Information Processing & Management Vol. 23, No. 5, pp. 419–432, 1987
Printed in Great Britain.

0306-4573/87 $3.00 + .00

SEARCH SUCCESS AND EXPECTATIONS WITH A COMPUTER INTERFACE*

DONALD MACGREGOR, BARUCH FISCHHOFF and LYN BLACKSHAW
Decision Research, A Branch of Perceptronics, 1201 Oak Street, Eugene, OR 97401

(*Received* 22 *July* 1986; *accepted in final form* 24 *March* 1987)

Abstract—People's expectations for success can affect their use of information sources in a variety of ways, including their willingness to search at all, their satisfaction (or frustration) with the success that they encounter, and their confidence in the completeness of their search for specific items. An earlier study using a pencil-and-paper format found people to be overconfident in their ability to locate various items in two entry-level menus to *The Statistical Abstract of the United States*. In addition, their performance was considerably better with a broad menu, comprised of the 33 chapters in the *Abstract*, than with a narrow one, comprised of 8 superordinate categories. The present study transferred this task to a computer-interactive format. Surprisingly, neither the transfer itself nor the introduction of performance feedback affected the realism of subjects' expectations. A review of the two studies, involving 481 subjects in all, in the context of the general psychological literature on confidence assessment, provides some suggestions regarding the design of interfaces for computerized data bases.

As computers permeate the home and workplace, the cognitive skills needed to use them become increasingly important. Some of those skills are relatively novel, such as developing mental models of how diverse system components interact when following complex formalized procedures (e.g. [1–3]). Others are refinements of everyday tasks, such as the need to think in precise Boolean terms to query computerized data bases (e.g. [4–6]). For still others, the tasks remain the same, but the skills need to be exercised in a new substantive domain, such as learning the names of objects and procedures that have been labeled by someone from a different subculture (e.g. [7–9]).

One familiar skill that needs to be exercised regularly with computerized systems is evaluating the chances that a procedure will be successful. On the basis of such expectations, users can determine whether to continue with the procedure, whether to ask for help, how much to invest in the effort, and how to plan for possible failures. As Bookstein [10] notes in the specific context of computerized data bases, "Uncertainty and incompleteness [are] intrinsic to both indexing and retrieval" (p. 122). Without a realistic assessment of uncertainties, users may overestimate their chances of success, leading to unwise investment of resources, premature frustration with systems that have failed to fulfill promises that they never made, uncritical acceptance of incomplete or erroneous system products, and inadequate attention to the sort of premonitors of failure that could be exploited to improve subsequent usage. Underestimating one's chances of success can lead to complementary difficulties.

Cognitive psychology has devoted considerable attention to the determinants and appropriateness of people's expectations [11–15]. This research typically finds that people have no difficulty expressing their confidence in succeeding at a task in quantitative terms (e.g., odds, probabilities), that people tend to be more knowledgeable when they are more confident, but that their absolute level of confidence is not a reliable indicator of their absolute level of knowledge.

Figure 1 shows one very common pattern of results. As confidence (indicated by judged probability of success) increases, so does knowledge. However, the rate of change

*This research was supported by the National Science Foundation under Grant SES 8312482 to Perceptronics, Inc. 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 National Science Foundation. Our thanks to RickReed, Mark Layman, Leisha Sanders, and Sonny Eberts for help in various stages of this process.

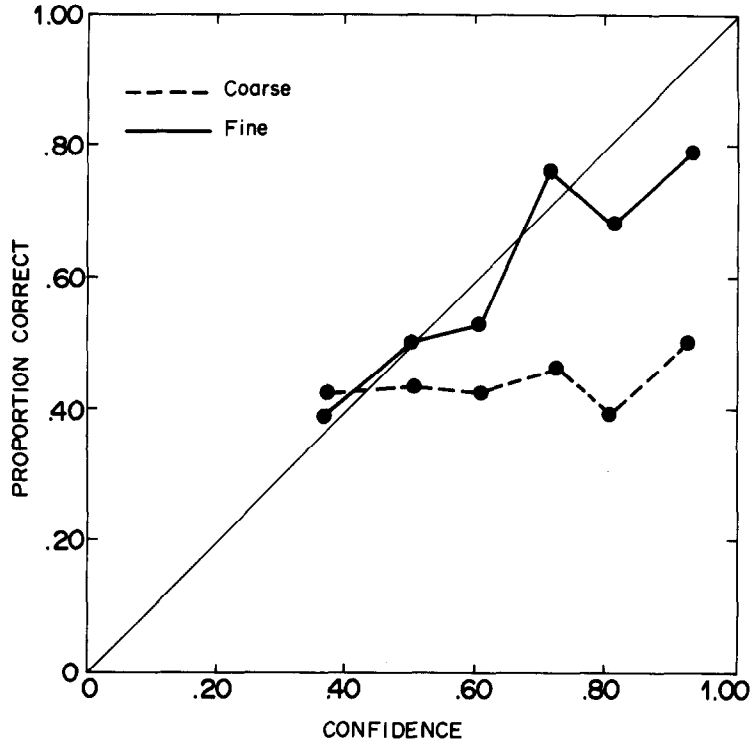

Fig. 1. Calibration of first choices for coarse partition and fine partition subjects. (Source: [24].)

is too slow; within an experimental group, a large increase in confidence is accompanied by a rather modest increase in knowledge. Where this "calibration curve" (as it is commonly labeled) falls depends primarily on the overall difficulty of the task. The "coarse" group in Figure 1 had a 44% success rate (on a task described immediately below), while the "fine" group had a 62% success rate. The latter group's curve lay above the former and showed much better performance, in the sense that the expected percentage of successes (as inferred from the probabilities) was much closer to the observed percentage. Indeed, the only significant discrepancy between success and expectations was at the (right-hand) extreme, where near-100% confidence met success only 80% of the time. In contrast, the coarse group's performance was almost unrelated to its performance, showing overconfidence over most of the range (in the sense that subjects expected to be successful more often than they were).

The particular results in Figure 1 came from individuals estimating the probability that they had identified the location of various items of information in the *Statistical Abstract of the United States*. They are strikingly similar to results observed with a variety of other tasks having similar difficulty levels. It appears as though people exercise the cognitive skills involved in confidence assessment similarly in information retrieval as elsewhere.

The difference between the two groups' tasks was the set of possibilities within which they were to locate 11 information items, such as "The percentage of physicians who are women" and "The average age of U.S. ships." The "fine" partition group received the 33 chapters appearing in the *Abstract*'s Table of Contents. The "coarse" partition group received 8 superordinate categories (of our own creation) without being told the chapters contained in each. The two partitions were used to explore a recurrent issue in the design of data-base systems with hierarchical informational structures, the tradeoff between breadth and depth in presenting entry-level menus. In some cases, the design is dictated by technical constraints, as when a computer screen or response selection device can accommodate only a limited number of alternatives. In others, the designer can weigh the additional information provided by a finer, broader partition against the compactness of a coarser partition [16].

The research literature shows a mixture of results in experiments with artificially constructed data bases of modest size presented on interactive computer systems, with apparent superiority accruing to finer partitions [17–20]. The results in Figure 1 extended that pattern of results to a richer, natural data base (i.e. the *Abstract*). Moreover, they also showed that the finer partition improved performance for both subjects' ability to identify the location of items and the realism of their expectations. These were called the *transparency* and *metatransparency* of the system, respectively.

The specific task faced by subjects in the study of Figure 1 was to identify the three most likely places to find each item, in order of likelihood, and then to distribute 100% of probability over the three options and the complementary "All Other Chapters [Categories]." Figure 2 depicts calibration for the second and third choices of the fine partition group in two different ways. The curves with closed circles in the lower left-hand corner show the actual probability judgments, which were necessarily lower for the (less likely) second choice than for the first, and lower for the third than for the second. Subjects' expectations were about as realistic here as with the first choices. The curves with the open circles are the result of examining how subjects allocated the probability remaining after expressing their confidence in preceding choices. For example, if a subject's probability distribution over the four alternatives was (.60, .30, .05, .05), then the implicit conditional probability for the second choice is .75 (= .30/(1.0 − .60)), and for the third it is .50. Looked at this way (called the *sequential choice* perspective), subjects' performance appears much poorer than with the other, *simultaneous choice* perspective. They were, it seems, roughly attuned to the level of success in those subsequent choices, but not to the details of how confident to be.

In point of fact, however, these subjects performed a simultaneous choice task. A separate study attempted to simulate a sequential search more closely by asking subjects to (a) pick a first choice, (b) assign a probability to it, (c) pick a second choice, imagining that the first was wrong, (d) assign a (conditional) probability to that choice, and (e) condition-

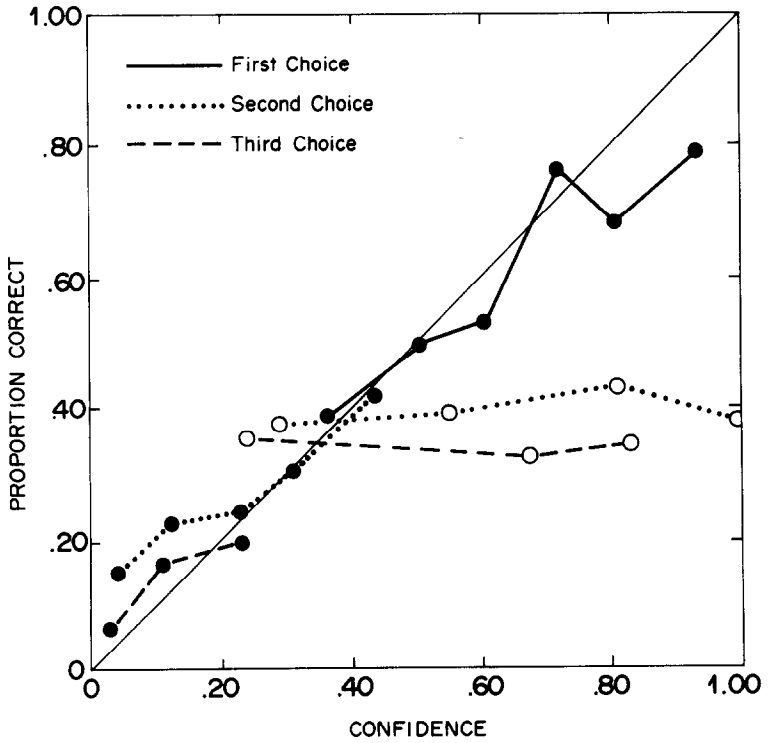

Fig. 2. Calibration of simultaneous probabilities (closed circles) and of sequential probabilities obtained by conditionalizing simultaneous probabilities for second and third choices (open circles). Fine partition subjects. (Source: [24].)

ally pick and evaluate a third choice. This manipulation had no effect on subjects' ability to pick correct locations. Surprisingly, however, these subjects were much more confident in their first choices, leading to considerable overconfidence, and to rather less confidence in their second and third choices. Apparently, focusing on the first choice without simultaneously considering alternatives made it seem particularly likely and the subsequently considered choices seem particularly unlikely. The recommendation implied by these results, encouraging searchers to consider several alternatives prior to beginning their search, fits well with existing psychological results [21, 22].[1]

The implications of these results for searching computerized data bases is limited by the pencil-and-paper format of the task. That limitation might seem especially severe for the supplementary study of sequential search in which subjects simply imagined having been wrong with previous choices prior to picking subsequent ones. Imagination is a plausible mechanism for producing a set of choices that will then be examined in batch processing. However, it would seem to be quite different from actually receiving negative feedback. The present study explores these issues by repeating the previous tasks in a computerized format. Specifically, it replicates directly the simultaneous search and sequential search tasks, and adds a new version of the sequential task in which subjects receive feedback on their choices, and proceed to subsequent choices only when their previous ones are wrong.

In one of the few direct comparisons of pencil-and-paper versus online presentation of questions, Newsted [23] found little difference in responses or reported satisfaction with the two formats, beyond what might be attributed to possible self-selection biases in subjects' choice of response mode. Newsted's task involved answering survey-type attitude questions (with the results indicating that a computer might offer a cost-effective way to collect data in a readily analyzed form). It did not involve cognitive skills that were peculiar to computers, beyond the need to manage simple procedures and whatever reticence might be evoked by the setting.

Were the Fischhoff and MacGregor [24] results (partially depicted in Figs. 1 and 2) to be replicated in the online format, then we would have evidence that these complex cognitive skills are not affected by the change of venue. Previous studies have found confidence assessment to be remarkably insensitive to experimental manipulations [25], with the same response pattern emerging despite diverse attempts to change it (e.g., by increasing the stakes involved, using odds rather than probabilities, providing lengthy lectures on the meaning of probabilities). Indeed, the basic pattern of Figure 1 (moderate sensitivity to the extent of one's knowledge, emerging as an overall tendency toward overconfidence) is so refractory to manipulations that it has been somewhat difficult to discern the psychological processes underlying confidence assessment (i.e., in the absence of contrasting patterns of results which emerge in different conditions).[2]

One manipulation that has made a difference [27] is providing intensive feedback about the adequacy of one's confidence assessments, in the form of calibration curves and performance statistics, accompanied by discussion of their meaning. A single round of such feedback produced a moderate improvement in calibration with moderate generalization to confidence assessment tasks using other response modes. Feedback on the correctness of location selection in information search could have a similar salutary effect by providing some indication of the appropriateness of selections and the associated confidence assessments. On the other hand, each bit of feedback is a fairly weak indicator. Since perfect performance is not expected with an uncertain system, it is hard to say, for example, that one is overconfident when a selection is wrong despite being 75% certain that it was right. Some (i.e., 25%) of one's 75% certain selections *should* be wrong (otherwise they should be 100% certain selections); why not this one? Over a set of selections, however, knowledge and confidence should be in line. It is this long-run perspective that is expressed

[1]Koriat, Lichtenstein, and Fischhoff [21] found that calibration could be improved by requiring subjects to list explicitly reasons why their favored choice might be right or wrong. Slovic and Fischhoff [22] reduced people's exaggerated belief that they would have been able to predict past events, had they been asked, by having them explicitly make the case for how events might have developed otherwise.

[2]For a somewhat exotic replication, see Henrion and Fischhoff [26].

by the calibration curve, which is deemed perfect if one is right *XX*% of the time when one is *XX*% certain of being right. Case-by-case feedback might achieve the same effect as pooled feedback [as in 27] or even enhance it by delivering its message repeatedly. Or, it could have no cumulative impact, with subjects failing to learn anything from the weak message that it provides.

It is possible to raise similarly conflicting hypotheses about the impact of the move to computers alone on performance. It might improve calibration by imposing a more formal and unfamiliar setting, thereby increasing self-reflection. Or, it might degrade calibration by instilling unjustified expectations regarding system performance or by introducing new sources of error (or success) about which users have little insight or sensitivity (e.g. [28]).

Whatever users' calibration turns out to be, having a quantitative assessment of it should provide some guidance in predicting and improving system design. Users' expectations from a system should affect their readiness to use it at all and their willingness to persist in the face of adversity. The appropriateness of those expectations should affect their frustration and satisfaction with the system. Ideally, a system should both generate and justify high expectations. Although most design attention seems devoted to increasing success rates, helping users to better understand a system's capabilities might make a significant contribution to how effectively they use it. Such help might come through improved technical design, online feedback, or training and instruction.

METHOD

Design

For each of 11 general knowledge items, subjects first selected the most promising locations in the *Statistical Abstract of the United States* and then assessed the probability (from .00 to 1.00) that each selection was correct. For roughly half of the subjects, the set of possible locations consisted of the 33 *chapters* appearing in the *Abstract*'s Table of Contents; for the remainder, the set consisted of 8 superordinate *categories* that we created. These correspond, respectively, to the fine and coarse partitions of Figures 1 and 2. Full listings of items, chapters, and categories appear in Fischhoff and MacGregor [24].

Category and chapter subjects were divided roughly equally across three experimental conditions: *Simultaneous search*, wherein subjects chose three possible locations in order of decreasing likelihood and then divided 1.00 of probability across them and "All Other Categories [Chapters]." *Sequential search*, wherein subjects chose three locations in turn, each time assessing the (conditional) probability that it contained the sought item, under the assumption that they had yet to find the correct location. These subjects received no feedback, but only imagined that they had failed previously when considering their second and third choices. *Feedback*, wherein subjects made their choices sequentially, but proceeded only if told that their previous selection was incorrect.

All responses were collected online in a computer-interactive format. There were thus six cells in this 3×2 factorial design, which crossed three experimental conditions with two sets of possible locations. In addition, the present simultaneous and sequential search conditions are directly comparable to those conditions in Fischhoff and MacGregor [24], where a paper-and-pencil format was used.

Procedure

Subjects participated in this experiment as the first in a series of unrelated experiments involving either computer-interactive or pencil-and-paper tasks, all having to do with judgment and decision making. The task was entirely self-administered and self-paced. A series of text screens introduced the task and gave instructions for responding. Subjects read the screens at their own pace, progressing to subsequent screens by pressing a designated key. An experimenter was on hand to answer technical questions about the procedure (but not

about the meaning of the items or contents of the *Abstract*). The task was sufficiently straightforward (and the computer implementation sufficiently accomplished) that there were relatively few questions of any sort. The computers were all IBM Standard PCs. Four were arranged around a large conference table so that several subjects could perform the task at once, but without being able to observe one another's work.

Subjects

A total of 261 individuals were recruited by an advertisement for paid subjects who were native speakers of English, appearing in the University of Oregon student paper. They were divided fairly evenly between men and women. In previous experience with subjects recruited in this manner where we have collected demographic data, the mean age of the males has been 24 and the females 21. Approximately two-thirds are students, with most of the remainder somehow involved in the university community. Although by no means representative of all U.S. adults, these individuals are not unlike the sorts of individuals who the developers of data bases in general and of the *Abstract* in particular would hope to be able to serve.

These subjects were assigned randomly to the six experimental conditions. In the analyses that follow, we will add selected results from the 220 subjects who completed Fischhoff and MacGregor's pencil-and-paper tasks. They were recruited in a similar manner.

Data analysis

Within each group, subjects' probability assessments were grouped into 12 ranges: 0, .01–.09, .10–.19, .20–.29, . . . , .90–.99, 1.00. Where this grouping left less than 20 responses in a range, adjacent categories were merged so as to produce more stable estimates of the percentage of correct responses associated with each probability (which was represented by the mean of all the probability responses in a range), and of the summary statistics described below.

A common way to characterize the performance associated with probability assessments is the *Brier Score* [29], which is used routinely by the U.S. National Weather Service to evaluate probability of precipitation forecasts [30].[3] It distinguishes three kinds of performance: *Knowledge*, how much one knows; *Resolution*, one's ability to discriminate different degrees of knowledge; *Calibration*, one's ability to assign appropriate levels of absolute confidence to different degrees of knowledge. Fuller expositions of these scores can be found in Fischhoff and MacGregor [24] and sources cited therein. Some alternative summary statistics are offered in [32] and [33].[4]

According to the computational scheme of the Brier score, the more one knows, the lower one's Knowledge score. Because this score is a direct function of the percentage of correct responses, it will not be used here, in deference to *percentage correct*, which is more readily interpreted. The better feeling that one has for when one knows more and when one knows less, the greater is one's Resolution score. It is, in effect, the variance of the percentages of correct responses associated with different probabilities, weighted by the number of responses involved. The more appropriate one's expressions of confidence, the lower one's Calibration score. It is, in effect, the mean squared difference between the calibration curve and identity line, weighted by the number of responses involved. The Brier score equals the sum of the Knowledge and Calibration scores minus the Resolution score, so that lower values indicate better overall performance. In practice, it is most heavily influenced by the Knowledge score, and will not be used here.

[3]These forecasters' performance is excellent [30,31], perhaps because they have ideal conditions for learning: a clear criterion, prompt feedback, and an incentive system that rewards them for candor.

[4]The formal expression of the Brier score is $= c(1-c) + (1/N)\sum_{t=1}^{T} n_t(r_t - c_t)^2 - (1/N)\sum_{t=1}^{T} n_t(c_t - c)^2$.

In this expression, c represents the overall proportion of correct answers. N is the total number of answers in the test set. T is the number of different probability values used by the respondent. Each n_t represents the number of times a particular response was used. c_t is the proportion of correct answers among all answers assigned probability r_t.

Table 1. Transparency (percentage of correct selections)

Choices	Categories			Chapters		
	1	2	3	1	2	3
Pencil and Paper[a]						
Simultaneous						
Conditional	43.7	37.9	34.6	61.6	39.5	33.8
Cumulative	43.7	65.1	77.2	61.6	76.1	82.4
Sequential						
Conditional	41.3	32.8	33.9	59.7	32.3	29.1
Cumulative	41.3	60.6	73.8	59.7	72.7	80.6
Computer Interactive						
Simultaneous						
Conditional	42.3	30.9	33.0	62.6	41.7	41.0
Cumulative	42.3	57.9	63.3	62.6	76.2	80.6
Sequential						
Conditional	41.6	37.8	37.5	58.3	40.1	34.7
Cumulative	41.6	63.6	77.3	58.3	75.0	83.7
Sequential With Feedback						
Conditional	43.2	35.7	37.8	58.5	43.8	39.5
Cumulative	43.2	63.4	77.3	58.5	76.6	85.8

[a]Results from [24].

RESULTS

Transparency

Table 1 shows the percentage of correct responses associated with first, second, and third choices for Fischhoff and MacGregor's [24] four pencil-and-paper groups (receiving categories or chapters in simultaneous or sequential mode) and the present six groups. The "conditional" rows refer to the percentages of subjects selecting the correct choice among those who had yet to choose correctly.

"Cumulative" refers to the percentage of subjects who had chosen the correct location by the end of that round. From both perspectives, the system was similarly transparent to subjects in all groups receiving the same set of locations. The ranges of conditional percentages correct were 4.3% and 2.2% for the first choices of the category and chapter conditions, respectively. They were 4.8% and 11.9% (i.e., somewhat more variable) for the third choices, where fewer responses were involved (given the high percentage of cases where subjects had already chosen correctly). The categories are, therefore, much more difficult to use, although by the third choice the cumulative percentages correct are relatively close (presumably due, in part, to the limited number of categories).

It is of mild substantive interest that putting questions on a computer did not improve this aspect of performance. Although the more involved computer setting might evoke greater attention, it does not do so in a way that helps subjects locate these items. Even the feedback condition, which might be expected to tell subjects something about how they were interpreting and misinterpreting the system, had no effect. One possible reason is that the items and data base were too heterogeneous to reveal much internal structure with only 3 choices for each of 11 items.[5]

Methodologically, however, the similarity is quite important, given the dependence of calibration on task difficulty [12]. The similar difficulty of the tasks allows direct comparisons across conditions within the category and chapter groups.

[5]Nor was there any tendency for feedback subjects to get higher percentages correct (relative to no-feedback sequential-search subjects in the computer-interactive format) on the latter items in the set, at which point they might have been able to glean some cumulative lesson from the feedback. If anything, the opposite was true, with a rank correlation of −0.25 between degree of "improvement" with the feedback and ordinal position.

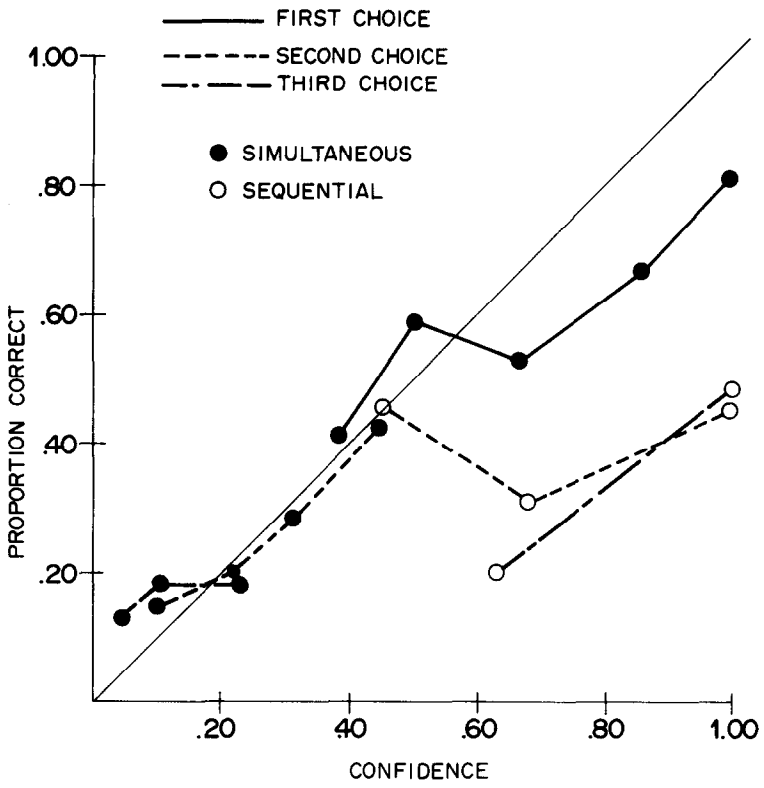

Fig. 3. Calibration of simultaneous probabilities (closed circles) and of sequential probabilities obtained by sequentializing simultaneous probabilities for second and third choices (open circles). Subjects using chapters in computer-interactive mode.

Metatransparency

Simultaneous search. Figure 3 presents the calibration curves for the simultaneous-search computer-interactive subjects that are directly comparable to the pencil-and-paper results in Figure 2. Visual comparison shows considerable similarities. The curves reflecting actual responses (closed circles) for all three choices slope upward and lie relatively close to the identity line; overall, there is a tendency toward overconfidence in first choices. Conditionalizing the probabilities assigned to second and third choices (open circles) shows, as before, a less attractive picture, with little relationship between Confidence and Knowledge.[6]

Calibration curves for responses to the category tasks were equally similar here (not shown) and in the pencil-and-paper tasks (the first choice curve of which appears in Fig. 1). The calibration curves for both the actual and conditionalized responses were quite flat, showing little sensitivity to the extent of subjects' knowledge. Overall, there was again a substantial tendency toward overconfidence, except for the actual probabilities assigned to third choices (where there was very little probability "left" to be overconfident with, given the overconfidence in the preceding choices).

Thus, none of the factors that might have made a difference in the shift from pencil-and-paper to computer-interactive mode did so regarding calibration, any more than it did with regard to transparency. This consistency is reflected in the statistical summaries of

[6]As mentioned in the introduction, the conditionalizing procedure looks at the percentage of the "remaining" probability that is allocated to a particular choice. Formally, the conditionalized probability for choice i is $P_i \Big/ \left(1 - \sum_j P_j\right)$, where $j = 1, \ldots, i-1$ and P_j equals the actual probability assigned to choice j in a simultaneous search mode.

Table 2. Summary on performance statistics for simultaneous search: paper and pencil vs. computer interactive

	Categories			Chapters		
	1	2	3	1	2	3
Actual Responses						
Paper and Pencil						
Proportion Correct	.437	.237	.155	.616	.249	.112
Mean Confidence	.599	.225	.112	.676	.201	.093
Over/under Confidence	.162	−.012	−.043	.060	−.048	−.019
Calibration	.056	.007	.005	.009	.005	.002
Resolution	.003	.002	.000	.024	.007	.003
Number of Responses	670	666	652	427	406	365
Computer Interactive						
Proportion Correct	.423	.204	.149	.626	.247	.177
Mean Confidence	.690	.270	.145	.735	.259	.153
Over/under Confidence	.267	.065	−.004	.109	.011	−.024
Calibration	.097	.012	.007	.024	.001	.004
Resolution	.003	.001	.001	.017	.011	.001
Number of Responses	532	460	221	366	275	141
Simultaneous Converted to Sequential						
Pencil and paper						
Proportion Correct	.437	.379	.346	.616	.395	.338
Mean Confidence	.599	.569	.607	.676	.617	.728
Over/under Confidence	.162	.190	.261	.060	.222	.391
Calibration	.056	.065	.110	.009	.094	.227
Resolution	.003	.003	.001	.024	.001	.000
Number of Responses	670	377	227	427	157	80
Computer Interactive						
Proportion Correct	.423	.309	.330	.626	.417	.410
Mean Confidence	.690	.796	.821	.735	.773	.905
Over/under Confidence	.267	.487	.491	.109	.357	.495
Calibration	.097	.298	.270	.024	.177	.277
Resolution	.003	.011	.003	.017	.004	.026
Number of Responses	532	269	88	366	120	39

Table 2 as well, the top half of which presents actual responses and the bottom half conditionalized responses (with first-choice statistics being the same for each).

Looking at the actual responses, one sees a steady decline (across choices) in proportion correct which is paralleled by an even steeper decline in confidence. The result is a shift from substantial overconfidence on first choices to mild underconfidence with third choices, as expressed in a change in the sign of the over/underconfidence statistic (which is equal to the difference between mean confidence and proportion correct). With Calibration scores, the smaller the better. Comparing the values for the pencil-and-paper tasks with the curves in Figures 1 and 2 provides some indication of the meaning of these numerical values. Thus, the score of .056 for the coarse cure in Figure 1 represents quite poor performance. A score of .00 represents perfect performance. It is approached most closely by the second and third choice curves with the chapters, as can be seen in Figures 2 and 3. Calibration scores also improve with the second and third choices for the categories. However, this, like the reduction in the absolute value of the over/underconfidence score, largely reflects the restricted range of probability responses.

The sequentialization of second and third choice probabilities shows a more dismal picture. The flatness of these curves in Figures 2 and 3 is one sign of this insensitivity. Its statistical reflection in the bottom half of Table 2 begins with similar proportions correct (about .40) with both locations sets and search modes. It continues with conditionalized probabilities of being correct (ranging from .569 to .905) that would lead one to expect a much higher success rate. The conjunction leads to substantial overconfidence (ranging

from .190 to .495) and very poor Calibration (from .065 to .298). From this perspective, subjects approached their subsequent choices the way that they approached their initial choice: they were reasonably certain of having gotten it right this time, an expectation that was not matched by their actual success.

In all four cases (second and third choices for categories and chapters), performance here was substantially worse with the computer. It is reflected in greater overconfidence and higher Calibration scores. A significant contributor is the much smaller number of responses involved in the computer-interactive results.[7] The numbers decline with each successive choice, in large part because subjects who answered correctly on previous choices no longer appear. A secondary reason is subjects who decline to give second and third choices, indicating by their probabilities that they are certain that the item could not be anywhere else. Given the similar cumulative percentages of correct choices (for the two response formats within each location set), the discrepancy is almost entirely due to the higher proportion of computer-interactive subjects behaving as though they had certainly gotten it right already. In the absence of a corresponding improvement in actually having gotten it right, the decreased number of choices erodes calibration in two ways. One is by increasing the mean confidence in those final choices (to 1.0) and, hence, overconfidence. The second is by increasing the variance in the proportions correct associated with different probability responses, simply due to reduced sample size. Such variability tends to increase Calibration scores.

Resolution, the final statistic in the table, measures the variance in proportions correct. The near-zero scores with the categories reflect the flatness of those curves. The larger (hence, better) scores with the chapters reflect their upward slope. Resolution is relatively small with third choices in simultaneous search despite the good calibration because proportions correct vary over such a small range (even though those changes match the corresponding changes in confidence). There were no systematic differences between the computer-interactive and pencil-and-paper versions in resolution.

Given the general similarity in performance patterns with the two response modes, we lean toward interpreting the smaller number of second and third choices for computer-interactive subjects as being more the cause than the effect of their great overconfidence and inferior calibration. Specifically, we believe that subjects found it slightly more inconvenient to register their responses with the computer than with pencil and paper. As a result, they were more likely to save time by choosing just one or two locations. Dividing their 1.00 of probability over fewer locations led to higher mean probability, greater overconfidence, and poorer calibration. If this account is correct, then it is somewhat surprising that designing a seemingly simple interface should have such unintended consequences and somewhat surprising that subjects' feelings of confidence for their second and third choices should be so easily disrupted. This interpretation seems in keeping with the picture of indifferent performance revealed in the conditionalized responses. Although subjects' task is to evaluate their chosen options simultaneously, they may do so, in part, by treating each subsequent choice like a new first choice.

Sequential search. The top third of Table 3 presents performance statistics for the pencil-and-paper subjects asked to simulate an explicitly sequential search. As mentioned before (and seen here), the heightened attention on the first choice increased the confidence placed in it. Without a corresponding increase in knowledge, the result was greater overconfidence and worse calibration, in comparison with the first choice in simultaneous search. Apparently as a result, subjects had less confidence in their second choice and even less in their third. The differences between these conditional probabilities and the conditionalized probabilities in Table 2 are quite striking, with much better performance being observed here. The tendency to give only one or two responses was not observed here

[7]Making this comparison requires normalizing the sample sizes in Table 2 to accommodate the different numbers of subjects in each group. For the categories, subjects made second choices 56% of the time and third choices 37% of the time for the pencil-and-paper task, while the corresponding percentages for computer-interacting subjects were 51% and 17%. For the chapters, pencil-and-paper subjects made second and third choices, 37% and 19% of the time, respectively; with computer-interactive subjects, the percentages were 33% and 9%.

Table 3. Summary of performance statistics: sequential search

	Categories			Chapters		
	1	2	3	1	2	3
Paper and Pencil						
Proportion Correct	.413	.328	.339	.597	.323	.291
Mean Confidence	.774	.519	.334	.797	.502	.317
Over/under Confidence	.361	.191	−.005	.200	.179	.027
Calibration	.154	.084	.021	.050	.083	.033
Resolution	.004	.003	.025	.023	.008	.006
Number of Responses	649	381	254	546	220	148
Computer Interactive						
No-feedback						
Proportion Correct	.416	.378	.375	.583	.401	.347
Mean Confidence	.780	.575	.348	.756	.526	.410
Over/under Confidence	.364	.197	−.027	.173	.125	.063
Calibration	.152	.055	.029	.035	.062	.048
Resolution	.006	.007	.004	.024	.006	.033
Number of Responses	462	270	168	472	197	118
Feedback						
Proportion Correct	.432	.357	.378	.585	.438	.395
Mean Confidence	.671	.563	.444	.743	.566	.457
Over/under Confidence	.239	.206	.065	.158	.128	.062
Calibration	.100	.061	.047	.044	.050	.027
Resolution	.004	.004	.018	.009	.002	.010
Number of Responses	528	300	193	492	203	114

because subjects were explicitly asked for three locations. Again, the superiority of chapters over categories is maintained.

The summary statistics for the computer-interactive version of the sequential task (the middle third of Table 3) look quite similar in all respects. Here, too, whatever factors might have been thought to distinguish the computerized format had no overall effect on subjects' exercise of these cognitive skills. Some of the similarities can be seen in Figure 4, which provides the (overconfident) calibration curves for the first choices with the two formats. The change in procedure changed these responses relative to the simultaneous search and those changes were faithfully recorded with both formats.

Feedback. The limited impact of the computer-interactive format on the responses reported thus far might be attributed to its involving no change in the tasks, beyond putting them on a computer. Providing feedback creates a new condition, and one well suited to exploit the potential of computers. The bottom third of Table 3 presents performance statistics for subjects receiving feedback. They are strikingly like the comparable statistics for the no-feedback subjects immediately above. The one possibly notable difference is the lower confidence (and, hence, reduced overconfidence and Calibration scores) for the first choices of feedback subjects. However, the calibration curve is essentially flat (as reflected in the small Resolution score) and in the absence of similar changes elsewhere, this seems like random fluctuation.

DISCUSSION

On *a priori* ground, there seemed to be a variety of reasons why transferring the set of information search tasks to the computer might have affected subjects' performance on these tasks, relative to their pencil-and-paper predecessors [24]. The shift could have affected either the confidence with which subjects approached the task as a whole, or the care with which they examined their knowledge regarding the location of particular items. However, except for the reduced number of choices made on the simultaneous search, computerization had no appreciable impact. Conceivably, it may have had a variety of effects

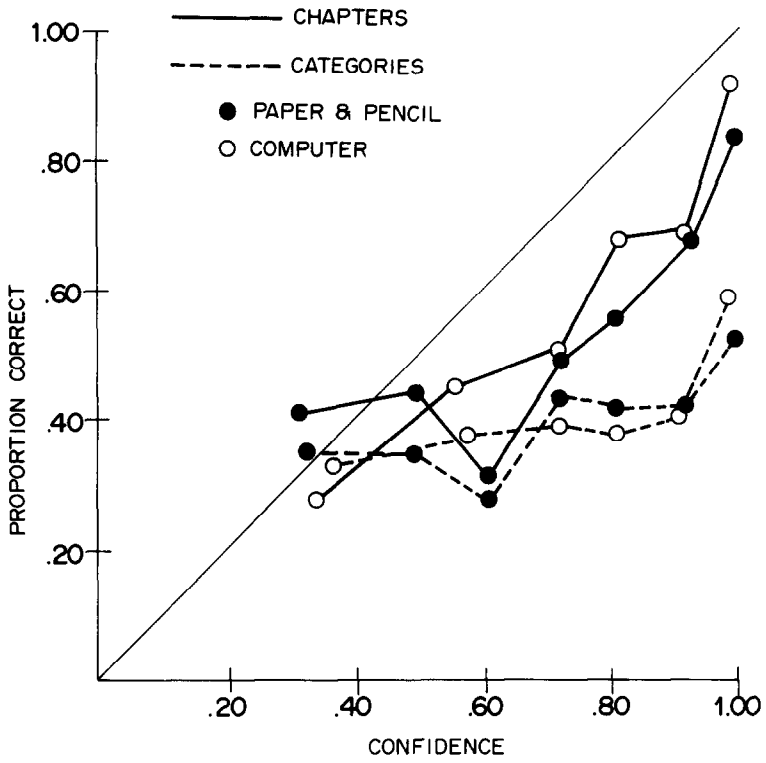

Fig. 4. First choices for sequential search subjects using computer-interactive (open circles) or pencil-and-paper (closed circles) format.

that cancelled one another in the aggregate. However, a more parsimonious account is simply that subjects approach this task with the same cognitive skills that they approach its paper-and-pencil version.

To summarize: people are only somewhat sensitive to the extent of their own knowledge, with the most common overall tendency being overconfidence. That overall tendency depends upon the gap between how well people expect to do on a task as a whole (one expression of which is their mean confidence) and how well they actually do (one expression of which is their percentage of correct responses). When a task is more difficult than subjects expected, their relative insensitivity to how much they know leaves them with inappropriately high confidence. Although not observed here, complementary processes can produce underconfidence with unexpectedly easy tasks. Within a task, the belief that they can distinguish widely different states of knowledge leads people to assign disparate probabilities to sets of items with quite similar percentages correct. Resolution scores, reflecting the ability to make such discriminations, set a limit to the range of probabilities that can be usefully offered. At the extreme, as with the flat "coarse" curve in Figure 1 (and its vanishing Resolution score), subjects would be better off consistently giving their mean confidence level than trying to distinguish levels of confidence. As represented by the categories, this data base has no metatransparency and not that much transparency, once one considers how well subjects would have done just by guessing.

The consequences of miscalibration should depend on the decisions based on confidence assessments. von Winterfeldt and Edwards [34] have shown that rational decision makers (i.e., ones following the expected utility rule) facing continuous decision options (e.g. invest $\$X$) will not pay a very large price for even moderate inaccuracy. The price paid for this protection is reduced ability to detect errors of estimation and, hence, to learn to make better estimates. It is an empirical question how miscalibration of this magnitude will affect people's satisfaction (or frustration) with a data base, their approach to particular search tasks, and the scrutiny they afford to search products.

If having inappropriate expectations does exact a price, then several responses are possible. The simplest, if it works, is to tell users what to expect from a data base, preferably with some indication of how their performance is likely to vary with experience and salient features of particular search tasks (e.g., success rates with author, title, and subject searches).[8] More ambitious, but perhaps more viable than presenting summary descriptions, is designing the interface so that it conveys an appropriate overall impression and encourages users to approach their task in an effective way.

The previous study [24] suggested one way of doing so, namely, having users nominate and evaluate several possible locations prior to beginning their search. The sequentialized perspective on the probabilities assigned to the second and third choices suggests that somewhat more attention might be directed at them. The increased tendency to be satisfied with just one or two choices in the computer version of this task indicates the importance of designing an interface that makes it seem easy and important to make those latter choices. The heightened overconfidence in first choices with the sequentialized search suggests discouraging a tendency to focus unduly on any favorite candidate (as in [21]). The similarity of the feedback and nonfeedback conditions suggests that feedback needs to be presented more effectively to be of any value (perhaps in the form of personalized performance statistics, as used by [27]). The robustly inferior performance with the categories reinforces the widely recognized importance of developing comprehensible entry menus [36–39].[9]

The similarity of these results to those observed in the most directly comparable studies of confidence assessment suggests that results observed elsewhere regarding these cognitive skills might be tentatively generalized to this context. It is an open, and important, question whether similar results would be found with more involved and involving computer-interactive systems. For example, are the operators of semi-automated process-control industries [40] similarly miscalibrated in their estimates of how well they understand how those systems are performing? One component of that study will be discovering the cues that determine users' overall confidence in a system as well as the cues that govern their decisions to trust or override a system that they are monitoring.

REFERENCES

1. Belkin, N.J. Cognitive models and information transfer. Social Science and Information Studies, 4: 111–129; 1984.
2. Hollnagel, E.; Mancini, G.; Woods, D., editors. Intelligent decision aids in process environments. Heidelberg: Springer-Verlag; 1986.
3. Vigil, P.J. The psychology of online searching. Journal of American Society of Information Science, 34: 281–287; 1983.
4. Bates, M.J. The fallacy of the perfect thirty-item online search. RQ, 24(1): 43–50; 1984.
5. Tenopir, C. To err is human: seven common searching mistakes. Library Journal, 635–636; 1984.
6. Wason, P.C.; Johnson-Laird, P.N. Psychology of reasoning: structure and content. London: Batsford; 1972.
7. Streatfield, D. Moving towards the information user: some research and its implications. Social Science Information Studies, 3: 223–240; 1983.
8. Collins, W.S. Indexing documents by Gedanken experiments. Journal of American Society of Information Science, 29: 107–119; 1978.
9. Murphy, G.L.; Medin, D.L. The role of theories in conceptual coherence. Psychological Review, 92: 289–316; 1985.
10. Bookstein, A. Probability and fuzzy-set applications to information retrieval. Annual Review of Information Science and Technology, 20: 117–151; 1985.
11. Fischhoff, B.; Slovic, P.; Lichtenstein, S. Knowing with certainty: the appropriateness of extreme confidence. Journal of Experimental Psychology: Human Perception and Performance, 3: 552–564; 1977.
12. Lichtenstein, S.; Fischhoff, B.; Phillips, L.D. Calibration of probabilities: state of the art to 1980. In: Kahneman, D.; Slovic, P.; Tversky, A., editors. Judgment under uncertainty: heuristics and biases. New York: Cambridge University Press; 1982: 306–334.

[8]Anecdotally, we found that users of a computerized library catalog were more satisfied with its performance (without any improvement in that performance) when told that it was developed primarily for inventory purposes and only secondarily for lay users. Satisfaction also seemed to be increased by being told that they had at their disposal all the information that was available to librarians [35].

[9]Fischhoff, MacGregor, and Blackshaw [40] approach these design issues using the present performance statistics as evaluative criteria.

13. Nelson, T.O. A comparison of current measures of the accuracy of feeling-of-knowing predictions. Psychological Bulletin, 95(1): 109–133; 1984.
14. Wallsten, T.S.; Budescu, D.V. Encoding subjective probabilities: a psychological and psychometric review. Management Science, 29(2): 151–173; 1983.
15. Wellman, H.M. The origins of metacognition. Metacognition, Cognition and Human Performance, 1: 155–205; 1985.
16. Savage, R.E.; Habinek, J.K. A multilevel user interface: decision and evaluation through simulation. In: Thomas, J.C.; Schneider, M.L., editors. Human factors in computer systems. Norwood, NJ: Ablex; 1984.
17. Kiger, J.I. The depth/breadth tradeoff in the design of menu-driven user interfaces. International Journal of Man-Machine Studies, 20: 210–213; 1984.
18. Snowberry, K.; Parkinson, S.; Sisson, N. Computer display menus. Ergonomics, 26: 699–712; 1983.
19. Snowberry, K.; Parkinson, S.; Sisson, N. The effects of help fields on navigating through hierarchical menu structures. International Journal of Man-Machine Studies, 22: 479–491; 1985.
20. Landauer, T.K.; Nachbar, D.W. Test of a model of menu-traversal time. Murray Hill, NJ: Bell Communications Memorandum; 1986.
21. Koriat, A.; Lichtenstein, S.; Fischhoff, B. Reasons for confidence. Journal of Experimental Psychology: Human Learning and Memory, 6: 107–118; 1980.
22. Slovic, P.; Fischhoff, B. On the psychology of experimental surprises. Journal of Experimental Psychology: Human Perception and Performance, 3: 544–551; 1977.
23. Newsted, P.R. Paper versus on-line presentations of subjective questionnaires. International Journal of Man-Machine Studies, 23: 231–247; 1985.
24. Fischhoff, B.; MacGregor, G. Calibrating databases. Journal of American Society of Information Sciences, 37; 1986.
25. Fischhoff, B. Debiasing. In: Kahneman, D.; Slovic, P.; Tversky, A., editors. Judgment under uncertainty: heuristics and biases. New York: Cambridge University Press; 1983: 422–444.
26. Henrion, M.; Fischhoff, B. Uncertainty assessment in the estimation of physical constants. American Journal of Physics, 54:791–798; 1986.
27. Lichtenstein, S.; Fischhoff, B. Training for calibration. Organizational Behavior and Human Performance, 26: 149–171; 1980.
28. Kiesler, S.; Siegel, J.; McGuire, T.W. Social psychological aspects of computer-mediated communication. American Psychology, 39: 1123–1134; 1984.
29. Brier, G.W. Verification of forecasts expressed in terms of probability. Monthly Weather Review, 75: 1–3; 1950.
30. Murphy, A.H.; Winkler, R.L. Can weather forecasters formulate reliable probability forecasts of precipitation and temperature? National Weather Digest, 2: 2–9; 1977.
31. Murphy, A.H.; Brown, B.G. A comparative evaluation of objective and subjective weather forecasts. Journal of Forecasting, 3(4): 361–394; 1984.
32. Swets, J.A.; Pickett, R.M. Evaluation of diagnostic systems: methods from signal detection theory. New York: Academic Press; 1982.
33. Yates, J.F. External correspondence: decompositions of the mean probability score. Organizational Behavior and Human Performance, 30: 132–156; 1982.
34. von Winterfeldt, D.; Edwards, W. Decision analysis and behavioral research. New York: Cambridge University Press; 1986.
35. Blackshaw, L. Decision making in searching computerized databases (DR report no. 86-8). Eugene, OR: Decision Research; 1986.
36. 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.
37. Furnas, G.W.; Landauer, T.K.; Gomez, L.M.; Dumais, S.T. Statistical semantics: analysis of the potential performance of key-word information systems. Bell System Technical Journal, 62(6): 1753–1826; 1983.
38. Jagodzinski, A.P. A theoretical basis for the representation of online computer systems to naive users. International Journal of Man-Machine Studies, 18: 215–252; 1983.
39. Fidel, R. Factors affecting online bibliographic retrieval: a conceptual framework for research. Journal of American Society of Information Science, 34(3): 163–180; 1983.
40. Fischhoff, B.; MacGregor, D.; Blackshaw, L. Creating catagories for databases. International Journal of Man-Machine Systems, in press.
41. National Research Council. Research and modeling of supervisory control systems. Washington, DC: Author; 1984.