Restr./ Unrestr.	0.93	−0.07	0.22	−0.31	No

tween judges and the effects of various question classes on retrieval are reported below.

The classification scheme consisted of five categories:

(1) *Domain:* Subject area of the question, measured by indicating in which DIALOG category does the question belong.
(2) *Clarity:* Semantic and syntactic, each measured by a scale from 1 (unclear) to 5 (clear), i.e., clarity involves two variables.
(3) *Specificity:* Of the query part and the subject part of the question, each measured by a scale from 1 (broad) to 5 (narrow); (two variables).
(4) *Complexity:* Measured in two ways: on a scale from 1 (low complexity) to 5 (high complexity) and by indicating the number of search concepts; (two types of measures).
(5) *Presupposition:* Presence of implied (not explicitly stated) concepts, measured in two ways: on a scale from 1 (nothing implied) to 5 (many implications) and by indicating the number of presupposition that can be expressed as search terms; (two types of measures).

What Was the Consistency in Assignment of Question Classification?

With this as with any other classification scheme, we seek assurance that the score assigned to a question is really a property of the question and not a property of the judge. We assigned 21 judges to score each of 20 questions, so that each of the 40 questions was scored approximately 10 times. We then asked: for each scale and for each question do the ten scores that have been assigned seem to be related or random? The measure of relatedness is the variance. If all judges exactly agree, the variance is zero. The null hypothesis is that the scores assigned by the judges are chosen at random from the numbers one through five.

To determine a 95% confidence limit, we generated 10,000 quasi-random sets of ten scores each to simulate the 10 judges of a question and determined the distribution of the 10,000 variances of those sets of scores. We found that there is less than a 10% chance that the variance will be less than one. We then asked: if judges were assigning scores randomly, what is the chance that N out of 40 questions would all show variance less than one? For $N = 7$ there is less than 10% chance that this will happen. For $N = 9$ there is less than 5% chance. Based on this test, we find that inter-judge agreement is significant at the 90% confidence level for clarity of semantics and clarity of syntax. Inter-judge agreement is significant at the 95% confidence level for specificity of subject, complexity and the presence of presuppositions. Only specificity of query fails this test, with $N = 6$.

This analysis is further supported by an analysis of variance which shows that, although there are large variations between judges, there are also significant differences between questions. We conclude that, with the possible exception of specificity of query, the measures in the question classification scheme represent reproducible properties of the questions themselves; in other words, we are confident (within described limits) that the classification scheme is a valid one.

For each question and each characteristic of that question, we took the mean of the 10 scores assigned by the judges. Table 9 presents the mean of these means for the 40 questions on all characteristics. This enabled us to establish a cut point for questions that are above or below the mean of means on the given characteristic for use in calculation of relevance, precision, and recall odds. Furthermore, we calculated as shown the mean variance for each characteristic over the 40 questions. We presume that when the variance is higher the judges had a harder time assigning that score or that the associated characteristic was more difficult to apply. Note that for two characteristics (number of search concepts in Complexity and number of presupposed concepts in Presupposition), the score was indicated as an actual number chosen by each judge and not restricted to a scale from 1–5. Thus, for these two the variance has to be higher and it is not comparable to variances of characteristics indicated on a scale.

TABLE 9. Classification of questions: mean judgements on different characteristics (classes) over all questions (N questions = 40; N judges = 21; N questions classified by each judge = 20).

Question Characteristic	Mean of Means	Stand. Dev. Means	Mean of Variance	Minimum for a Question	Maximum for a Question
Clarity					
of semantics (1–5)	3.44	0.56	1.19	1.80	4.67
of syntax (1–5)	3.57	0.40	1.28	2.50	4.33
Specificity					
of query part (1–5)	2.70	0.68	1.25	1.20	4.50
of subject part (1–5)	3.47	0.77	1.20	1.70	4.67
Complexity					
indic. on scale (1–5)	3.31	0.58	1.13	1.90	4.55
by no. of concepts	5.64	1.49	3.00	3.50	9.43
Presupposition					
indic. on scale (1–5)	2.68	0.54	1.19	1.63	3.75
by no. of presup.	2.82	1.08	2.50	1.17	5.67

The variances did not differ very much. The highest disparity of judgment was on clarity of syntax and specificity of the query part of the question. In assessment by number of concepts, the variance was higher in respect to complexity than in respect to presupposition. In general, this may indicate that searchers (as represented by our judges) have a harder time in distinguishing between the query and subject part of the question. Among themselves, they also see specificity somewhat differently and assess to some extent a different number of search concepts as present in the same question.

Let us interpret these findings on a more general level. As many other researchers have found, when judges are assigned a conceptual task there is a substantial "declustering" that takes place. Although none of the judges report difficulty in understanding or dealing with the concepts involved in question classification, it is clear that when they tried to apply those concepts to specific questions significant and substantial disagreement can occur. Similar phenomena have been found in indexing and abstracting, and are found in the analysis of overlap in search terms and items retrieved in this project (see Part III). We interpret it to mean that no matter how clear a concept may be in the abstract, when it is processed by a particular human intelligence it is transformed into a specific representation. The set of specific representations form a cluster which must be supposed to have the concept at its center. We believe the study of this "declustering" process may provide an important tool for understanding which characteristics are essential and which characteristics are accidental in the individual representations. We will refer to this again in the analysis of overlap results.

What Was the Relationship between Question Classes and Relevance Odds?

While there were five question classes, relevance odds and precision and recall odds could be calculated only for four of them. Classification on domain had to be left out, because it had more than two values; these could not be broken into high and low values as required by the method. For the remaining four classes, the relevance odds (and precision and recall odds) were calculated separately for two dimensions of each class involving *clarity*–semantic and syntactic clarity; *specificity*–of the query part and of the subject part of the question; *complexity*–indicated on a scale from 1 to 5 and by number of concepts present in a question; and *presupposition*–indicated on a scale from 1 to 5 and by a number of concepts presupposed or implied in a question. The odds tables are organized accordingly.

Table 10 is a summary of the relation between the question characteristics as expressed by assessments of classification judges and the odds that a retrieved item be relevant or partially relevant as opposed to not relevant. For each characteristic the questions were broken into two classes: those questions assessed to have high or above mean values and those that had low or below mean values. The description presented with Table 5 is applicable.

The following question was asked: What were the odds that the retrieved items be relevant or partially relevant as opposed to not relevant in questions for which a given characteristic was judged as high (above mean) as opposed to questions for which it was judged low?

The analysis includes 200 outside searches only, because the outside searchers performed the searches on the basis of the written question only and the classification judges assessed the question characteristics on the basis of the same source. Project searches were done on the basis of the written question and an additional source (i.e., problem statement) as well, thus they were not appropriate for this particular analysis.

As can be seen on the characteristic of clarity, relevance odds in question with high semantic clarity decreased by 21% (1–0.79) and with high syntactic clarity decreased by 26%. Expressed in opposite terms: relevance odds in questions with low semantic clarity were enhanced by a factor of 1.27 (1/0.79) and in those with low syntactic clarity by a factor of 1.35 (1/0.74). Thus, clarity of the questions, be it semantic or syntactic, had a significant impact on relevance

TABLE 10. Summary of the relation between question characteristics as assessed by judges and the odds that a retrieved item be relevant or partially relevant (*N* questions = 40; *N* judges = 21; *N* outside searches = 200; statistical significance at 95%).

Question	Cut Point (Mean)	Odds Ratio	Log Odds	Stand. Error +/−	*t*-Value	Stat. Signif.
Clarity						
of semantics (1–5)	3.44	0.79	−0.24	0.07	−3.60	Yes
of syntax (1–5)	3.57	0.74	−0.31	0.07	−4.61	Yes
Specificity						
of query part (1–5)	2.70	0.97	−0.03	0.07	−0.40	No
of subject part (1–5)	3.47	0.55	−0.60	0.07	−8.75	Yes
Complexity						
indic. on scale (1–5)	3.31	1.71	0.54	0.07	8.02	Yes
by no. of concepts	5.64	1.93	0.66	0.07	9.85	Yes
Presupposition						
indic. on scale (1–5)	2.68	1.12	0.12	0.07	1.76	No
by no. of presup.	2.82	1.45	0.37	0.07	5.53	Yes

odds: high clarity a negative one, and low clarity a positive one.

On the characteristic of specificity, the specificity of the query part of the questions had no significant relation with relevance odds, while the specificity of the subject part was quite significant. For a narrow question (i.e., with high specificity of the subject) the odds that an item be relevant or partially relevant decreased by 45% (1 − 0.55) or in opposite terms, for broad questions (i.e., with low specificity of the subject) relevance odds increased by a factor of 1.82 (1/0.55). The different picture on relevance odds between the query and subject specificity is connected with significantly different assessments (as expressed by respective means) on the level of their specificity. As suggested, it may be that judges have difficulty in distinguishing between the two parts of the question.

Complexity of questions assessed by either indication on a scale or as to number of concepts present, was quite significant. For questions of high complexity as indicated on a scale (or as to number of concepts) the odds that a retrieved item be relevant or partially relevant increased by a factor of 1.71 (1.93).

The characteristic of presupposition presented dissimilar results. When judges assessed presupposed or implied concepts in the questions on a scale, the relation to relevance odds was not significant, but when they indicated presupposition as to the number of implied concepts, the relation was significant. For questions with a high number of implied concepts relevance odds increased by a factor of 1.45. The difficulty in assessing presuppositions and the type of measure for recording of assessments may have been the contributing factor in this disparity.

In general, relevance odds were enhanced for questions of low clarity, low specificity, high complexity, and with a high number of presupposition. Or in other words, when questions were either not very clear, broad, complex, and/or left a lot implied, the odds that retrieved items be relevant increased. Intuitively this is to be expected for clarity and specificity, but not for complexity and presupposition. We can see that "unclear" and "broad" may be related and in both cases more items are acceptable for answers. In other words, for very clear and/or narrow questions relevance judgments by users may be more strict and thus producing lower relevance odds. However, we cannot easily explain why should complex questions (those with a higher number of concepts present) or questions with a higher number of presuppositions have better relevance odds. Is it because complexity and specificity are closely related as different sides of the same coin? Are these results applicable only to this experiment? Are complexity and presuppositions in the minds of judges seen as the same characteristic? These and similar speculations about whys on this and other variables are just that—speculations. Explanations require further research.

What Was the Relationship between Question Classes and Precision and Recall Odds?

The following question was asked in this analysis: What were the odds that questions with above mean assessment on any of the question characteristics had searches with above mean precision or recall? Table 11 provides the answers. The description presented with Table 6 is appropriate.

As can be seen, the only characteristics that were significantly related to precision odds were specificity of the subject part of the question and complexity as indicated by either the scale or the number of concepts. Not a single characteristic was significantly related to comparative recall of searches.

As to characteristics of specificity of the subject part of the question the relation is negative. For questions that are assessed as narrow (high specificity) the odds that a search had above mean precision declined 53% (1–0.47). Put in opposite terms, for broad questions (i.e., for questions with low specificity) the odds that precision of searches be above mean increased 2.14 times (1/0.47).

As to characteristic of complexity, for questions of high complexity indicated on a scale (or as to number of terms),

TABLE 11. Summary of the relation between question characteristics as assessed by judges and the odds that precision and recall be above average (N questions = 40; N judges = 21; N outside searches = 200; statistical significance at 95%; mean precision for outside searches = 0.54; mean recall = 0.20).

Question Characteristic	Cut Point (Mean)	Odds Ratio	Log Odds	Stand. Error +/−	t-Value	Stat. Signif.
Precision						
Clarity						
of semantics (1–5)	3.44	0.75	−0.29	0.29	−1.00	No
of syntax (1–5)	3.57	0.96	−0.04	0.29	−0.14	No
Specificity						
of query part (1–5)	2.70	1.00	0.00	0.28	0.00	No
of subj. part (1–5)	3.47	0.47	−0.76	0.29	−2.58	Yes
Complexity						
indic. on scale (1–5)	3.31	2.27	0.82	0.28	2.82	Yes
by no. of concepts	5.64	2.16	0.77	0.28	2.67	Yes
Presupposition						
indic. on scale (1–5)	2.68	0.85	−0.16	0.28	−0.57	No
by no. of presup.	2.82	1.65	0.50	0.29	1.73	No
Recall						
Clarity						
of semantics (1–5)	3.44	1.29	0.25	0.30	0.85	No
of syntax (1–5)	3.57	1.41	0.34	0.30	1.14	No
Specificity						
of query part (1–5)	2.70	1.11	0.11	0.29	0.37	No
of subj. part (1–5)	3.47	1.57	0.45	0.30	1.49	No
Complexity						
indic. on scale (1–5)	3.31	0.64	−0.45	0.29	−1.54	No
by no. of concepts	5.64	0.82	−0.20	0.29	−0.69	No
Presupposition						
indic. on scale (1–5)	2.68	0.55	−0.59	0.29	−1.97	No
by no. of presup.	2.82	0.91	−0.09	0.30	−0.30	No

the odds that precision of searches be above the mean increased by a factor of 2.27 (2.16).

To generalize, questions with low specificity and high complexity have twice the odds that precision of searches be high. High specificity and high complexity of questions are highly related to odds for precision, but each in an opposite way. As discussed for relevance odds, this may be intuitively clear for specificity but not for complexity.

We see here again that the item-wise or micro level of analysis provided a sharper probe for relations than the search-wise or macro level of analysis. (The relation between these two levels of analysis is further discussed in conclusions presented at the end of Part III.) As expected, when both levels indicate a significant relation with odds on the same characteristic, then both point in the same direction. Precision odds more often show a significant relation than comparative recall odds, a fact that will be discussed later.

The findings suggest that no matter what other variables are involved, the characteristics of questions by themselves (as specified here) may have a significant impact on the outcome of searching. It seems that all questions are not created equal in qualities other than their subject.

Precision and Recall

What Were the Figures for Precision and Recall?

As mentioned, precision is defined for questions and for searches:

(1) Precision for a *question:* fraction of relevant or partially relevant ($R + pR$) items in relation to all items submitted to the user i.e. in relation to the union output of 9 searches for a question.
(2) Precision for a *search:* fraction of relevant items in a given search in relation to all items retrieved by that search.

Recall was calculated only as a *comparative* measure for searches of the same question, but not for a question as a whole. It is a fraction of relevant or partially relevant ($R + pR$) items in a search in relation to all $R + pR$ items in the union of all 9 searches for a question. An overall recall for a question cannot be established because we do not know what relevant items were left unretrieved in the file.

Table 12 provides the mean, standard deviation and range of overall precision for questions and precision and

TABLE 12. Precision and recall for questions and searches (N questions = 40; N all searches = 360; N outside searches = 200; N project searches = 160).

	Precision				Recall			
	Mean	Stand. Dev.	Min.	Max.	Mean	Stand. Dev.	Min.	Max.
Overall for 40 question	0.51	0.24	0.1	1.0		Not applicable		
All (outside + project) searches	0.57	0.34	0.0	1.0	0.22	0.21	0.0	0.90
Outside searches	0.54	0.34	0.0	1.0	0.20	0.20	0.0	0.81
Project searches	0.61	0.32	0.0	1.0	0.25	0.23	0.0	0.90

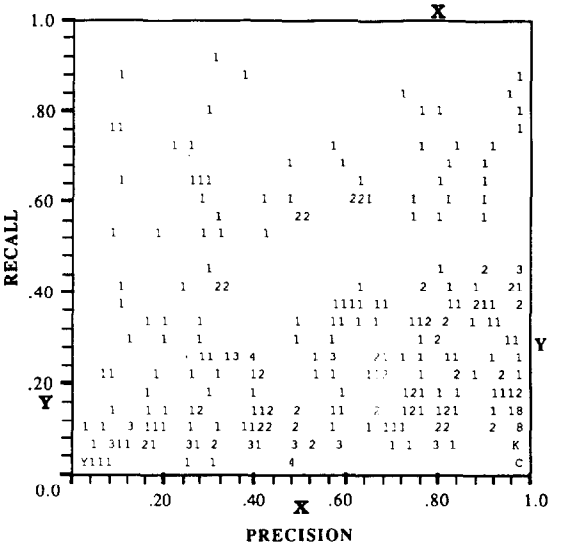
recall for all, outside and project searches. The mean number of items retrieved per question and per search is given in Table 2. A comparison of results between outside and project searches is discussed in greater detail in Part III of this article in the section "Project Searches."

For questions the mean overall precision was 51%. For all searches precision was 57% and recall was 22%; for outside searches precision was 54% and recall was 20%; while for project searches precision was 61% and recall 25%. The range for both was very wide. The distribution is not normal — the results tended to accumulate toward the end points of scales, for instance, for search precision the result accumulated at the high end and for recall at the low end of the scale. Our precision and recall results correspond to precision and recall performance figures found in many studies summarized in [3], however, without correlation with other variables precision and recall figures by themselves have as little meaning as any other single, unrelated statistic.

What Was the Correlation Between Precision and Recall?

Table 13 provides a scatter plot of recall vs. precision for all 360 searches. For each search recall and precision are plotted against each other as one point, resulting in 360 points plotted in the graph. Two linear regression lines are plotted: (i) for precision as the independent and recall as the dependent variable (connecting Y to Y on vertical lines) and (ii) for recall as the independent and precision as the dependent variable (connect X to X on the horizontal lines).

TABLE 13. Plot for precision and recall for all searches (N searches = 360; entries in the plot: 1 = one search with the given precision and recall value; 2 = two searches with the given value and so on until 9; after 9, A = 10 searches with the given value, B = 11 searches with the given value and so on. For regression lines connect X and X on the ordinates and Y and Y on the abscissas).



	MEAN	ST. DEV.	Regression Line	Res. Ms.
Precision (X)	= 0.572	0.335	$X = 0.249 \cdot Y + 0.517$	0.109
Recall (Y)	= 0.219	0.213	$Y = 0.101 \cdot X + 0.162$	0.004

The results are quite remarkable. The correlation between recall and precision was about 16%. An enormous amount of scatter is shown. It is often said that there is an inverse relation between recall and precision. In real search situations it has never been clear that this relationship should exist, a point that was raised even by Cyril Cleverdon already in 1972 [4]. With our data the opposite was observed. No matter which variable is chosen as independent, there was a weak but positive relationship between recall and precision; as one rose slowly so did the other. In our 360 searches those searches with higher recall tended to have higher precision and vice versa.

A study of relations between precision and recall was *not* one of our objectives and we did not study why the observed relations did occur in our data. Nevertheless, it is interesting and unavoidable to speculate why our observations are contrary to many other observations and to what has become one of the accepted and taught "principles" of searching. However, this should be a question for further research, rather than speculation.

What Variables Explain Precision and Recall?

The key variables of precision and recall are both bounded by 0 and 1 (that is they are percentages). In this situation it is sometimes useful to perform the so called logistic transformation. Each variable is replaced by the logarithm of the corresponding odds ratio. For example, the value 40% is transformed to logarithm of 40/60. We have performed multiple regression analysis of the transformed values of precision and recall against nine variables: four cognitive variables, the users' estimate on the existence of public knowledge, the searchers' frequency of using DIALOG, and the number of terms, commands, and cycles used in the search.

For precision the most important explanatory variables were: (i) the users' estimate of the probability of existence of public knowledge on the subject of their question, which explained about 10% of the variance, and (ii) the Remote Associates Test score of the searchers (test of word association) which explained about an additional 5% of the variance. None of the other possible variables passed the F test for entering the regression. Thus, altogether these two variables explained about 15% of the observed variation in precision. As in other cases with a low R-squared value, we must conclude that the bulk of effect on precision (about 85%) was not explained by variables included in this analysis.

The situation for explaining recall is substantially worse. Only one variable entered the regression, the combination score AC-CE (Abstract Conceptualization minus Concrete Experience) on the Learning Style Inventory (this score indicates the extent to which an individual emphasis abstractness over concreteness as learning style). It explained somewhat less than 5% of the observed variation in recall. The bulk of effect on recall (about 95%) was not explained by variables included in this analysis.

A large number of additional regression analyses (reported in ref. 2) were performed, involving altogether 344 combinations of variables, that is, involving every meaningful combination of variables in the study. (The above regression involved $9 \times 2 = 18$ combinations.) The results were disappointing in that no significant explanations were found on this search-wise level of analysis.

These negative regression results invite some speculatory explanations. Three different points can be raised. First, we cannot, of course, exclude the possibility that the bulk of the variation in both precision and recall was due to essentially random factors highly specific to users, questions and/or the searchers. Second, the effects which may be observed on micro- or item-wise levels (using relevance judgment of each item to calculate relevance odds) are not strong enough to predict the values of precision or recall at the macro or search-wise level using regression analysis (which in turn in itself has definite and considerable limitations as an analysis tool). Third, the measures of precision and recall themselves may not be the most sensitive and thus appropriate measures. In themselves they may need a reexamination as to what they are showing and what they can show; this particularly applies to recall. After all they are macro measures with all the ensuing limitations of all macro measures. In any case, if the analysis was restricted to regression on a search-wise level, not much would have been learned.

Utility

What Were the Figures for Utility Measures?

Precision and recall were based on relevance judgment on each item evaluated by users. Measures of utility on the other hand, indicate a user's assessment of all items provided together in the response to a question. In other words, measures of utility are a judgment on the totality of items provided, rather than judgment on each answer separately, providing grounds for separate analyses and comparison between the two.

The number of users which assigned given values to each of the five utility measures is given in Table 14. As can be seen

- 70% of the users considered their participation in the project and the information that resulted as worth "much more" or "somewhat more" than the time it took; 20% said it was worth "about the same" as the time it took, and 10% said it was worth "less" than the time it took;
- 45% of the users could not assign a dollar value to the information provided; 28% assigned less than $50; 20% assigned between $50 and $200; and 7% assigned over $200;
- 25% of the users spent less than an hour on evaluation; 30% spent between one and two hours; and 45% spent more than two hours;
- 45% of the users scored the contribution made by the information supplied to the resolution of their problem as high (upper two points); about 23% were in the middle and 32% were in the lower two points on the scale; actually, only two users (5%) said "nothing" was contributed;
- 60% of the users scored their satisfaction with the results of the search high (upper two points); 20% scored in the

TABLE 14. Utility measures: distribution of user assignments to each measure (N users = 40).

Worth Scale

Was your participation in this project and the information which resulted:

	No. of Users
5 Worth much more than the time it took	16
4 Worth somewhat more than the time it took	12
3 Worth about as much as the time it took	8
2 Worth less than the time it took	4
1 Practically worthless	0

User's Time

How much time did you spend reviewing these items?

	No. of Users
Less than 1 hour	10
1–2 hours	12
2–3 hours	8
3–4 hours	7
Over 4 hours	3

Dollar Value Assigned

What is the dollar value of these items?

	No. of Users
I cannot assign a a dollar value	18
Less than $50	12
$50–$100	3
$100–$200	4
Over $200	3

Problem Resolution Scale

What contribution has this information made toward the resolution of your research problem:

	No. of Users
5 Substantial contribution	3
4	15
3	9
2	11
1 Nothing contributed	2

Satisfaction Scale

How satisfied were you with results of the search:

	No. of Users
5 Satisfied	11
4	13
3	8
2	6
1 Dissatisfied	2

middle; and 20% scored on the lowest two points. Actually, only two users (5%) said they were "dissatisfied."

It is on interest to note that the problem resolution scores did not parallel the satisfaction scores. Six (15%) more users scored satisfaction high than scored problem resolution high (upper two scores), and five (12%) more users scored problem resolution low than scored satisfaction low (lower two scores). This may show that users made a distinction between the two concepts as measures and/or that they interpreted the scales differently.

What was the Relationship between Utility Measures and Relevance Odds?

The following question was asked: What were the odds that retrieved items be relevant or partially relevant (as opposed to not relevant) in questions for which the utility was assessed by users as high (above mean) as opposed to those for which it was assessed low? The answers are presented in Table 15 which contains the summary of the relation between the five utility measures and the odds that a retrieved item is relevant or partially relevant. The description

TABLE 15. Summary of the relation between utility measures as assigned by users and the odds that a retrieved item be relevant or partially relevant (N users = 40; N questions = 40; N types of utility measures = 5; N all searches = 360; N unique items retrieved for all questions = 5411; statistical significance at 95%).

Utility Measure	Cut Point (Mean)	Odds Ratio	Log Odds	Stand. Error +/−	t-Value	Stat. Signif.
Worth Scale (1–5)	4.0	1.42	0.35	0.06	5.80	Yes
Evaluation Time (hrs)	2.49	0.88	−0.13	0.05	−2.40	Yes
Dollar Value ($)	75.25	1.34	0.29	0.06	4.52	Yes
Problem Resolution Scale (1–5)	3.15	1.88	0.63	0.05	11.43	Yes
Satisfaction Scale (1–5)	3.63	1.92	0.65	0.05	11.64	Yes

presented with Table 6 is appropriate here. The analysis involves 5,411 retrieved items, that is, unique items retrieved by all 360 searches. Inclusion of output for both, outside and project searches is warranted here because we are relating the items to users' assessment of utility, and the users did not care by what searches the items were retrieved.

All five utility measures had a significant impact on relevance odds. When the worth of users' time spent in relation to outcome was assessed as high (above mean), the odds that an item be relevant or partially relevant were higher by a factor of 1.42. Time spent on evaluation had a small negative effect: when the time was long (above mean) relevance odds were lower by 12%. When dollar value was considered high, relevance odds were higher by a factor of 1.34. When contribution to resolution of the problem was assessed as high, relevance odds increased by a factor of 1.88. Finally, when the satisfaction was high, relevance odds almost doubled (increased 1.92 times).

In general, relevance odds for questions were higher when users assessed the answers as being highly worth their time, having a high dollar value for them, contributing a lot to problem resolution, and when they were highly satisfied with the whole thing. They were lower when users took a lot of time to evaluate the answers.

For the most part, high marks on utility and higher relevance odds coincide to a great degree, however, this cannot be interpreted as one causing the other.

What was the Relationship between Utility Measures and Precision and Recall Odds?

User judgments were responsible for both utility measures and relevance indications used for calculation of precision and recall. However, the users were asked to use different criteria in judgment of each. For the former the criterion used was overall usefulness (value, utility of the provided output) and for the latter the relevance of individual items. We did *not* raise the question which one of these is the "proper" or "more significant" criteria for evaluation (a question of long standing debate in information science). Instead, we investigated the relation between the two sets of ensuing measures. This was done by using the cross product ratio method of analysis. The following question was asked: What were the odds that precision or recall be above mean in cases when utility measures were above mean?

Table 16 presents the summary of the relation between the five utility measures and the odds that precision and recall be above average. The description associated with Table 6 is applicable here. The analysis involves all 360 searches, that is, both the outside and project searches, for the same reasons as explained above in connection with utility measures and relevance odds.

The results indicate that

- Searches in questions for which users indicated that their participation was worth somewhat or much more time than it took were 2.4 times more likely to have high (above mean) precision;
- The time it took users to review the answers had no relation to precision, however, this was the only variable that had relation to recall and the relation was negative: when users took longer than average to evaluate the items retrieved for their question, the searches were 39% (1 − 0.61) less likely to have high recall, or in opposite terms, when they took less time, recall odds increased by a factor of 1.64 (1/0.61). Actually, it may be more appropriate to say that when recall was low the evaluation of retrieved set took longer;
- Searches in question with results that were assigned by users a higher (above mean) value in dollars, (in our case

TABLE 16. Summary of the relation between utility measures as assigned by users and the odds that precision and recall be above average (N users = 40; N questions = 40; N types of utility measures = 5; N total searches = 360; statistical significance at 95%; mean precision for all searches = 0.57; mean recall for all searches = 0.22).

Utility Measure	Cut Point (Mean)	Odds Ratio	Log Odds	Stand. Error +/−	t-Value	Stat. Signif.
Precision						
Worth Scale (1–5)	4.0	2.40	0.88	0.23	3.71	Yes
Evaluation						
Time (hrs)	2.49	0.96	−0.04	0.21	−0.20	No
Dollar Value ($)	75.25	1.69	0.52	0.25	2.02	Yes
Problem Resolution						
Scale (1–5)	3.15	3.21	1.17	0.22	5.24	Yes
Satis. Sc. (1–5)	3.63	2.49	0.91	0.22	4.14	Yes
Recall						
Worth Scale (1–5)	4.0	0.76	−0.28	0.23	−1.20	No
Evaluation						
Time (hrs)	2.49	0.61	−0.49	0.22	−2.23	Yes
Dollar Value ($)	75.25	0.70	−0.35	0.26	−1.33	No
Problem Resolution						
Scale (1–5)	3.15	0.98	−0.02	0.22	−0.08	No
Satis. Sc. (1–5)	3.63	0.92	−0.09	0.22	−0.40	No

above $75.25) were 69% more likely to have higher precision;
- Searches in questions where users indicated that the contribution to their problem resolution was high (above mean) were a dramatic 3.21 times more likely to have higher precision;
- Searches in questions where users indicated high overall satisfaction with results of the whole exercise were 2.49 times more likely to have higher precision.

In general, retrieved sets with high precision increased the chance that users assessed that the results were "worth more of their time than it took," were "high in dollar value," contributed "considerably to their problem resolution," and "were highly satisfactory." On the other hand, high recall did not significantly affect the odds for any one of those measures. But, low recall increased the odds that the users took a long time to evaluate the results.

The generalization could be expressed the other way as well: When users were satisfied and valued the results highly the chances that the searches had high precision increased. When users took a long time to evaluate the results, the chances that the associated searches had low recall increased.

These are interesting findings in another respect. They indicate that utility of results (or user satisfaction) may be associated with high precision, while recall does not play a role that is even closely as significant. For users, precision seems to be the king and they indicated so in the type of searches desired. In a way this points out to the elusive nature of recall: this measure is based on the assumption that something may be missing. Users cannot tell what is missing any more than searchers or systems can. However, users can certainly tell what is in their hands, and how much is *not* relevant.

Items Retrieved: Output Size

The output size here refers to the following six quantities or variables: (i) number of relevant items retrieved; (ii) number of partially relevant items retrieved; (iii) number of not relevant items retrieved; (iv) total number of evaluated items retrieved ($R + pR + N$); (v) number of items not evaluated; and (vi) total number of retrieved items (evaluated + not evaluated).[4]

We analyzed the relations between the size of output variables per question and odds of relevance, precision, and recall. Clearly it is to be expected that some of the relations be high, e.g., it is expected that for questions with a high number of items judged relevant or partially relevant the precision odds will be high. Some of the relations to relevance, precision and/or recall odds seem to be intuitively clear to the point that such analysis is confirmation of the obvious. Nevertheless, intuitions need to be confirmed. Moreover, even if some of the relations are intuitively clear, others may not be. As it turned out, not all of our own intuitions were confirmed and some findings were even counterintuitive.

What Was the Relationship between Output Size and Relevance Odds?

The following question was asked: What were the odds that items retrieved for questions with high (above mean) values of given size of output were relevant as opposed to not relevant? Table 17 provides a summary of the relation between six variables included under size of output and relevance odds. The calculation involves 5411 unique items contributed by all 360 searches. The cut points are the mean number of items retrieved per question and not per search. The description provided for Table 5 is appropriate.

All six sizes of output variables had a significant relation with relevance odds and all but one were positive. For questions with a high (above average) number of relevant, partially relevant, and total number of evaluated items, the

[4]Note in explanation of items (v) and (vi): as mentioned, for questions exceeding 150 items in total unique retrievals, only the first 150 items were sent to users, thus for a number of questions there were more items retrieved than evaluated. See section "Overall Retrieval" for exact quantities.

TABLE 17. Summary of the relation between size of the output for questions and the odds that a retrieved item be relevant or partially relevant (N questions = 40; N searches = 360; N unique, items retrieved = 5411; statistical significance at 95%).

Size of Output Characteristic Per Question	Cut Point (Mean)	Odds Ratio	Log Odds	Stand. Error +/−	*t*-Value	Stat. Signif.
No. of relevant items	33.58	2.95	1.08	0.05	19.11	Yes
No. of partially relevant items	36.20	2.43	0.89	0.05	15.95	Yes
No. of not relevant items	65.50	0.22	−1.52	0.05	−25.87	Yes
Total no. of evaluated items	135.28	1.21	0.19	0.08	2.46	Yes
No. of items NOT evaluated	159.83	1.39	0.33	0.05	6.03	Yes
Total no. of retrieved items	295.10	1.39	0.33	0.05	6.03	Yes

respective relevance odds increased 2.95 times, 2.43 times, and 21%, while with a high number of not relevant items they decreased 78% (1–0.22). In questions with a high (above mean) number of items retrieved, not evaluated, (i.e., not sent to users), relevance odds increased 39%.

To generalize, as expected, relevance odds increased for questions with a high number of relevant items and decreased in questions with a high number of not relevant ones. Questions with high or an above average number of items submitted to users for evaluations and/or with a high number of items not submitted, that is, questions with high total retrievals, had a significant (albeit not a very large) impact on relevance odds. In other words, in questions with a larger number of retrievals, the likelihood that an item is relevant improved, and in a question with small number of retrievals, it declined.

This suggests that the total size of output for a question, (in this study), i.e., the total number of items retrieved as answers, has a positive relation with relevance odds. It would be of interest to explore at what point in the increase in the size of output (e.g., by broader and broader searches) will this relation change to a negative one, as it clearly must.

What Was the Relationship between Output Size and Precision and Recall Odds?

The following question was asked: What were the odds that searches in questions with high (above mean) of given size of output had above mean precision or recall? Table 18 is a summary of the relations between output size for questions and the odds that precision and recall of contributing searches was above mean. As in relevance odds, the cut points for various output sizes are the mean numbers retrieved for the questions and not for the searches (description with Table 6 is appropriate).

For precision, three variables were significant, as expected. The odds that a search had high precision increased 4.43 times in questions with high numbers of relevant and 2.9 times with high numbers of partially relevant items, and decreased 83% (1–0.17) in questions with a high number of not relevant items. The size of output per question, be it evaluated or not, had no significant relation with precision.

For recall, we could study only comparative recall for searches with questions of high and low output, but not questions as a whole. For such recall, three variables were

TABLE 18. Summary of the relation between size of the output for questions and the odds that precision and recall of searches be above average (N questions = 40; N searches = 360; statistical significance at 95%; mean precision for all searches = 0.57; mean recall for all searches = 0.22).

Size of Output Characteristics per Question	Cut Point (Mean)	Odds Ratio	Log Odds	Stand. Error +/−	t-Value	Stat. Signif.
Precision						
No. of relevant items	33.58	4.43	1.49	0.21	6.41	Yes
No. of partially relevant items	36.20	2.90	1.07	0.21	4.86	Yes
No. of not relevant items	65.50	0.17	−1.80	0.21	−7.71	Yes
Total no. of evaluated items	135.27	1.33	0.28	0.24	1.16	No
No. of items NOT evaluated	159.83	1.18	0.17	0.22	0.78	No
Total no. of retrieved items	295.10	1.18	0.17	0.22	0.78	No
Recall						
No. of relevant items	33.58	0.83	−0.18	0.22	−0.82	No
No. of partially relevant items	36.20	0.63	−0.47	0.22	−2.14	Yes
No. of not relevant items	65.50	0.72	−0.32	0.22	−1.49	No
Total no. of evaluated items	135.27	0.64	−0.44	0.25	−1.80	No
No. of items NOT evaluated	159.83	0.58	−0.55	0.22	−2.45	Yes
Total no. of retrieved items	295.10	0.58	−0.55	0.22	−2.45	Yes

significant, and all three were negatively correlated; searches in questions with above mean numbers of partially relevant, not evaluated, and total retrieved items were 37% (1–0.63) (42%, 42%) less likely to have high (above mean) recall, or to put it in opposite terms: searches in questions with low number of partially relevant, not evaluated and total retrieved items, were 58% (1/0.63) (81%, 81%) more likely to have high recall.

In general, as expected, precision odds were positively affected by a high number of relevant and partially relevant items in a question, and negatively by not relevant ones. Recall odds were negatively affected for questions with a high number of partially relevant items—this is unexpected; it seems that the question for which the users found a lot of marginally relevant answers have a pronounced negative effect on recall. Even more unexpectedly, the number of relevant items had no effect on recall odds of searches one way or another. Precision odds were dependent on a high number of relevant items for a question, but recall odds were not.

Surprisingly, the size of total output, be it small or large, evaluated or not, had no effect on precision, nor did the number of items in a question submitted to users for evaluation have any effect on recall. What was left unevaluated had a negative effect on recall; however, we cannot make any interpretation of this precisely because the items were not evaluated.

This suggests that precision may be governed by some aspects of size of output that are quite obvious (number of items in a question judged as relevant, partially relevant, or not relevant) and not affected by others that are not so obvious, particularly the total number of items submitted to users. At the same time recall seems to be impervious to the same factors, with the exception of items judged partially or marginally relevant—these affect recall in a negative way. Precision and recall were not affected the same way by any aspects of the output size provided to users.

As with other variables, the macro or search-wise level of analysis was not as sharp as the micro or item-wise level. Nevertheless, we still gained some non-trivial insight on how the output size affects precision and recall.

General conclusions from results of all the variables presented here and in the next part are stated at the end of Part III, the last part in this series of articles about a study on information seeking and retrieving.

Appendix: Questions–Summary of the Subject, Database Searched and Number of Items Retrieved

(R = no. of items judged relevant; pR = partially relevant; N = not relevant)

1. The relationship and communication processes between middle aged children and their parents. (Searched in DIALOG File 11; $R = 27$, $pR = 46$; $N = 75$)
2. Design, structure, and organization including overall integration of the acute care nursing department in the hospital. (DIALOG File 218; $R = 37$; $pR = 36$; $N = 156$)
3. Stereotypes which affect the diagnosis of child abuse by health care providers. (DIALOG File 64; $R = 36$; $pR = 47$; $N = 68$)
4. Effects of controlled lung hyperinflations, before and after endotracheal suctioning, on the cerebrovascular status of adults with severe closed head injuries. (DIALOG File 154; $R = 60$; $pR = 58$; $N = 33$)
5. Rules-of-thumb, industry by industry. (DIALOG File 148; $R = 16$; $pR = 23$; $N = 48$)
6. Prevention of carbon dioxide crystal growth on the interior surfaces of reactors. (DIALOG File 6; $R = 11$; $pR = 5$; $N = 134$)
7. Factors which impede strategic human resource management. (DIALOG File 75; $R = 70$; $pR = 39$; $N = 40$)
8. Effects of an aerobic interval training program on the physical and psycho-social health of menopausal women. (DIALOG File 154; $R = 2$; $pR = 5$; $N = 54$)
9. Alternatives for delivery of human services other than the classical model of individual casework in an agency based office. (DIALOG File 37; $R = 18$; $pR = 48$; $N = 84$)
10. Motivations of adults choosing to discontinue chemotherapy. (DIALOG File 154; $R = 4$; $pR = 15$; $N = 130$)
11. Psycho-emotional and psycho-social responses of parents and surviving siblings to an infant's death due to Sudden Infant Death Syndrome (SIDS) (DIALOG File 154; $R = 9$; $pR = 25$; $N = 115$)
12. Chemical reactivity of silicon carbide and silicon nitride ceramic powders at low (room) temperatures especially in aqueous environments. (DIALOG File 13; $R = 6$; $pR = 21$; $N = 121$)
13. Definition and measurement of effectiveness in non-profit human service organizations. (DIALOG File 15; $R = 7$; $pR = 36$; $N = 106$)
14. Changes in the function of hospital information systems due to the advent of prospective payment systems. (DIALOG File 151; $R = 35$; $pR = 71$; $N = 51$)
15. Occurrences, causes, treatment, and prevention of retrolental fibroplasia. (DIALOG File 154; $R = 28$; $pR = 86$; $N = 36$)
16. Retirement activities including pre-retirement indicators of retirement activity patterns. (DIALOG File 11; $R = 25$; $pR = 37$; $N = 108$)
17. Pumps and control systems for drug delivery in animal experiments and clinical applications. (DIALOG File 5; $R = 36$; $pR = 26$; $N = 88$)
18. Managerial competencies especially as applied to physician-managers. (DIALOG File 15; $R = 66$; $pR = 38$; $N = 46$)
19. Perceived impact of the 1977 Institute of Internal Auditors Standards. (DIALOG File 75; $R = 27$; $pR = 49$; $N = 74$)
20. Presentation of financial statements, especially the disclosure requirement form of the SEC. (DIALOG File 15; $R = 26$; $pR = 43$; $N = 81$)
21. Social support networks and the physical and mental health of never married older women. (DIALOG File 37; $R = 19$; $pR = 6$; $N = 77$)

22. Space commercialization forecast. (DIALOG File 108; $R = 15$; $pR = 135$; $N = 0$)
23. Sintered powder metal or powder metal parts infiltrated with copper or bronze. (DIALOG File 32; $R = 29$; $pR = 9$; $N = 51$)
24. Meaning of the cat in Italian renaissance (1450–1600) religious paintings. (DIALOG File 191; $R = 1$; $pR = 4$; $N = 30$)
25. Relationship between oral and written language and communication of basic writers. (composition students) (DIALOG File 1; $R = 30$; $pR = 26$; $N = 94$)
26. Policies of creating administrative agencies for purposes of compensating industrial workers accidentally killed or injured in Ohio or Ontario from 1915 to 1935. (DIALOG File 38; $R = 37$; $pR = 39$; $N = 8$)
27. Principles and design of miniature high pressure sensors. (DIALOG File 13; $R = 35$; $pR = 69$; $N = 46$)
28. History from 1800 of University Circle in Cleveland focusing on philanthropy, city planning and public vs. private development. (DIALOG File 38; $R = 5$; $pR = 23$; $N = 39$)
29. Firing or sintering of ceramic material using microwave radiation. (DIALOG File 8; $R = 36$; $pR = 34$; $N = 80$)
30. Creative evasion of censorship in South Africa. (DIALOG File 71; $R = 57$; $pR = 25$; $N = 13$)
31. Budgeting, especially automated acquisition budgeting, in law libraries. (DIALOG File 61; $R = 14$; $pR = 15$; $N = 85$)
32. Engineering properties and various utilizations of fly ash as a construction material. (DIALOG File 8; $R = 113$; $pR = 19$; $N = 18$)
33. Volume-averaged equations used to determine friction factors of 2-phase slurry flow in pipelines. (DIALOG File 8; $R = 44$; $pR = 57$; $N = 49$)
34. Expert systems directed by the user and not by an inference engine. (DIALOG File 13; $R = 10$; $pR = 39$; $N = 100$)
35. Music therapy for the chronically ill, especially cancer patients. (DIALOG File 154; $R = 31$; $pR = 20$; $N = 14$)
36. Industrial policy in Austria and Western Europe related to technological innovation, restructuring of industry, the EEC, and corporatism. (DIALOG File 90; $R = 62$; $pR = 49$; $N = 39$)
37. Training of employees on the right to know (RTK) laws, OSHA hazard compliance laws, chemical safety, and handling of hazardous materials. (DIALOG File 16; $R = 78$; $pR = 18$; $N = 54$)
38. Future of document acquisition, cataloging, storage, and information dissemination in the automated technical reference library. (DIALOG File 61; $R = 79$; $pR = 29$; $N = 42$)
39. Environment of a corporation as it affects organizational structure. (DIALOG File 15; $R = 26$; $pR = 38$; $N = 102$)
40. Known or proposed techniques for bacterial cloning and the commercial activity surrounding the technology. (DIALOG File 16; $R = 77$; $pR = 40$; $N = 32$)

Total in 40 questions:

$R = 1343$; $pR = 1448$; $N = 2620$

$R + pR + N = 5411$ evaluated items

Acknowledgments

The complex and lengthy process of assembly of data for this project, including communication with users, administration of searchers and searching, and compilation of "raw" data was managed by Alice Y. Chamis and Donna Trivison. The programming for statistical analyses was done by Jun-Min Jeong, J. J. Lee, Moula Cherikh, and Altay Guvenir. Their contribution is fully and gratefully acknowledged.

References

1. Saracevic, T.; Kantor, P.; Chamis, A. Y.; Trivison, D. "A Study of Information Seeking and Retrieving. I: Background and Methodology," *Journal of the American Society for Information Science*. 39(3): 161–175; 1988.
2. Saracevic, T.; Kantor, P.; Chamis, A. Y.; Trivison, D. *Experiments on the Cognitive Aspects of Information Seeking and Retrieving. Final Report for National Science Foundation Grant IST-8595411*. National Technical Information Service; (PB87-157699/AS). Educational Research Information Center; (ED 281530). 1987.
3. Sparck-Jones, K., Ed. *Information Retrieval Experiment*. London: Butterworths; 1980.
4. Cleverdon, C. W. "On the Inverse Relationship of Recall and Precision," *Journal of Documentation*. 28(3):195–201; 1972.