Decision Making in Online Searching

Lyn Blackshaw
University of Oregon, DeBusk Memorial Center, 1675 Agate Street, Eugene, OR 97403-1215

Baruch Fischhoff
Carnegie-Mellon University, Pittsburgh, PA 15213

A set of methods and results is offered for characterizing how people make decisions in the course of using computerized databases. In general, their performance resembles that revealed in studies of decision making in other contexts. In particular, people are only moderately sensitive to the likelihood of their succeeding, being overconfident for all but the easiest of tasks. These results are discussed in the context of previous research in information science and decision science, and with regard to their implications for the design of databases and the adaptation of users to them.

Using a database forces one to make a variety of decisions under conditions of uncertainty. The first of these decisions might be whether a particular database is the best place to look for some information, not knowing exactly what it contains nor how readily it will yield its contents. A second decision might concern which of several possible search tactics to try. A third might concern which offering in an entry-level menu or free-text lexicon provides the best gamble for getting one closer to the goal. A fourth might be whether to go it alone or seek help. A fifth is whether to accept the products of a search as containing or exhausting the sought information.

Facing such choices, an idealized individual who adhered strictly to the tenets of decision theory would begin a systematic evaluation of the alternatives. Its steps would include: (a) identifying the goals of the search, considering the relative importance of its possible positive and negative outcomes (e.g., getting the right information but being unsure of it, spending resources in the search, getting some wrong information but not realizing it); (b) identifying all feasible alternatives for achieving the desired outcomes; (c) assessing the likelihood of receiving the different outcomes with the different alternatives; (d) evaluating the options in terms of their likely outcomes; (e) evaluating the decision-making process, considering the need for refinement and the trust to place in its products. In principle, one could even use the formal procedures of decision analysis for identifying the best gamble among the options [1,2]. (Extensions of the logic to information retrieval may be found in several other studies) [3,4,5]. In practice, however, such procedures are seldom used even by scientists or research administrators in the course of their attempts to choose the studies that offer the best expected information yield [6]. Lay users of databases may not even address all the stages in a comprehensive decision-making process (a–e above), neglecting perhaps to consider in any seriousness the consequences of being wrong or any alternatives beyond the first one or two that come to mind.

Thinking of database search as a decision-making process may be particularly important when designing and operating heterogeneous databases intended for diverse user populations (e.g., Prestel, Dow–Jones News Retrieval). However conscientious indexers might be, there is little chance of making the location of all items clear to all users. With a small database, individual users might try to reduce the uncertainty about item location by learning the indexers' frame of reference. However, as database size and diversity increase, users might reconcile themselves to the presence of uncertainty and concentrate on gambling well in their choice of location. In such cases, facilitating those gambles would be a vital feature of system design.

As part of a project examining intuitive decision-making processes in information search, we have examined performance in a number of simple, well-structured search tasks [7,8,9]. Here, we move to more complex and less structured search, using the computerized catalog of our local library. We offer a number of techniques and some suggestive results regarding the extent to which people approach search as a decision-making process and the success with which they do so. A particular focus is how people assess the uncertainty surrounding their decisions. Having realistic expectations is essential for knowing how hard to work on a search, how soon to become frustrated if it is unsuccessful, how carefully to scrutinize its products, and how quickly to ask for help.

Received September 14, 1986; revised December 15, 1986; accepted January 26, 1987.

Method

Subjects

Sixty subjects responded to advertisements placed alongside the computer catalog terminals in the Eugene Public Library, in a free entertainment newspaper distributed outside supermarkets, and in the University of Oregon student newspaper. Each subject was offered eight dollars to take part in a 90-minute experiment. All claimed to have held a Eugene Public Library borrower's card for at least a year (mean = 6.5, median = 3) and to have used the computerized catalog. They reported visiting the library 4.6 times per month on average. There were 41 women and 19 men, with a mean age of 30.3 years. Subjects reported a high level of formal education, with 55 having gone beyond high school and 30 currently in higher education. Thirty-one reported working; a number were unemployed.

Materials and Equipment

The Eugene Library online catalog is a Canadian system, ULISYS, which was originally purchased for inventory control and circulation and later updated to include a public access module. The catalog contains 253,000 entries listed under Library of Congress subject headings, as well as under Author, Title, Author and Title, and Call Number. In addition to books, the catalog includes magazines, recordings, and framed prints. The Eugene Public Library provided us unrestricted daytime access to their vendor's telephone line. Our experiments used a Leading Edge (Model D) computer and Hayes Smartmodem 1200; together, they provided response times like those experienced in the library.

The experiments took place in a well-lit and ventilated office with subjects seated at the computer. The experimenter (Lyn Blackshaw) faced them for an initial interview, while asking questions about their personal background, use of the library, and expectations about the catalog. After this attempt to establish rapport, the experimenter sat behind subjects but in view of the computer screen.

A tape recorder was used as soon as the search began, making about 70 hours of recording available for coding subject responses. These were also recorded in writing during the experiment by the interviewer. A clock with a sweep second hand was placed in an unobtrusive place to avoid time pressure.

Design

As they arrived, subjects were assigned in order to one of four experimental groups which differed in the advice that they received regarding how to exploit the system. One group received a two-page abbreviated version of the advice provided by the library itself; one group received a page of advice culled from the suggestions of pretest subjects; one group received a page of advice that we created, stressing how to view the search as a decision-making process; the control group received no advice.

All subjects participated in the same six-part experiment: (1) *introduction:* a set of background questions asks subjects about themselves and about their library experiences. (2) *timed pretest:* a set of three standardized test items (one subject, one author, and one title) which subjects were to find as best they could; it was designed to establish (we hoped) the four groups' equal initial ability levels and to provide an opportunity to eliminate any problems that subjects had with the mechanics of the system. (3) *initial unstructured search:* in which subjects searched for three items of their own choosing, in whatever way they chose, while the experimenter recorded their responses. (4) *semi-structured search:* in which subjects sought books of their own choosing while under instruction to describe their thought process concurrently, particularly with regard to their perceived chances of success and search strategies. (5) *timed post-test:* in which subjects sought three specified items while being timed and encouraged to think aloud, although without any specific instructions regarding thoughts to emphasize. (6) *abstract categories:* an unrelated task in which subjects attempted to identify the location of 11 items of information in the *Statistical Abstract of the United States;* comparing responses here with those of subjects who (in an earlier study) had completed this questionnaire alone allowed an assessment of the overall impact of the present procedures.

Procedure

Four two-hour slots were scheduled on three days a week between 8:00 a.m. and 5:00 p.m. Times for the four experimental groups (three kinds of advice and one control) were pre-allocated equally to the time slots in order to balance slight variations in computer response time due to differences in concurrent terminal usage at the library. Subjects telephoned our office to volunteer and were assigned to times according to their convenience, so that no further effort was made to balance sex or age. On entering, subjects were told about the experiment, invited to ask questions, and then requested to sign an informed consent form.

Introduction. The interviewer began with "a few background questions" about the subject's schooling, employment, experience with computers, and library usage. Subjects then estimated various aspects of their personal search behavior. The former questions included: (a) how often they typically knew what books they wanted before entering the library; (b) how likely they were to browse the shelves; (c) how likely they were to use the hard-copy card catalog; (d) how likely they were to use the computerized catalog; (e) how they distributed their searches over those three sources. Subjects were also asked some open-ended questions regarding their search habits. The search experience questions asked for estimates of the likelihood (in percentage) that (a) they would find the books they want; (b) that an entry they sought in the computerized catalog would be in there; (c) that they would find that entry, assuming that it was there; (d) that they would be able to find a book on the shelf, if they had found its call number.

Timed Pretest. Subjects then received "three items to search for to help you familiarize yourself again with the system." The items included one author (Tennessee Williams), one title (*Tender Is the Night*), and one subject item (salmon fishing in the Northwest Pacific), in that order. Minimal instruction was given regarding the location of essential keys. Timing began as soon as the card was received and ended when the search item came up on the screen. Before beginning an item, subjects answered two questions: "How confident are you, from 0 to 100, that this is listed in the computer catalog?" and, after being assured that it was listed, "How confident are you, from 0 to 100, that you will find this by yourself without any help from me?"

After completing the first search, subjects were shown how to type inquiries without returning to the initial entry-level menu each time; this had been a major irritant in experimental pretests. During the searches, the experimenter remained silent and transcribed the entries that subjects typed. After the three timed searches, subjects were given the opportunity to talk about their experience. Only basic instruction in system mechanics was given in response to requests for help (e.g , do not type until the query prompt appears, do not leave a space between the entry command and the equals sign).

Experimental Manipulation. The control group received no further advice. The *expert* group received two pages of advice prepared by city library staff on the best way to use the computer catalog. The *subject* group received advice compiled from pretest subjects. The *theory* group received written advice based on decision theory. After reading the statement, subjects rated it on a 10-point scale for understandability (0=not at all understandable; 10=totally understandable). The advice statements appear in Appendix A.

Unstructured Search. The experimenter then said:

> I am now going to ask you ____ (name) to look for some entries that arise out of your interests
>
> This time I won't be answering any questions or giving any help; there'll be an opportunity to discuss eVerything later.
>
> Tell me exactly what you decide to look for, what alternatives you thought of, and how confident you are that you'll be successful.
>
> Talk aloud any concerns, questions, successes at each step of your search. I'll be putting the tape recorder on."
> Each subject was also handed the following written instructions:
> I'd like you to look, in any order you like, for:
> * A book on a topic that interests you but that you haven't searched for previously.
> * A book by a particular writer who interests you.
> * The title of a book you'd like to read.
> Remember to talk aloud and tell me what you're doing and thinking from the moment you read this.

As each subject searched, the researcher wrote down verbatim what was spoken as well as describing nonverbal behaviors (e.g., commands typed, browsing, choosing options on the computer, outcomes). The transcript of verbal behavior was checked against the audio recording later.

After subjects completed the three searches to their own satisfaction, the experimenter asked several follow-up questions. For each search goal, subjects estimated how confident they had been, before starting, (a) "that it was listed in the computer" and (b) that they "would be able to find it." They were then asked in an open-ended fashion about "any further comments on your personal method of search" and "any comments on the way that information was organized or presented on the computer." After rating the advice statements (which they were encouraged to reread if appropriate) on a 10-point usefulness scale (0=not at all useful; 10=extremely useful), they were asked to comment on the statements and to provide any advice that they themselves would give to "someone else searching for books in the computerized catalog."

Semi-Structured Search. In the next set of three searches, subjects were asked to talk aloud all their thoughts, to give confidence ratings and, when a search strategy had failed, to describe their plan for what to do next. No help or advice was given, with the few requests for help being redirected to the subject: "What would you do?"

In order to motivate the semi-structured search, an effort was made to find search goals of personal interest to the searcher. For the first search, subjects were asked for the name of a book that they had enjoyed reading from the public library, and then asked to find "another book on the same or a similar topic that you haven't read before." For example, a person who described Barbara Deming's *We Cannot Live without Our Lives* chose to look for another book on nonviolent activism. After identifying a topic, subjects were asked where they were going to look and how confident they were (a) that it was listed in the computer and (b) that they would be able to find it without help from the experimenter. The experimenter recorded verbal and nonverbal behavior as before. If subjects became silent they were invited to "talk aloud, tell me everything that goes through your thoughts, it doesn't matter how irrelevant you might think it is; it might be very useful in our study." If subjects chose another entry option, then they were asked again for their confidence level.

The second and third searches followed the same procedure and were initiated as soon as a subject either claimed success or gave up on the preceding search. The second search goal was derived from the subject's leisure interest. They were asked to find a book they'd like to read on one of those stated interests The third search goal was derived from work, paid or nonpaid, that subjects reported doing or thought they might enjoy. They were asked to find an author that they had not heard of before who had written two or more books on one of those work interests.

Timed posttest. Three search goals were selected from topics sought by pretest subjects. They were formulated so that they could not be accessed directly under the subject headings offered to participants Subjects were told that they would be timed in searching for topics that were not necessarily listed in the catalog under that name. Once again, subjects were asked to "talk aloud" all their thoughts. Timing started when they received a card giving their goal and ended when they abandoned or completed the search (in the sense of identifying a book that they thought satisfied the goal). The three goals were: 1) a book on snorkeling, 2) a book about fatness, and 3) a book on cabinet making.

When the last search had been completed, subjects were asked the same follow-up questions as after the Unstructured Search. Finally, they rated their confidence that "assuming that a book is listed somewhere in the computer catalog, that you will find it."

To conclude the session, subjects who expressed an interest were given feedback by the reseacher and some tips and short cuts that would help them in their use of the computer catalog when they next visited the Eugene Public Library. They were then given the *Abstract* categories questionnaire to complete in a self-paced manner in a separate room.

Abstract Categories. The questionnaire asked subjects to determine the location of 11 items of information (e.g., the mean annual temperature in Bismark, ND) among seven categories (e.g., Economy, Foreign Affairs and Immigration) summarizing the 33 chapters of the *Statistical Abstract of the United States* [10]. Both the contents of the categories and the labels assigned to them were determined by groups of lay subjects according to an involved procedure described by the present authors [8]. Subjects' specific task was, first, to choose the three most likely locations, in order of likelihood and, then, to assign each the probability that it would be correct (with the residual probability going to "All Other Categories"). No feedback was provided regarding the appropriateness of their choices (the researchers had found that this plausibly potent manipulation, provision of outcome feedback, had no effect on performance [9]).

Responses can be scored in terms of (what we have called) their *transparency,* the extent to which subjects can see which items are included in a category, and their *metatransparency,* the extent to which subjects can see how transparent a set of categories is (and how successful they will be in using it). As argued by several authors [7,8,9, 11,12], both properties are important for effective use of any imperfect database.

Results and Discussion

Subjects

Table 1 summarizes subjects' reports regarding their background and experience with the library. The means of the groups show the kinds of individuals who participated in this study and to whom its results can be most confidently generalized. The differences among the means show the extent to which our subject allocation procedure succeeded in equating the individuals in the four experimental groups. Although there were moderate differences among the groups on some of the demographic variables, these were not associated with any very large differences in their reported patterns of locating library materials. Almost all browse, while just over half use the card catalog at some times. (Having used the computer catalog was a requirement for participation.) The mean frequencies with which they report relying on each of these three sources roughly parallel the frequencies with which they report relying on each source at all.

Expectations

Prior to engaging in any searches, subjects expressed their expectations regarding the system in several different ways. Such expectations might affect their willingness to use a database at all, their satisfaction or frustration with the progress of searches, their readiness to ask for help, or their willingness to accept the results of a particular search as

TABLE 1 Self descriptions of subjects.

	Advice Group				
Measure	None (Control)	Expert	Subject	Theory	All
Mean age	31.6	31.8	29.5	28.2	30.3
Female (%)	80.0	73.3	53.3	66.7	68.3
Monthly Library Visits	5.2	4.1	4.7	4.3	4.6
Working (%)	60.0	33.3	60.0	53.3	51.7
Library Card (Years)	5.0	6.2	8.2	6.4	6.5
Enter Library With Purpose (%)	70.4	85.7	70.6	66.1	73.2
Browse (%)	93.3	86.7	93.3	93.3	91.7
Use Card Catalog (%)	66.7	53.3	60.0	53.3	58.3
Use Computerized Catalog (%)	100.0	100.0	100.0	100.0	100.0
Rely on (%)*					
Browsing	41.4	25.4	34.6	35.6	34.4
Card Catalog	27.0	31.9	24.0	20.0	25.9
Computerized Catalog	37.5	49.3	43.7	49.2	44.9

*Mean percentage of reliance on each source among subjects who reported relying on it at all.

definitive (e.g., is what they found what they want? If they were unable to find anything, does that mean there is nothing in there?)

As shown in Table 2, subjects thought that there was about a 70% chance that a typical book that interested them would be in the catalog and an 87% change that they would be able to find it if it were there, making for a (conjoint) probability of .61 (= .70 × .87) of finding a typical book in the computer catalog. Thus, they had a relatively high opinion of their own ability to use the catalog, but a less sanguine perception of the catalog's completeness (and, by implication, of the library's completeness, for those who knew, or believed, that the catalog included the entire collection). Additionally, only 77% expected to be able to find a sought book on the shelves. Whether such expectations are adequate to motivate a computer search should depend upon its importance, its costs, and its alternatives. (The similarity of the means for the four groups indicates that they entered the experiment with similar expectations in this respect.)

Subjects' expectations of being able to find individual items varied considerably. In the pretest, they were most confident in being able to find a book by a particular author (Tennessee Williams), given that it was in the catalog, and least confident in being able to find a book on a particular topic (salmon fishing). As discussed below and found elsewhere in studies of search performance [3,11,13], these differing expectations for the different kinds of search may have some justification.

One sign of subjects' ability to distinguish these two probability judgments is the weak correlation (tau = .19) between the mean judged probabilities of items being in the catalog and of being found if they were there.

The final row of Table 2 shows a second estimate of the likelihood of finding a typical item in the database, made at the very end of the experimental session. It is directly comparable to the question in Row 3. In each group there was a decrease in confidence, with the overall probability going from .87 to .79. As reported below, these subjects experienced an 85% success rate on the intervening tasks. Assuming that subjects had an accurate perception of their success rate [14] this decrease in confidence, despite performance that approximately met subjects' expectations, suggests that they either viewed these tasks as easier than their typical task or felt somewhat lucky to have done so well.

Outcomes

An obvious measure of search success is whether subjects locate the item that they seek. It is, however, sometimes less than obvious whether they have met that criterion, particularly with topic searches where it may be arguable whether a given book actually represents the sought category [5,12,15]. As a result, we have distinguished clear-cut successes from ambiguous ones (i.e., situations in which subjects were satisfied that they had attained their goal, but

TABLE 2. Expectations (mean probability).

	Advice Group				
Confidence in:	None	Expert	Subject	Theory	All
General Questions					
Finding book in library	58	76	70	64	67
Sought book being in computer catalog[a]	72	73	73	61	70
Finding book if in catalog[b]	87	87	87	86	87
Find book on shelf	72	74	88	76	77
Pretest					
Tennessee Williams					
in catalog	99	91	96	97	96
will find	95	91	94	99	95
Tender Is the Night					
in catalog	85	83	82	79	82
will find	88	88	90	97	91
Salmon Fishing					
in catalog	86	87	87	79	85
will find	82	78	80	82	81
Combined					
in catalog	90	87	88	85	88
will find	88	86	88	93	89
At end					
Find a book if in catalog[c]	76	84	81	76	79

[a] $F(3,56) = 1.09, p = 0.36$
[b] $F(3,56) = 0.03, p = 0.99$
[c] $F(3,56) = 1.07, p = 0.37$

TABLE 3 Search outcomes (all groups; %)

	Outcome			
Task	Success	Ambiguous	Satisfied	Gave Up
Pretest				
Tennessee Williams (AU)[a]	97	—	97	3
Tender Is the Night (TI)	100	—	100	—
Salmon Fishing (SU)	88	—	88	12
Unstructured Search				
Author (AU)	87	2	89	11
Title (TI)	85	2	87	13
Subject (SU)	62	13	75	25
Semi Structured				
Enjoyed (SU)	75	8	83	17
Leisure (SU)	76	9	85	15
Work (SU)	53	—	53	47
Posttest				
Snorkeling (SU)	67	15	82	18
Fatness (SU)	85	2	87	13
Cabinet Making (SU)	75	20	95	5
All Tasks				
Author	92	1	93	7
Title	93	1	94	6
Subject	73	9	81	19
Combined	79	6	85	15

[a] Abbreviations refer to kind of search: AU = author; TI = title; SU = subject.

we were not). Thus, in Table 3, the percentage of cases in which subjects were satisfied includes both clear-cut "successes" and "ambiguous" ones.

Subjects had little trouble with the author and title tasks in either the pretest or the unstructured search. In the latter case, most difficulties came from uncertainty about subjects' own goals (e.g., for "the title of a book you'd like to read"), rather than from difficulty with the system. As shown in the bottom section of the table, subjects found their author and title goals in over 90% of cases, and seldom did so in a way that caused us to think that they might have been too easily satisfied.

Topic searches were a different matter. With the well-defined pretest and posttest topics, subjects gave up in 12% of cases, apparently reflecting the difficulty of finding the synonyms under which these items were located. An additional sign of this difficulty is the relatively high rate of ambiguous successes, particularly for "snorkeling" and "cabinet making," where subjects may have been uncertain about the meaning and appropriateness of related terms given the relative unfamiliarity of the focal topics. There were fewer such ambiguous successes in the four cases (in the unstructured and semi-structured searches) where subjects defined their own goals. On the other hand, subjects gave up in 26% of those cases (compared with the 12% where we set the goal). Apparently, defining one's own goal adds an additional element of difficulty, even if it is somewhat more motivating. Nonetheless, these are considerably higher rates of success than have been observed in other studies of nonexpert subjects' search behavior [11,13,16,17,18]. In part, it seems due to our using a laxer criterion; asking whether casual readers can use the system to find something useful on a topic, compared with other investigators' focus on scholarly users' ability to find the most relevant material on their topic [19]. On the other hand, the greater education, subject matter expertise, and perhaps motivation of those subjects might have been expected to produce superior performance there.

One distinctive feature of the current experiment is that in the pre- and posttests subjects were timed, although not in a particularly obtrusive fashion. That feature does not seem to have created a tendency to give up (e.g., by accentuating feelings of failure) or to settle for ambiguous successes (i.e., prompting a speed-accuracy tradeoff). Table 4 shows the amount of time invested by subjects (in seconds). They worked relatively hard, averaging two minutes on the pretest and four minutes on the posttest, before expressing satisfaction or giving up. In each case, mean search time was larger than the median, indicating a distribution of search times that was skewed somewhat by a number of subjects willing to work quite a long time, relative to others. The right-hand column of the table shows the results of analysis of variance tests for the difference in search times for subjects in the different conditions (the skew in the distributions did not seem large enough to violate badly the assumption of normality underlying these inferential statistics). In only

TABLE 4. Search times (mean seconds; median in parentheses).

	Advice Group					
	None	Expert	Subject	Theory	All	P
Pretest						
Tennessee Williams	106	199	122	109	134	0 02
	(95)	(165)	(85)	(95)	(95)	
Tender Is the Night	57	52	65	62	59	0 76
	(50)	(50)	(50)	(60)	(50)	
Salmon Fishing	179	205	132	197	178	0.22
	(175)	(210)	(115)	(180)	(145)	
Combined	114	152	106	123	122	0 12
Posttest						
Snorkeling	333	247	211	286	269	0.28
	(210)	(205)	(155)	(265)	(208)	
Fatness	156	295	241	197	222	0.09
	(140)	(320)	(160)	(150)	(165)	
Cabinet Making	232	201	196	256	221	0 66
	(200)	(120)	(140)	(290)	(178)	
Combined	240	248	249	246	238	0 81

one case (the first), did the difference reach customary significance levels. Insofar as that pattern was not repeated with any consistency elsewhere, it provides only a weak suggestion that subjects receiving the "expert" advice performed more poorly. This weak suggestion received slight additional support in Table 5, which shows a slightly lower overall success rate for this group in the pretest, meaning that they worked longer and had less to show for it. On the other hand, in the posttest this group had the highest success rate, despite working an equivalent amount of time. All in all, these results seem to provide further evidence of the similarity of these conditions, within the limits of statistical variability.

Realism of Expectations

In addition to success rates, Table 5 provides a measure of how well subjects appraised their ability to use the system in the pretest. Called *over/underconfidence,* it is the signed difference between subjects' mean estimate of their probability of success and the proportion of successful searches (the expectations themselves appear in Table 2). A positive difference shows subjects to be correct less often than they would have expected. The preponderance of negative signs indicates widespread underconfidence. For example, subjects receiving no advice were 100% successful in seeking the first pretest item, but only 95% confident on the average,

TABLE 5 Performance on timed searches (%)

	Advice Group				
	None	Expert	Subject	Theory	All
Pretest					
Tennessee Williams					
success	100	93	93	93	97
over/underconfidence	−5	−2	1	6	−2
Tender Is the Night					
success	100	100	100	100	100
over/underconfidence	−12	−12	−10	−3	−9
Salmon Fishing					
success	93	73	93	93	88
over/underconfidence	−11	5	−13	−11	−7
Combined					
success	98	89	96	98	95
over/underconfidence	−10	−3	−8	−5	−6
Posttest (success alone)					
Snorkeling	73	100	80	73	82
Fatness	87	93	80	87	87
Cabinet Making	93	100	87	100	95
Combined	84	98	82	87	88

representing slight (i.e., 5%) underconfidence. Subjects were underconfident in 10 of the 12 comparisons in the table and by 6% overall.

Although the differences in success rates were not large across groups, there was still a positive correlation between success rate and confidence (tau = .46 ignoring ties, .29 considering ties). These results repeat a pattern observed elsewhere, in which confidence increases with performance but the correlation is imperfect [20] and the overall tendency is toward overconfidence with all but the easiest tasks, where underconfidence is observed. In this light, the present searches were quite easy for subjects both in an absolute sense and relative to their overall expectations of success in locating books (Table 2; top). Subjects realized that the present items were relatively easy, but not by how much.

The small differences in over/underconfidence among the groups seem better attributed to differences in their success rates, rather than to any differences in how they assessed the extent of their abilities. We hoped that the "theory" advice, which emphasized comparing the viability of competing alternatives, would have increased subjects' realism about their chances. By these measures, at least, it did not.

Table 6, which breaks down confidence and realism for the semi-structured search, shows a different pattern. Two

TABLE 6. Confidence and realism in semi-structured search.

	Advice Group				
Topic	None	Expert	Subject	Theory	All
Another book on an enjoyed topic					
confidence in finding					
first goal	65	79	73	53	67
all goals	62	70	66	44	61
success in matching					
first entry	53	80	40	36	52
all entries	57	71	47	44	55
success in reaching goal[a]					
with first entry	40	53	27	20	35
over/underconfidence[b]	25	26	46	33	32
Book on leisure topic					
confidence in finding					
first goal	80	85	78	79	80
all goals	70	83	74	71	75
success in matching					
first entry	58	77	75	77	72
all entries	48	83	65	76	68
success in reaching goal					
with first entry	50	46	50	46	48
over/underconfidence	30	39	28	33	32
Two books by same author on work topic					
confidence in finding					
first goal	59	71	65	69	66
all goals	58	64	62	61	61
success in matching					
first entry	54	69	55	50	57
all entries	47	52	44	58	50
success in reaching goal					
with first entry	54	15	27	7	26
over/underconfidence	5	56	38	62	40
All					
confidence in finding					
first goal	68	78	72	67	71
all goals	63	72	67	59	65
success in matching					
first entry	55	75	57	54	60
all entries	51	69	52	59	58
success in reaching goal					
with first entry	48	39	34	24	36
over/underconfidence	20	39	38	43	35

[a] Searches are considered successful here if subjects were satisfied with their result. Thus, "percentage success" includes both the clear-cut and the ambiguous successes in Table 4.

[b] Over/underconfidence is computed as the signed difference between subjects' confidence in finding a book meeting their first goal and the percentage of successes, as defined by subjects (see note *a*)

measures of confidence are given. The first refers to subjects' confidence that they will find a book on the topic under the first entry that they chose, considering both the possibility that the system will not include a book on their topic and the possibility that they will not be able to find it. The second estimate includes confidence ratings for any subsequent entries that subjects produced in the course of their search. In each case, confidence in all entries was less than confidence in the initial entry, indicating that changing entries was perceived as a sign of trouble. Often, it involved a broadening of the search goal which should, in principle, increase the chances of finding something appropriate. Less commonly, it involved a refinement of the goal, in .pa which case there might have been a lower probability of identifying a more suitable goal.

A necessary condition for succeeding with an entity is that it be included in the Library of Congress lexicon that underlies the database. Over all tasks, subjects found that only 60% of their first entries and 58% of all their entries were recognized. These "matches" were considerably more likely in the search for leisure topics than for the other two topics, indicating a closer match there between the mental representations of these users and of the indexers. However, the low match rate meant that subjects' chances of finding the book (or books) they were looking for under their first entry was less than their assessed probability of finding it there. This discrepancy indicated a high degree of overconfidence, which the realism scores showed to be substantial. Somehow, subjects' experiences prior to the experiment and on the preceding topic-search tasks had not shown them how different the Library of Congress view of the world was from their own.

Had subjects made but a single entry, their performance would have been dismal (36% success overall). However, in the setting that we had created, most persisted. As shown in Table 3, subjects eventually reached an outcome that satisfied them in 79% of cases in the three semi-structured tasks. The lower success rates observed in previous studies (cited above) may reflect, in part, lesser willingness or opportunity to persist. We speculate that subjects' lack of realism, which emerges as the recurrent disappointment of confident choices, may be a discouraging factor, which was somewhat reduced here by the presence of an encouraging experimenter and by payment for a fixed time period.

As elsewhere, there was a weak overall relationship between confidence and success (tau = .25 ignoring ties, .23 including ties).

Summarizing across items, there were moderate differences among the experimental conditions in both confidence and success. Moreover, those differences went in somewhat different directions. As a result, there were some appreciable differences in the realism of the different groups, with subjects receiving no advice having less confidence and more success, resulting in reduced overconfidence. In the absence of intergroup differences elsewhere, this seems like a chance result.

In the pretest searches, subjects' confidence in their ability was conditioned on each item being in the database (which it always was). Here (and in the immediately preceding unstructured search), the probability question was unconditional. Although subjects had no apparent difficulty distinguishing the two question before, there is perhaps some chance of subjects confusing the two questions here. If so, the confusion would have increased their reported confidence and artifactually increased the observed overconfidence. We do not believe this to be substantial problem.

Retrospective Expectations

In keeping with the effort to interfere as little as possible with the unstructured searches, subjects were not asked to give their expectations prior to these tasks. Rather, once the three searches were over, subjects estimated how confident they *had been* that an entry satisfying each goal was in the catalog and that they would be able to find it. The means of these estimates appear in Table 7, along with an overall probability of success which is inferred by multiplying these two probabilities. They are different from the comparable estimates in Table 6 in two ways. They refer to the chance of success on *some* attempt (not just the first one) and they are provided after subjects learn whether they have succeeded or failed. The former difference in procedure should increase estimates—if subjects' estimates are internally consistent. The latter difference should have no effect—if subjects can reconstruct their presearch perspective. There is, however, a good deal of evidence suggesting that, unbeknownst to them, people have difficulty undoing the effects of knowing how things turned out. As a result, they believe that they (and others) knew more in foresight than was actually the case [21,22]. In the present context, such a "hindsight bias" would increase the probabilities assigned to successful searches (because it is hard to see how they could have gone wrong) and decrease the probability of success attributed to unsuccessful ones. The aggregate effect should be to bring the estimated probability of success closer to the actual probability of success.

Direct comparisons between the results in Tables 6 and 7 are inappropriate because different search goals were involved. For example, the mean (inferred) probability of finding an item at all in the unstructured searches was only slightly larger than the mean overall probability of finding an item on the first try in the semistructured search (.73 vs. .71), even though subjects frequently made several attempts. That could be a sign of inconsistency or simply reflect the greater difficulty of the unstructured search tasks.

Perhaps a safer comparison is between relationships created within the two sets of data. For example, there was again a weak positive correlation between knowledge and confidence (tau = .28 ignoring ties, .26 considering ties). However, whereas the mean absolute levels of confidence and knowledge were very far apart in the semi-structured

TABLE 7. Confidence and success in unstructured search.

	Advice Group				
Goal	None	Expert	Subject	Theory	All
Topic of Interest					
Confidence in					
entry being in catalog	82	83	82	69	79
finding it if there	89	89	94	88	90
success (inferred)	73	74	77	61	71
Success					
with first entry	47	50	63	50	52
with any entry	93	50	88	67	74
in matching all entries	61	70	70	50	62
Over/underconfidence	−20	24	−11	−6	−3
Writer of interest					
Confidence in					
entry being in catalog	91	84	87	79	85
finding it if there	96	90	93	97	94
success (inferred)	87	76	81	77	80
Success					
with first entry	60	53	50	80	60
with any entry	87	65	70	80	75
in matching all entries	68	76	69	71	71
Over/underconfidence	0	11	11	−3	5
Title of a Book					
Confidence in					
entry being in catalog	83	84	72	71	78
finding it if there	96	90	86	85	89
success (inferred)	80	76	62	60	69
Success					
with first entry	75	65	56	50	61
with any entry	81	65	81	65	73
in matching all entries	71	70	70	58	67
Over/underconfidence	−1	11	−19	−5	−4
All					
Confidence in					
entry being in catalog	85	84	80	73	80
finding it if there	94	90	91	90	91
success (inferred)	80	76	73	66	73
Success					
with first entry	61	56	56	59	58
with any entry	87	60	79	70	74
in matching all entries	66	72	70	57	66
Over/underconfidence	−7	16	−6	−4	−1

search, resulting in greater overconfidence (mean = .35), there was no overall difference with the unstructured search (mean = −.01). Subjects succeeded 74% of the time and should have expected to succeed about 73% of the time (an estimate obtained, as mentioned, by multiplying their mean judged probability of the target being in the database by the mean judged conditional probability of being able to find it if it is there). Most of this reduction in overconfidence could be due to the greater chance of success with the unstructured problems (73% vs. 36%, where the former represents success at all and the latter represents success with the first entry). As noted earlier, the weak positive relationship between confidence and knowledge leads to a shift from over- to underconfidence as ease increases. The crossover point could be in the 75% success range observed here, eliminating any overall trend to over/underconfidence. Hindsight bias could have given an additional boost, pushing people toward feeling that they would have expected roughly the success that they experienced.

Overall, subjects were quite confident in the catalog's containing something related to their topic and in their ability to find it eventually if it were there. That confidence was rewarded by a 58% success rate on the first entry (compared with only 36% on the semi-structured searches that followed). For those who persisted, the success rate increased to 74%. That increase suggests that people can produce additional useful terms, even though the terms that they tried matched those in the system only 66% of the time. One "secret" to this success was that subjects changed goals in 10.4% of searches. Although these new goals were no more likely to meet success than the goals that were maintained throughout, the change may have released subjects from searches having little chance of success. Depending upon how one looks at them, these successes might be viewed as a sort of failure.

As before, there were few consistent differences across the groups in confidence or initial success. The group receiving the expert advice did, however, show the least overall success (and least improvement with subsequent choices), even though it was the most successful group in terms of ability to produce entries that matched those in the computers (unfortunately, those matching entries did not contain useful material).

The overall realism of subjects' expectations may have conferred some resistance to frustration, one sign of which was the rarity (two cases, to be exact) of subjects giving up their search after providing a first entry that failed to match. Subjects' lack of overconfidence was due more to doubts about the system (mean confidence = .80) than to doubts about their ability to use it (mean confidence = .91). Almost everything sought was, in fact, in the database. Some failures to find it were due to inability to think of the right synonym, which is clearly a user-system interface failure. Many others were due to more subject-centered errors, such as misspelling, misplacing words in a title, and mismatching authors and titles. Perhaps, however, it is easier to doubt the system than oneself. It is not clear how long-term experience with a system would affect subjects' ability to use it or their perceptions of its adequacy (and the realism those perceptions).

Evaluation of Advice

Subjects found all three instructions to be quite understandable (mean = 8.6 on 0–10 scale) and none to be particularly useful (mean = 3.8 on 0–10 scale). There were no differences in understandability ratings (range of means = 8.5 to 8.7). Subjects found the statement produced from other subjects' recommendations to be the most useful (4.3) and that from the library to be the least (3.1), although the differences were not large.

Abstract Categories Task

Purpose. Subjects' final task was to complete a questionnaire that had also been completed by subjects in an earlier series of experiments which differed from the present one in a number of respects. Its participants were recruited entirely through an advertisement in the University of Oregon student newspaper, hence were somewhat younger than the present subjects (mean age = 23 versus 30). Their experiment was conducted in a group setting, rather than individually, hence may have evoked somewhat less involvement. They completed but one information search task, rather than having first completed the present series of computer–interactive tasks with its element of tutoring. If the present intensive experience had some general effect on subjects' search strategies, then one might expect their performance on this task to be superior. If observed, such an effect might be attributed, in part, to the motivating effect of the individualized attention. As mentioned earlier, replication of that previous study in an individualized computer-interactive setting found no difference in response patterns [9]. An improvement might also be attributed to the greater life experience of the present subjects. However, as described below, the effects of greater knowledge should be discernible on the basis of results from earlier studies.

Results. Table 8 presents a statistical summary of present subjects' performance, contrasting it with that for the earlier group. The first three lines express the transparency of the categories (for these subjects) in three different ways. "Conditional" refers to the proportion of subjects choosing correctly on a round among those who had yet to do so. "Cumulative" refers to the proportion of subjects who had answered items correctly by the end of each round. "Proportion correct" refers to the proportion correct among

TABLE 8 Performance statistics on *Abstract* task

	Library Group			Original Group[a]		
	1st	2nd	3rd	1st	2nd	3rd
Transparency (proportion correct)						
Conditional	.584	.526	.333	.604	.647	.500
Cumulative	.584	.797	.858	.604	.844	.905
Metatransparency						
Proportion correct	.584	.220	.071	.604	.256	.078
Mean confidence	.729	.195	.067	.727	.187	.073
Over/underconfidence	.145	−.025	−.004	.123	−.069	−.005
Calibration	.035	.012	.004	.042	.008	.006
Resolution	.003	.007	.001	.006	.002	.000
n/responses	620	599	537	422	402	376

[a] Called the "elaborated subject partition" group in Fischhoff, MacGregor, and Blackshaw (1986).

all subjects, including those who had already chosen correctly (and would presumably not be making subsequent selections, unless they were asked to make their selections in advance—a strategy that produced better performance than sequential selection [7,9].

From all three of these interrelated perspectives, the two groups were fairly similar. After three choices, both groups had answered correctly in about 90% of cases. The slightly inferior performance of the library group may reflect either reduced knowledge of the material in the *Abstract* or a lesser tendency to share the perspective that guided the subjects whose judgments shaped the categories and their labels.

Previous research has shown that differences in transparency have a predictable effect on metatransparency [23]. Namely, in comparable tasks, subjects who are less successful will tend to have a poorer feeling for the extent of their own knowledge. However, the differences in transparency here are too small for there to be much of an associated metatransparency effect. As a result, any difference in metatransparency might best be attributed to subjects' immediately preceding experience. In any case, there was not very much. Subjects were similarly confident in their ability to locate items, as indicated by their mean probabilities. If subjects' expectations were realistic, then, over a long run, their proportion correct should equal their mean probability. The positive difference indicates overconfidence in first choices, which was similar in both groups. Confidence was so high here that there was little probability "left" for the remaining two choices, where subjects in both groups were slightly underconfident.

"Calibration" and "resolution" are statistics used, among other places, by the US National Weather Service to validate probabilistic forecasts [24,25]. The former is the mean squared difference between each probability response (e.g., .7, 1.0) and the associated proportion of correct answers (weighted by the number of responses involved). "Resolution" equals the variance in the proportions correct associated with different probability responses. It reflects people's ability to discriminate different,levels of knowledge. Calibration reflects people's ability to assign appropriate (absolute) levels of confidence to those levels. (Fuller expositions of these statistics and examples of representative values may be found in [23,26].) The values here represent average performance and, more important for present purposes, are quite similar for the two groups. Thus, whatever effect the present experience had on subjects did not extend to their responses to this seemingly related task.

Protocol Analysis

Background. Summary statistics like those in Tables 2–7 provide one kind of information regarding subjects' performance, showing the successes and expectations emerging as the end result of all the complex cognitive processes that are evoked by such search tasks. They can enable one to predict how well users of a system will do on future tasks, where they will experience frustration because of unrealistically high expectations, and where they will underutilize a system because of unrealistically low ones. They can show where users need help and how that help should affect them (e.g., to lower expectations, to sharpen the evaluation of how much they know). Those statistics are, however, less informative regarding how that help is to be provided. Telling people about potential problems and urging them to do better is an obvious form of advice. However, experience elsewhere suggests that information about cognitive difficulties is unlikely to be helpful unless accompanied by some relevant instruction in how to use one's mind more effectively [22,27].

Providing that kind of instruction requires some understanding of how people think without it and in what ways they are capable of manipulating their thought processes. Such understanding can be sought indirectly by inferring performance that has been observed in different situations. For example, overconfidence in the extent of one's knowledge might be due to failure to scrutinize one's favored beliefs critically enough. Evidence affirming this hypothesis may be found in the reduced overconfidence observed with subjects who list explicitly reasons why favored answers may be wrong [28].

An alternative to seeing how various manipulations affect the product of thought processes is to attempt to capture those processes themselves. Various techniques exist for measuring byproducts of those processes, such as eye movements and physiological arousal levels [29,30]. An appealing alternative is to have subjects describe their own thought processes. Such introspections have a long, troubled history in psychology [31,32,33,34,35]. Current thinking seems to afford them some credence, as long as they are collected concurrently with the thinking (rather than retrospectively) and are validated by some convergent operations.

Protocol Design. As a supplement to our summary performance measures, we undertook a combination of having subjects describe their thoughts and observing associated behavior. Because any behavior, whether explicitly verbalized or simply observed, can be described in many different ways, some theoretical basis is needed for instructing subjects regarding what to reveal and for coding what emerges. Given our interest in decision-making processes, we asked subjects in the semi-structured search tasks to describe their goals, their favored alternatives, and their confidence levels (in quantitative terms), as well as whatever else they chose to relate. A more fully structured search could have requested other elements of a comprehensive decision-making process, such as the full set of (seriously considered) alternatives and the utility (or disutility) of different search outcomes and operations (e.g., computer costs, the value of time spent searching). Our experience in preliminary tests suggested that any further structure would have constituted an imposition that substantially changed how subjects approached the task. Although that might be an interesting manipulation, it would be a different enterprise than asking subjects to produce a somewhat ordered version of their natural thought processes. The semi-structured search instructions seemed to achieve about the right balance. Even the unstructured search was not entirely free of pressure. The instructions suggested similar concerns (although without requiring their description as directly) and followed tasks that elicited both general and specific expectations. The gist of the instructions emphasized reporting behavior and functional thoughts, rather than emotions. The use of a timer in the pretest, however unobtrusive, was one of many other contextual features that may have suggested to subjects how to approach and describe their tasks.

Attempts to trace information-search processes with other central theoretical concerns may be found in several reports [11,17,18,32,33,36,37,38,39,40]. The procedures in these studies emphasize topics such as the mental models that database users and compilers have of a substantive domain and stylistic differences in how users approach their task.

Coding Behavior. Two interrelated schemes, both reflecting a decision-making perspective, were developed for characterizing subjects' observed behavior. The *decision process* scheme attempts to describe behavior in terms of the elements in a normative decision-making process (that is, one following decision theory's dictates regarding how decisions *should* be made). The *action* scheme attempts to describe behavior as emerging from or serving such a decision-making process. The decision process scheme is meant to be general enough to be used with any information search task. The action scheme is adapted to the particular actions relevant to this database. Both schemes are still under development.

Table 9 shows the current version of the decision-making process scheme. It distinguishes five kinds of behavior that appear in most normative theories of decision making: iden-

TABLE 9 Coding scheme for decision-making processes in information retrieval.

Stage	Expression
1	*Goal*
	a) Searches for a goal without specifying any
	b) States goal in own words
	c) Justifies the selection of that goal
	d) Explicitly evaluates goal
	e) Defines the goal in a different way
	f) Chooses a different goal
	g) Restates original goal
2	*Determining option set*
	a) Searches for options without specifying any
	b) Browses for options (by talk or behavior)
	c) Talks aloud options
	d) Selects an option
	e) Types an option on the computer
	f) Writes an option on paper
	g) Justifies option chosen
	h) Talks about options that are elsewhere
3	*Evaluating options*
	a) Explicitly evaluates options (before or after selecting them)
	b) Explicitly rejects options
	c) Reconsiders options
	d) Abandons search
	e) Successful outcome
	f) Person claims the outcome as successful although they have not reached the goal
4	*Recognition of uncertainty*
	a) Mentions confidence level
	i. Information is in the system
	ii He/she is able to find it
	b) Comments on problems encountered
	c) Comments on amount of time taking or likely to take
	d) Mentions own frustration or lack of motivation
	e) Refers to possible outcomes of own actions
	f) Refers to what might happen regardless of own actions
	g) Comments on own lack of knowledge
	h) Seeks help
5.	*Evaluates exercise*
	a) Notices possible and actual mistakes
	b) Fails to notice own mistakes
	c) Blames self (inability, mistakes or lack of motivation)
	d) Blames system (software, hardware)
	e) Indicates awareness of how to avoid problems
	f) Indicates awareness of how to recover from mistakes and difficulties
	g) Evaluates own strategy
	h) Evaluates system

tifying goals, determining an option set, evaluating options, recognizing uncertainty, and evaluating the overall process. Each is then operationalized in six to eight observable kinds of behavior. In formulating these categories, we tried to be sufficiently detailed to describe think-aloud verbal proctocols and nonverbal search behaviors, but broad enough to ensure reliability among coders.

(1) *Goal.* The seven codes under this heading include the person's search for a goal, any justification or evaluation of that goal, definition of the goal, and choice of a new goal.
(2) *Determining option set.* Coding covered looking for options, talking, writing and typing options, selecting and browsing options that arose on or off menu.
(3) *Evaluating options.* These six codes describe a person's evaluation of options prior to, during, and after completion of a search.
(4) *Recognition of uncertainty.* Under this category are included expressions of confidence and the probability of success, comments on time consumption and difficulties, awareness of possible outcomes, and efforts to seek help.
(5) *Evaluates exercise.* Eight codes were identified: awareness or ignorance of mistakes, blame directed at self or others, knowledge of how to avoid or to recover from problems, and explicit evaluations of the person's search strategy or of the system.

The left-hand side of Figure 1 shows the current version of the action scheme. It orders actions by rough proximity to the initiation and termination of the search. The top entries are overt actions associated with goal setting, followed by those actions involved in explicit consideration of options (i.e., possible locations) and in the decision to pursue one (in a system that requires options to be examined serially). The next entry is the system's response, either providing no match to the entered option or offering a set of headings for subjects' consideration. In the former case, subjects can choose among the options in Stage 4. Unlike the choice among possible entries, this set of alternative actions is quite heterogeneous, including even "offline" alternatives (e.g., seek help). If there is a match but at too high a level of generality, then the system will show headings within the chosen category within which titles may be sought. In this system, only ten headings (or titles) can be shown at a time, with an indication of how many others are available. Once a title has been chosen, the system provides the choice of bibliographic and location information, which can provide additional cues to the book's contents. The final stage is termination of the search, with different possible degrees of satisfaction, as judged by searcher or observer.

Data Analysis. Each scheme can be used independently to describe (a) the percentage of all search behaviors that fall within the scheme and (with somewhat greater difficulty) their centrality to the information–retrieval process; (b) the relative frequency of the different kinds of behavior covered by each scheme; (c) the sequencing of different behaviors within the search process; (d) the relationship between types of behavior and aspects of performance (e.g., confidence, success, realism, particular errors). In combination, the schemes can be used to describe the kinds of

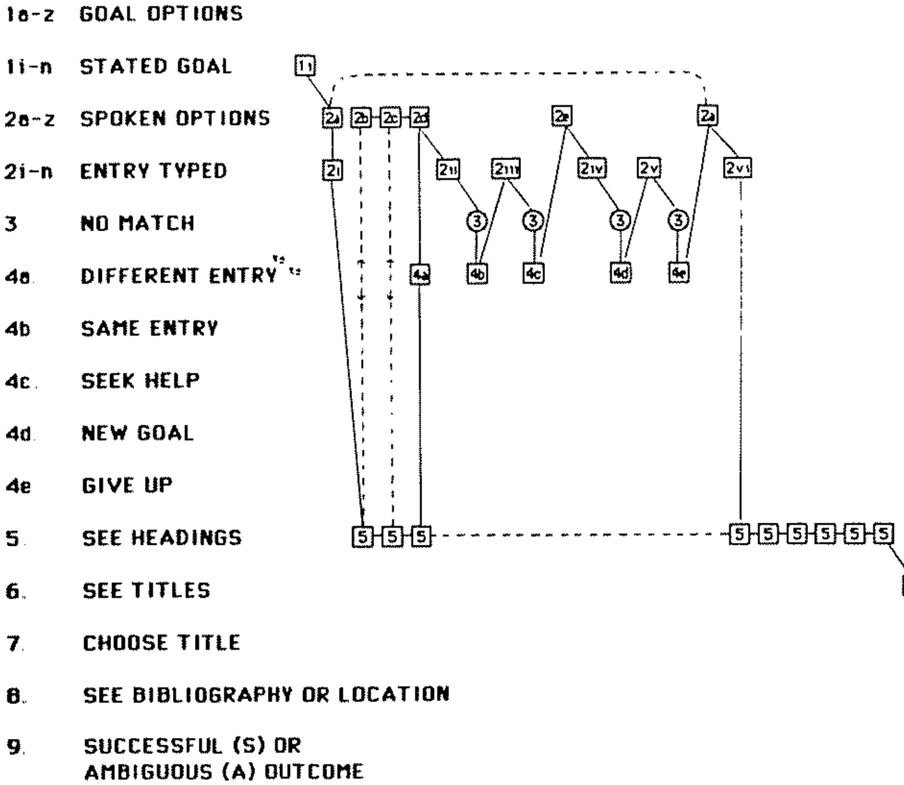

FIG 1. Graphic representation of subject 11Es semistructured search for "another book on a similar topic," following the transcript in Table 10. Squares indicate overt actions by subject. Circles indicate actions by database system. Actions are ordered from left to right from beginning of search.

deliberations that precede and follow different actions. All of these questions can be asked conditional on various experimental manipulations.

Some of the questions that can be asked of these data include: (a) How extensive are subjects' deliberations regarding alternative search strategies? (b) Are more protracted initial deliberations associated with better performance? (c) If so, what instructions or task structuring will encourage such deliberations? (d) Does devoting more time to clarifying search goals enhance the effectiveness of the time spent on clarifying search alternatives? (e) How does such clarification affect the probability of giving up after failure to match an entry or after finding nothing suitable under it? (f) Does publicly stating goals enhance or reduce search success, flexibility and persistence? (g) Are public statements of confidence (or other evidence of considering uncertainty) associated with better performance? (h) Subjects had more realistic expectations when they stated in advance what they would try next if a favored option proved unsuccessful [7]. Would this result be replicated in the present context? If so, what insight can be provided into the source of its effectiveness, particularly with regard to mechanisms that might be invoked to improve searches? (e.g., is there a greater tendency to consider sources of uncertainty? Is there more refinement of goals?) (i) Does providing justifications for choices (code 2g) improve performance (by forcing commitment to a particular way of thinking)? (j) Which sequences of events (or systems, or experimental manipulations) increase the likelihood that subjects will attempt to evaluate their search experience, looking for some general lessons to derive from it? (k) What are the precursors to attributing successes and failures to oneself and to the system? How realistic do those attributions seem to an outside observer?

One intent of each such question is to provide an accurate descriptive account of decision making in online search. That account would facilitate predicting search behavior and make some contribution to the general literature on decision making, which has not looked extensively at this sort of sequential, interdependent decision. A complementary goal is to help subjects make better decisions (and searches) through identifying problems and the natural thought processes that can (and must) be built upon to effect changes.

An Example. Table 10 and Figure 1 show the application of these analytical schemes to one representative example. In it, the eleventh subject in the group receiving the expert advice conducts a semistructured search for a book on a topic similar to one that she had enjoyed in the past. The left-hand side of Table 10 records all observed actions as expressions of the decision process scheme, while its right-hand side codes them in terms of the action scheme. Figure 1 provides a graphic representation of the search in terms of the proximity of these actions to its initiation or termination.

In it, the subject begins with what proves to be an unduly vague goal (a book about New York). Such generality allows great confidence that there will be something on the topic. However, the resulting list of headings is, on the one hand, so extensive that the subject browses only a portion of it (30 headings) and, on the other hand, formulated in terms that leave the subject unsatisfied with any of them. As a result, the subject is left musing about options that might be more adequate (2b–2d). When one is chosen, it is mistyped. Apparently not realizing that the intended entry has not actually been tried, the subject moves on to different topics (steps 14–16). The next entry is typed correctly. When it fails to match, the subject changes the topic, rather than reformulating it (steps 17–20). This failure is met by a reformulation, with the subject trying to guess at the system's semantics, rather than seeking a synonym (steps 22–24). The next failure produces reversion to the original actions (steps 25–36). Finally, a title of doubtful usefulness is chosen.

Figures 2 and 3 present graphic representations of this subject's other two semi-structured searches. A full reading requires the accompanying decision-process protocol. However, these figures alone readily show quite different patterns. With "leisure" (Figure 2), apparently a better defined topic for this subject (and for others), the subject is able to mull goals without recourse to the system. A match is found readily and the headings reviewed until an appropriate one is found and a title located in it. The third search (Figure 3) begins similarly. However, the subject is unsatisfied with the title and returns to the headings and eventually to the superordinate entry. The title found this time proved more adequate, but not before the bibliographic information was examined as a double check.

As mentioned, we believe that characterizations like Table 10 and Figures 1–3 can provide useful insights into subjects' decision-making behavior. They offer sufficient detail to allow underlying processes to emerge, but in a sufficiently standardized way to allow comparisons and statistical summaries. Beyond coding the individual actions, it may be possible to identify recurrent response patterns and the conditions that prompt them. For example, Figure 1 might present elements that are common to situations where people fail to clarify their goals in advance, or where a technical mistake is misinterpreted as a substantive message from the system about its internal organization.

General Discussion

The present results may be interpreted on several levels. The summary performance statistics allow prediction of system utilization by individuals resembling those in the study. They show how likely users are to attain different kinds of goals, how long successful and unsuccessful searches are likely to take, and how such outcomes are related to people's reliance on different library resources (Table 1). Such estimates could provide precise design specifications regarding issues such as the number of terminals to provide, the effects of various changes in system response time, and the value of maintaining a parallel card catalog.

TABLE 10 Subject #11E's semi-structured search for a book on a similar topic.

	Decision Behavior	As Coded on Visual Model
1.	States goal of search: a book on New York	[1i]
2.	Talks aloud an entry option: subject New York	[2a]
3.	Types su = New York City on the computer keyboard	[2i]
4.	Is 100% confident that it will be under that heading	[100]
5.	Browses first 10 headings in the catalog	[5]
6.	Browses second 10 headings in the catalog	[5]
7.	Reconsiders options and talks aloud another entry: New York lifestyles	[2b]
8.	Browses a further 10 headings	[5]
9.	Talks aloud another option: night life in New York	[2c]
10.	Selects another entry option	[4a]
11.	Redefines entry option	[2d]
12.	Selects entry but mistypes	[2ii]
13.	Computer responds that there is no matching heading	[3]
14.	Reconsiders entry option	[4a]
15.	Types su = (entertainment) New York City	[2iii]
16.	Computer responds that there is no matching heading	[3]
17.	Reconsiders entry option	[4a]
18.	Selects another entry option: night clubs	[2e]
19.	Types on the computer su = New York (night clubs)	[2iv]
20.	Is 50% confident that it will be under that heading	[50]
21.	Computer responds that there is no matching heading	[3]
22.	Reconsiders entry option	[2v]
23.	Types on computer su = New York City night clubs	[2iv]
24.	Computer responds that there is no matching heading	[3]
25.	Reconsiders entry option	[4a]
26.	Selects first entry option: New York City	[2ar]
27.	Types a shorter version of the first option: su = New York	[2vi]
28.	Browses the same first 10 headings call up after 2i	[5]
29.	Browses the next 10 headings	[5]
30.	Browses the next 10 headings	[5]
31.	Browses the next 10 headings	[5]
32.	Browses the next 10 headings	[5]
33.	Browses the next 10 headings	[5]
34.	Browses the next 10 headings	[5]
35.	Selects titles for the heading: music hall	[6]
36.	Selects one title: "They all sang"	[7]
37.	Stops searching, unsure whether it is the desired book but choosing to look at it	[A]

More general design guidance can be found in the identification of current system features that appear to be problematic. The obscurity of the Library of Congress subject classification used by this system would seem to be one. Its cumbersomeness is evidenced by the seemingly low probability (58% in Table 6) of attempted entries even being recognized by the system and the similar conditional probability of recognized entries producing titles that subjects judged to satisfy even the somewhat modest criterion of being on the target topic. The incidence of ambiguous successes may indicate imprecision in subjects' thinking or the lowering of standards in response to a recalcitrant system (as seems to have been the case in the example of Figure 1).

More general still is evidence of the thought processes evoked as subjects decide what to look for, where to look, and how far to trust what they have found. Although most general, such evidence also provides a point of departure for initiating more fundamental design changes. For example, a focus here has been on the realism of subjects' expectations regarding their search success. We found a consistent modest positive correlation between predicted and experienced success with a overall tendency toward overconfidence except for the easiest tasks (and for the reconstructed expectations of success on the unstructured search). Although this pattern replicates that observed in studies of confidence assessment conducted in other settings, the relationship between confidence and knowledge seemed particularly weak with the unstructured and semi-structured searches here. Some of that lack of meta-transparency seems due to lack of goal clarity and some to particularly confusing properties of the choice options offered by the present system. Subjects often seemed surprised (and somewhat unnerved) by the discovery that their choices were not even recognized and, when recognized, appeared to have some alternative meaning to the system.

When user and system mismatch, one might try to change either, capitalizing on whatever is known about the thought processes involved. For example, we observed poor

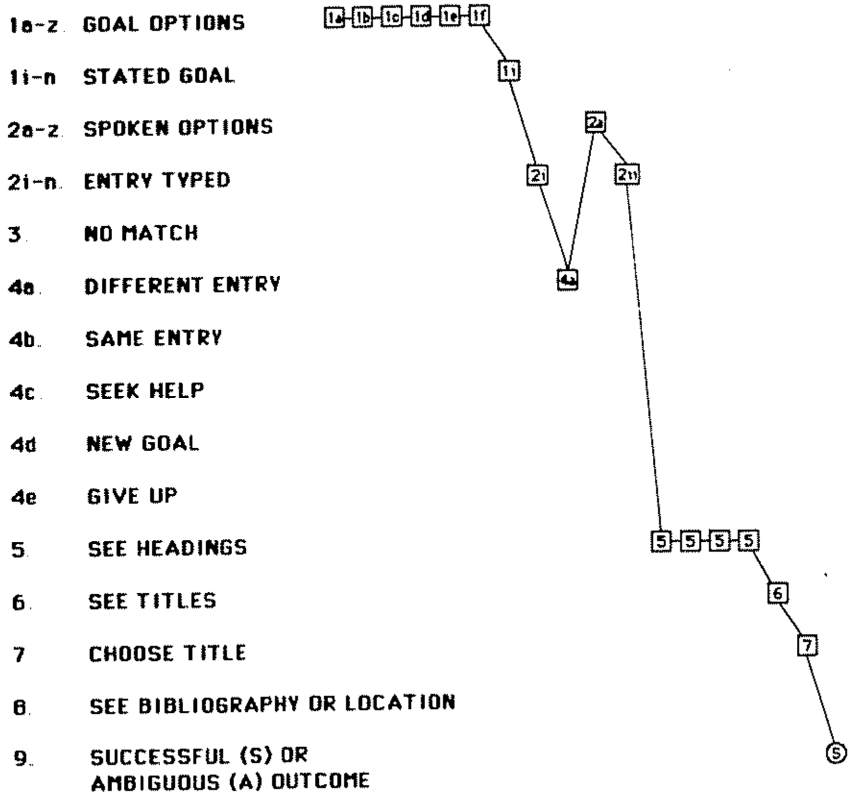

FIG 2 Graphic representation of subject 11Es semi-structured search for a book on leisure

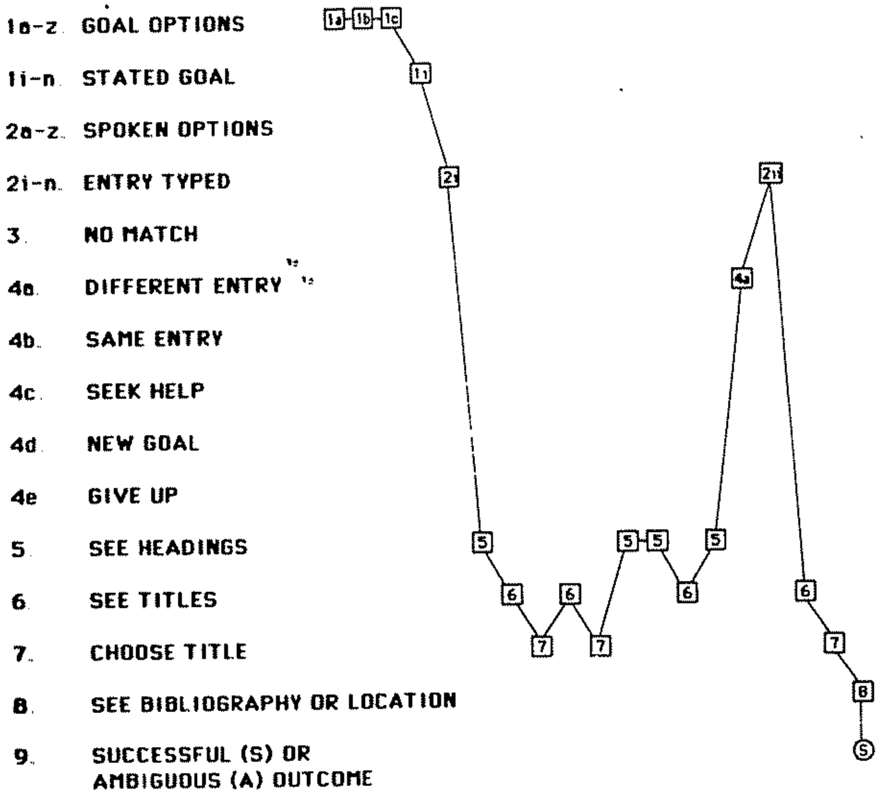

FIG 3 Graphic representation of subject 11Es semi-structured search for two books by the same author on a work topic.

transparency and metatransparency with a predecessor to the set of categories used in the present *Abstract* categories task [7]. Having no access to users' mental representations of this domain (which are likely to be both diverse and poorly articulated [41], we adopted several standard psychometric techniques for asking potential users first to create categories from the chapters of the *Abstract* and, then, to produce labels for their creations. The result was a substantial improvement in transparency (to the level seen in the right-hand side of Table 8) which carried with it a corresponding improvement in metatransparency (as would be expected from the generally observed relationship between these two aspects of performance).

With so large, diverse, and fixed a system as that of the Library of Congress, the opportunities for such redesign are limited. What might be changed is the face it presents to users. Here, we attempted to augment the system with 1–2 pages of advice. The nature of the system forced this advice to address general issues of system design and operation, rather than the particulars of its contents. Subjects seemed to appreciate the effort. However, they claimed that it was of little use, an appraisal that was borne out by the similar performance of the four experimental groups. Conceivably, better instructions, a larger set of subjects and tasks, or more refined data analysis techniques would reveal more demonstrable effects (Appendix B shows an interim report prepared for the staff of the Eugene Public Library, based on the data of this report, an impressionistic analysis of the decision process and action protocols, and unsystematic observations from the experimental sessions).

The relatively weak relationship between expectations and performance indicates that users need to learn not only about the system, but also about themselves in relation to it. Describing the performance of other users would be a direct approach to providing such insight. Whether people can use such summary information would be an empirical question. If not, one might attempt to describe the thought processes leading to problems and how users might redirect their own thoughts. It may be that such knowledge is better conveyed by doing than by telling. Intensive personalized feedback has proven effective in improving the realism of expectations on other tasks [19,26]. Perhaps a training module or individualized performance tracking system could be incorporated in a computerized system. The present experiment, which was not designed as a learning experience, had the effect of lowering subjects' estimates of the probability of finding a book that they wanted from 87% at its beginning to 79% at the end. Although we have no direct evidence regarding typical searches, such lowered expectations seem in keeping with the performance observed here. We look to the forthcoming analysis of the protocol data for hypotheses and evidence regarding effective manipulations.

Acknowledgment

This research was supported by the National Science Foundation under Grant IST 8312482 to Perceptronics, Inc. Its support is gratefully acknowledged. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Foundation. We thank the staff of the Eugene Public Library for their continued help, without which the experiment would have been impossible. We also thank Mark Layman, Leisha Sanders, Geri Hanson, and Steven Baker for their usual skilled assistance.

Appendix A: Three Advice Statements

Advice Given by the Eugene Public Library

YOU MAY FIND MATERIALS IN FIVE DIFFERENT WAYS

BY AUTHOR	EXAMPLE	AU=MICHENER JAMES
BY TITLE	EXAMPLE	TI=HAWAII
BY AUTHOR AND TITLE	EXAMPLE	AT=MICHENER/HAWAII
BY SUBJECT	EXAMPLE	SU=HAWAII
BY CALL NUMBER	EXAMPLE	CN=919.69 C226

To make your inquiry, use this terminal by following the instructions given. Type your request using the keyboard provided, making sure that when you have completed typing your inquiry, you press the key marked RETURN to send your request to the computer.

To search the catalog by AUTHOR, first type AU= then type the author's last name, optionally followed by first name or initial. The more information you provide, the faster the response to your search will appear.

If the computer finds more than one AUTHOR name that matches your request, it will list the matching names on the screen. Each will be identified by a number (called the LINE NUMBER) listed beside it on the left side of the screen. The number shown in parentheses to the right of the author's name indicates the number of titles in the collection written by that author.

To find available titles written by a specific author, FOLLOW the INSTRUCTIONS provided on the screen below the list of matching author names.

To search the catalog by TITLE, first type TI= then type in the title that you are looking for. You need not type the entire title, as the computer will perform a TRUNCATED search and display all the titles in the catalog that match your search. However the more information you provide, the faster the response to your search will appear.

Some titles would normally result in many unnecessary matches being displayed... For example if you enter the search TI=ROOTS the computer will find all the titles that begin with the word ROOTS. If the title you are looking for is a single word like ROOTS, you can get a faster response by adding the qualifier/E to the end of your search... i.e., T=ROOTS/E will only look for those titles which match your search exactly.

To search the catalog by SUBJECT, first type in SU= followed by the subject that your are looking for. The computer will display all the subject headings in the catalog that match your request... the more detailed you make your search the faster the computer's response will be.

If the computer finds more than one SUBJECT HEADING that matches your request, it will list the matching headings on the screen. Each will be identified by a number (called the LINE NUMBER) listed beside it on the left side of the screen. The number shown in parentheses to the right of each heading indicates the number of titles in the collection that refer to that subject.

To find out which titles are related to a specific subject, FOLLOW the INSTRUCTIONS provided on the screen below the list of matching subjects.

Some subjects may result in more matching entries than you want . . . for example, if you search by SU=TENNIS, the computer will list all subjects that begin with the word TENNIS. To reduce the search time, if you enter your request followed by the qualifier/E, the computer will look only for those subjects that match your request exactly

Sometimes when the computer lists a subject, it may tell you to SEE or SEE ALSO some other heading that the librarians consider may be appropriate to your search.

NOTE: Only valid LIBRARY OF CONGRESS subject headings are used by this library.

Advice from Others Who Have Used the Library's Computer Catalog

Go directly as possible to what you want. Just keep trying. Go from the specific to the more general; it takes too long the other way. If you get it right, you save a lot of time. Generalize your search if starting with the specific doesn't work. Start specific and then scale up unless you've got a feel for the right category from the start in which case look for a general category that is large enough to definitely be there and small enough that it won't be a haystack for the needle you're looking for. From there, narrow down your search.

Try to balance intuition, instinct and being scientific about it. If either one is getting frustrating, then try the other way.

Don't be put off if you don't succeed first of all. You may have to alter your usual approach to fit in with the program. Don't take it for granted that becausë you don't find it under the first heading that it's not there. You have to use your head. Exhaust the heading titles to find what you want. Don't give up too easily. You get better and more confident as you go on.

It's better if you know what you're looking for before you start. Just use one word for subjects. Read the screen. Things that go wrong are due to your stupid mistakes. Follow what it says on the screen. Check your spelling and go back and check for errors. Don't forget your = sign or to put the first name last. Pay attention to abbreviations. Have patience. This is primitive software. It's slow. You have to keep going back and forth. Have patience.

Advice from Researchers

One way to think about this task is that it's like making a series of decisions. At each stage, you want to make the best choice among a set of alternatives, each of which might contain the book(s) that you want. For example, your initial choices are author, title, subject, and author title. If you chose "subject," then your choices would be any of the categories into which your book(s) could be placed. If you picked a very general subject like "animal," then you would have further choices under it. And so on.

Ideally, you would like to make choices that will have a high probability of getting the book(s) in a short period of time. Here are some pieces of advice regarding how to make efficient decisions.

- Before starting, think thoroughly about all the different ways in which your book might have been categorized.
- Reduce the set of possibilities to a few alternatives, which you can turn to if your first choice doesn't work out.
- In making that first choice, remember that a broadly defined option is more likely to contain what you want, but also will leave you further from your goal.
- When you've settled on a first choice, before trying it, give an extra moment's thought to why it might not work out. (e.g., Is that really what you want? How else might the library have categorized it? If it doesn't work out, where will you look next?).
- If a particular choice doesn't work out, think about what that tells you about the organization of the system as a whole. Ask how surprised you were at what happened.
- Before making your choice, make a prediction regarding the chances that it will work out. Over time, try to get a feeling for how realistic your expectations are. If you find that you're too confident of succeeding, try to think harder about possible surprises.
- If you are uncertain about where to look, think about your choice as being partly an experiment. Ask which choice will tell you the most, even if it turns out to be wrong.

Appendix B:

Summary of Difficulties Experienced by 60 Users of the On-line Catalog and Recommendations Discussed at the 5/29/86 Meeting of Departmental Heads at the Eugene Public Library

Difficulties Experienced by Users

(1) Problems in being able to identify the correct match for subject headings.
(2) Spelling weaknesses.
(3) Difficulty in being able to assess whether subject headings were relevant to the initial goal.
(4) Slowness in learning the potential of the system and using it creatively.

Feedback from Users

(1) Complaints about the apparent lack of logic in the system:
   i) Many double listings of authors, titles and subject headings.

ii) Titles not alphabetized.
iii) No ordered sequence to call numbers.
iv) Books listed that they knew were missing.
v) Books listed with "no holdings".

(2) Comments on the clumsiness or inefficiency of the system:
i) Constant return to title page.
ii) Only 10 headings at a time.
iii) Inconsistency of numbering for options.
iv) Two steps required to access title
v) Too slow a display.
vi) Having to wait for the screen to print before typing requests.

(3) Problems in using the subject catalog:
i) Having to second guess the compilers of the Library of Congress subject headings.
ii) Insufficient See Alsos.
iii) The computer indicating there was no match when they had already taken books out on that subject (i.e. they had used the "incorrect" term).
iv) Wanting the title catalog to be accessed at the same time as the subject catalog, in order to include more common subject names.

(4) Irritation and embarrassment at the audio signal accompanying input errors.

Recommendations

(1) Improvements to the System and in its Use:
i) Continue to add more See Alsos and cross references to the subject catalog.
ii) Have more subject heading books readily available.
iii) Facilitate two levels of users: giving the more advanced information on how to avoid returning to the title page, how to call up status without returning to the previous screen, how to go across categories in a search and any other shortcuts used by the librarians.
iv) Give demonstrations to patrons on the best ways to use the catalog; give people alternatives such as typing in title when the subject term has no match, etc.
v) Have a member of staff available at stated times close to the terminals watching for difficulties and helping people.
vi) Provide seating at some of the terminals; even though this may increase the time spent at the terminals, it might also increase the success rate.
vii) Reconsider whether the public terminals should take more than 10 names into memory at a time.
viii) Eliminate the beep that announces incorrect entries.
ix) Standardize the numbering system for the options; if an option is unavailable, leave the number out.

(2) Promotion of the computer catalog:
i) As many of the criticisms by users arise out of the catalog being an inventory system, it is preferable to state clearly that this is the system used by the librarians to catalog books as they arrive. Promote the fact that the public has access to the same information, rather than allowing people to believe that it is a custom-made online catalog for public use.
ii) One of its major strengths over a card catalog or even some on-line systems is that it gives the status of each holding. In our experiments, 21 people (35%) did not use the Availability information and in consequence were only 62% confident that they would find a book on the shelves after locating its call number on the computer. Nine people (15%) used the Availability information some of the time and reported a confidence about finding the books on the shelves of 71%. In contrast, 30 people, (50%) said they looked at Availability and reported a 91% success rate in locating books on the shelves. This option needs to be actively advertised and demonstrated to users.

References

1. Raiffa, H. *Decision Analysis*. Reading, MA: Addison Wesley; 1968.
2. von Winterfeldt, D.; Edwards, W. *Decision Analysis and Behavioral Research*. New York: Cambridge University Press; 1986.
3. Blair, D. L. "Searching Biases in Large Interactive Document Retrieval Systems," *Journal of the American Society for Information Science*. 31:271–277; 1980.
4. Bookstein, A.; Swanson, D. "A Decision Theoretic Foundation for Indexing," *Journal of the American Society for Information Science*. 26:45–50; 1975.
5. Cooper, W. S. "Indexing Documents by Gedanken Experiments." *Journal of the American Society for Information Science*. 29:107–119; 1978.
6. National Research Council. *Priority Mechanisms for Toxicity Testing*. Washington, DC; 1983.
7. Fischhoff, B.; MacGregor, D. "Calibrating Databases," *Journal of the American Society for Information Science*. 37:222–233; 1986.
8. Fischhoff, B.; MacGregor, D.; Blackshaw, L. *Creating Categories for Databases* (Report No. 86-9). Eugene, OR: Decision Research; 1986.
9. MacGregor, D.; Fischhoff, B.; Blackshaw, L. *Search Success and Expectations with a Computer Interface* (Report No. 86-10). Eugene, OR: Decision Research; 1986.
10. U.S. Department of Commerce. *Statistical Abstract of the United States*. Washington, DC; 1983.
11. Bates, M. J. "Factors Affecting Subject Catalog Search Success," *Journal of the American Society for Information Science*. 28:161–169; 1977.
12. Bookstein, A. "Probability and Fuzzy-Set Applications to Information Retrieval," *Annual Review of Information Science and Technology*. 20:117–151; 1985.
13. Matthews, J. R.; Lawrence, G. S.; Ferguson, D. K., Eds. *Using Online Catalogues: A Nationwide Survey*. New York: Neal Schuman; 1983.
14. Hasher, L.; Zacks, R. "Automatic Processing of Fundamental Information: The Case of Frequency of Occurrence," *American Psychologist*. 39(12):1372–1388; 1984.
15. Dumais, S. T.; Landauer, T. K. "Describing Categories of Objects for Menu Retrieval Systems," *Behavioral Research Methods, Instruments, and Computers*. 16(2):242–248; 1984.

16. Cochrane, P. A.; Markey, K. "Preparing for the Use of Classification in Online Cataloging Systems and in Online Catalogs." *Information Technology and Libraries*. 4:91–111; 1985.

17. Croucher, C. "User Studies and Interface Design." *Programs* 20(2):211–214; 1986

18. Fenichel, C. H. "The Process of Searching Online Bibliographic Databases: A Review of Research." *Library Research*. 2:107–127; 1980.

19. Bates, M. J. "The Fallacy of the Perfect Thirty-Item Online Search." *RQ*. 24(1):43–50; 1984.

20. Lichtenstein, S.; Fischhoff, B. "Do Those who Know More also Know More about how much They Know? The calibration of probability judgments." *Organizational Behavior and Human Performance*. 20:159–183; 1977.

21. Fischhoff, B. "Hindsight ≠ foresight: The Effect of Outcome Knowledge on Judgment under Uncertainty." *Journal of Experimental Psychology. Human Perception and Performance* 1:288–299; 1975.

22. Fischhoff, B. "Debiasing." In: D. Kahneman, P. Slovic and A. Tversky, Eds. *Judgment Under Uncertainty: Heuristics and Biases*. New York: Cambridge University Press; 1982.

23. Lichtenstein, S.; Fischhoff, B.; Phillips, L. D. "Calibration of Probabilities: State of the Art to 1980." In: D. Kahneman, P. Slovic and A. Tversky, Eds. *Judgment under Uncertainty: Heuristics and Biases* New York: Cambridge University Press; 1982.

24. Murphy, A. H.; Brown, B. G. "A Comparative Evaluation of Objective and Subjective Weather Forecasts." *Journal of Forecasting*. 3:361–394; 1984.

25. Murphy, A. H.; Winkler, R. L. "Probability Forecasting in Meteorology." *Journal of the American Statistical Association*. 79:489–500; 1984.

26. Lichtenstein, S.; Fischhoff, B. "Training for Calibration." *Organizational Behavior and Human Performance*. 26:149–171; 1980.

27. Fischhoff, B.; Bar–Hillel, M. "Diagnosticity and the Base-Rate Effect." *Memory and Cognition*. 12(4):402–410; 1984.

28. Koriat, A.; Lichtenstein, S.; Fischhoff, B. "Reasons for Confidence." *Journal of Experimental Psychology: Human Learning and Memory*. 6:107–118; 1980.

29. Atkinson, R. C.; Herrnstein, R. J.; Lindzey, G.; Luce, R. D., Eds. *Stevens' Handbook of Experimental Psychology* (2nd ed.). New York; Wiley; In press.

30. Sokols, D.; Altman, I., Eds. *Handbook of Environmental Psychology*. New York: Wiley; 1987.

31. Ericsson, K. A.; Simon, H. A. *Protocol Analysis: Verbal reports as data*. Cambridge, MA: MIT Press; 1984.

32. Fidel, R. "Towards Expert Systems for the Selection of Search Keys." *Journal of the American Society for Information Science*. 37(1):37–44; 1986

33. Hoc, J. M.; Leplat, J. "Evaluation of Different Modalities of Verbalization in a Sorting Task." *International Journal of Manmachine Studies*. 18:283–306; 1983

34. Nisbett, R.; Wilson, T. "Telling More than We Can Know." *Psychological Review*. 84:231–259; 1977.

35. Svenson, O. "Process Descriptions of Decision Making." *Organizational Behavior and Human Performance*. 23:86–112; 1979.

36. Belkin, N. J. "Cognitive Models and Information Transfer." *Social Science and Information Studies*. 4:111–129; 1984.

37. Fidel, R. "Individual Variability in Online Searching Behavior." *Proceedings of the ASIS Annual Meeting*. 22:69–72; 1985.