Int. J. Man–Machine Studies (1987) **27**, 33–63

Creating categories for databases

BARUCH FISCHHOFF, DONALD MACGREGOR AND LYN BLACKSHAW
Decision Research, 1201 Oak, Eugene, Oregon 97401, U.S.A.

(Received 15 July 1986 and in revised form 25 March 1987)

The value of a database is bounded by the accessibility of the information it contains. The present studies provide a multifaceted approach to designing and evaluating entry-level menus using, as a case in point, the *Statistical Abstract of the United States*. They consider different ways of organizing material into categories, developing labels for those categories, and presenting them to users. As performance criteria, the studies consider both the *transparency* of the resulting system, how easily users can identify the location of items, and its *metatransparency*, how well users can assess the system's transparency. The latter criterion, which measures the realism of users' expectations regarding their success with the system, is relevant to how willing users are to attempt a search, how carefully they scrutinize its products, and how satisfied (or frustrated) they are with their progress. Aside from demonstrating a general method, these studies provide some potentially useful substantive results. One is the persistent superiority of the *Statistical Abstract's* 33 chapters as an entry-level menu, as compared with various attempts to create superordinate categories. A second is subjects' relatively poor ability to predict success in locating individual items. A third is the relatively good performance obtained with superordinate categories whose internal structure and labels were determined by individuals like the eventual users. These results replicate and amplify results using more restricted and artificial databases, and offer some promise for designing interfaces as well as some insight into subjective categorization processes.

Introduction

To borrow a pair of terms from cognitive psychology, the ideal database will earn high marks for both *availability* and *accessibility* (Tulving & Pearlstone, 1966). The former criterion refers to the amount of information that it contains. The latter criterion refers to the ease with which users extract information from it. Typically, the two criteria are in conflict. As the size of a database increases, the effort of extraction will grow as well. Greater size is likely to mean more elaborate search procedures, more ways to go wrong, more possible places to look, and more levels of internal organization (e.g. nested menus) to traverse before getting to actual information (correct or incorrect). As long as databases compete on the basis of comprehensiveness, there will be increasing threats to accessibility (e.g. Roth, 1985).

The primary mode of access to any database is some set of categories. These may partition some universe of content, as do the elements of an on-screen menu or table of contents. Or, they may be interrelated, as are the elements of a free-search

Please address correspondence to Baruch Fischhoff at the Departments of Engineering and Public Policy or Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA 15213.

0020-7373/87/070033 + 31$03.00/0

lexicon. The more satisfactorily these categories organize and communicate the contents of the database, the more accessible those contents will be. As a result, creating such categories is a major element in the design of information systems (Cooper, 1978; Fidel, 1983; Furnas, Landauer, Gomez & Dumais, 1983; Kiger, 1984; Lee, Whalen, McEwan & Latremouille, 1984; McDonald, Stone, Liebelt & Karat, 1982; Pejtersen, 1980; Witten, Cleary & Greenberg, 1984).

Two paradigms may be discerned for the creation of categories, one focusing on the substance of the search, the other on its process. Although they could be complementary, most design efforts that are not entirely ad hoc seem to focus on one and rely on good sense to take care of the other. A pure "substance" approach attempts either: (a) to divine the underlying logic of the domain and express it in terms of an efficient taxonomy; or (b) describe the (modal) user's mental model of the domain and devise categories that conform to it. The former leaves it to the user to discern that logically superior system. The latter leaves it to the system's designer to align categories with users' potentially idiosyncratic perspectives. A pure "process" approach attempts to facilitate movement within the system, trying to ensure that users always know where they are, even if they are not exactly where they want to be. Being oriented within the system allows users to make good decisions (or gambles) about which category members to select, to realize when they are wrong, and to retrace and correct their steps. It accepts fallibility and uncertainty as inevitable and attempts to deal with them effectively (Bookstein, 1985). By contrast, the substantive approach holds out the hope of always getting the desired information on the first try.

The present article offers a methodology for a substantively informed process approach to creating categories. It develops a procedure for evaluating category systems, a procedure for creating consensually valid categories, and a perspective on presenting those categories. These methods are developed in the context of psychological research into decision-making, concept comprehension, and information processing. They are illustrated in the context of one representative database, the *Statistical Abstract of the United States* (U.S. Department of Commerce, 1983).

WHAT KINDS OF PERFORMANCE ARE DESIRED?

For the user of a database, each set of categories represents a set of alternative courses of action, one (or more) of which must be selected and its contents investigated. Such a choice among alternatives whose outcome is not entirely predictable constitutes a decision under conditions of uncertainty. Attractive decisions are those in which a desirable outcome is likely. This occurs either when all alternatives lead to good outcomes or when those that do are readily discerned. Thus, databases present attractive decisions when it is easy to discern which category contains the desired information. We have called this property *transparency* (Fischhoff & MacGregor, 1986). It can be measured, simply, in terms of the percentage of chosen alternatives that prove to contain the desired information.

When transparency is incomplete, each choice represents a gamble. Under those circumstances, optimal use of a database means choosing those categories that have the best *expected outcome*. Much of decision theory (Raiffa, 1968; Watson & Buede,

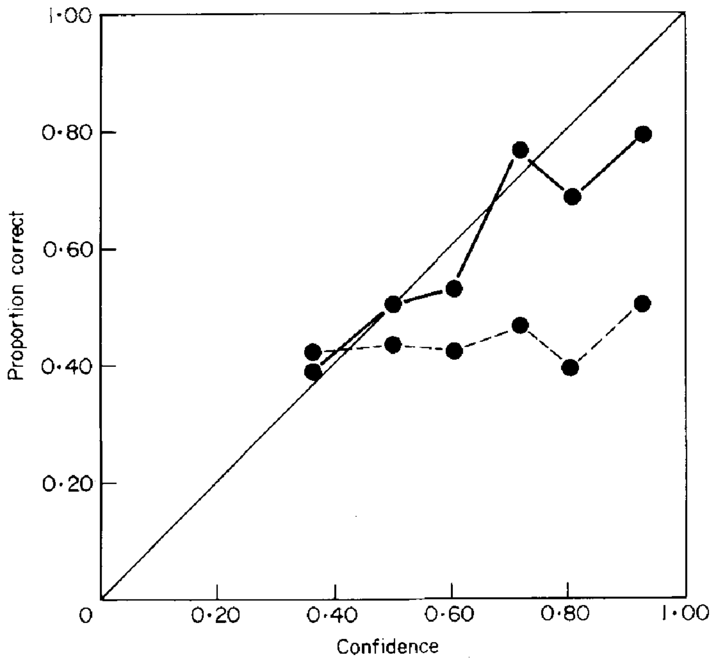

FIG. 1. Calibration of first choices for coarse partition and fine partition subjects. (Source: Fischhoff & MacGregor, 1986). – – –, coarse; ——fine.

in press; von Winterfeldt & Edwards, 1986) is devoted to devising procedures for characterizing the expected outcome of alternatives. In these schemes, an action's attractiveness depends upon the attractiveness of its possible outcomes and on the probability that it will produce them. The primary outcome of choosing a database option is either identifying or not identifying the correct category for a sought item of information.† From this perspective, it is essential that the users of a database be able to assess the probability of success with each option. We have called the property of letting users know where they stand, in this sense, *metatransparency*. One common measure of metatransparency is *calibration*, which looks at the difference between expected and experienced success over a series of predictions. Perfectly calibrated users would choose the correct category on XX% of the occasions on which they believe that they have an XX% chance of success (Lichtenstein, Fischhoff & Phillips, 1982; Murphy, 1972).

Figure 1 provides a graphic representation of metatransparency in the form of calibration curves, contrasting expected and experienced degrees of success (i.e. subjects' probability of being correct as a function of their confidence in their choices). The specific example concerns two sets of categories providing access to the data stored in the *Statistical Abstract of the United States*. The fine partition is the 33 chapters in the *Abstract's* Table of Contents; the coarse partition is a set of eight

† More sophisticated decision-making models (e.g. Katz, Murphy & Winkler, 1982; Krzysztofowicz, 1983) could consider such additional outcomes as the cost of search (e.g., the time spent doing it) and the cost of different errors (e.g. not getting the desired information, mistakenly accepting another item). With transparent systems, the users' main task is evaluating the relative usefulness of different kinds of information. Even the most carefully designed database will leave users unsatisfied if they are confused about what information they really want, or misestimate the costs of getting the wrong information (Fischhoff, Slovic & Lichtenstein, 1980; March, 1978).

TABLE 1
Investigator-produced categories and labels for the Statistical Abstract of the United States.

Category name	Subordinate chapters†
(A) Census Information	1 Population 2 Vital Statistics 3 Immigration and Naturalization 32 Outlying Areas under US Jurisdiction 33 Comparative International Statistics
(B) Health & Welfare	4 Health & Nutrition 11 Social Insurance & Human services
(C) Environment	7 Geography & Environment 8 Public Lands, Parks & Travel
(D) Government	6 Law Enforcement, Courts & Prisons 9 Federal Government Finances 10 State & Local Government Finances 12 National Defense & Veterans 16 Elections
(E) Education & Science	5 Education 21 Science
(F) Commerce	17 Banking, Finance, & Insurance 18 Business Enterprise 19 Communications 22 Transportation—Land 23 Transportation—Air & Water 30 Domestic Trade & Services 31 Foreign Commerce & Aid
(G) Personal Finance	13 Labor Force, Employment & Earnings 14 Income, Expenditures & Wealth 15 Prices
(H) Industry	20 Energy 24 Agriculture 25 Forest & Forest Products 26 Fisheries 27 Mining & Mineral Products 28 Construction & Housing 29 Manufactures

† Numbers refer to the order of each chapter in the *Abstract*.
Source: Fischhoff and MacGregor (1986).

superordinate categories of our own creation (see Table 1). In an experimental test (Fischhoff & MacGregor, 1986), these two sets showed moderate differences in transparency; on their first choice, subjects correctly identified the coarse category holding 11 test items 44% of the time and the correct fine category 62% of the time. There was, however, an enormous difference in metatransparency. With the fine categories, as confidence in having selected the correct category increased, so did the likelihood of having done so, as evidenced by the upward slope of the calibration curve. On the other hand, coarse category subjects were correct about 40% of the time regardless of their confidence. Such poor calibration should not only reduce users' ability to gamble wisely when choosing categories, but also lead to frustration with the database, which seems to behave unpredictably.

Thus, transparency and metatransparency are separate criteria, both of which need to be considered when evaluating databases (and the categories they use) and when predicting how people will interact with them. At times, the database designer may be forced to make a tradeoff between these two criteria, perhaps even selecting a design that is less transparent, but gives users more realistic expectations.

HOW CAN INTUITIVELY MEANINGFUL CATEGORIES BE CREATED?

Having sharply defined criteria carries, of course, no assurance of having attractive alternatives to evaluate. There are several possible sources of category sets (Savage & Habinek, 1984; Snyder, Haap, Malcus, Paap & Lewis, 1985). One source is the supplier of the data who, presumably, has some traditional way of organizing it. That organization will be effective only if it also reflects how most users think about that content area. That is likely if the supplier has somehow shaped public thinking. It is less likely if the supplier is so removed from its intended users that it has little opportunity to see how they think or what difficulties they are having with the database. Such mismatches may be a common occurrence with current attempts to provide wide access to databases that were previously used only by specialists. A second source of categories is categorization experts, such as librarians. They attempt to optimize a number of criteria, including machine search time, storage requirements, logical distinctiveness, and usability. It is not clear to what extent these criteria support or conflict with one another. However important user accessibility is, it is most likely to be achieved where categorization experts either know or have shaped users' thinking.†

A third source of categories is the users themselves. Users can be involved reactively, seeing how well they can use categories produced by others. If their performance is poor, then designers can change the system according to their intuitions about what went wrong. Or, users can be involved actively, seeing how they would organize the raw material of the database (Dumais & Landauer, 1984; McDonald *et al.*, 1982; Nakamura, Sage & Iwai, 1983; Snyder *et al.*, 1985). If they can articulate their own perceptions and if those perceptions are shared by other users (Lee *et al.*, 1984), then the result should be categories that match users' mental representation of the domain—although these categories might not serve the other goals of the system. The main threats to this strategy would be if users could not introspect on their own mental processes or anticipate how the categories would actually be used (Broadbent, Fitzgerald & Broadbent, 1986; Ericsson & Simon, 1984; Nisbett & Wilson, 1977).

The present studies offer two general approaches to eliciting potential users' mental representation of the categories underlying a domain. The more structured approach asks people to assign elements to categories whose labels have been determined by the investigator; the success of this categorization is tested by other users' ability to tell which categories contain various items of information. In this case, the "elements" are the *Abstract's* chapters and the "items" are facts it

† Suggestive research regarding gaps between the categorization processes of experts and laypeople may be found in Adelson (1984), Chi, Feltovich and Glaser (1981), and Murphy and Wright (1984). Within the information-science literature, there is considerable evidence of disagreement among experts themselves regarding categorization (including indexation) schemes (see reviews in Cooper, 1978; Furnas *et al.*, 1983; Lee *et al.*, 1984).

contains. A potential advantage of this approach is allowing substantive experts to propose well-defined category labels that divide the universe of content in a logically sound and efficient way. A potential disadvantage is the possibility that experts and users see this universe so differently that the experts' labels have little intuitive appeal.

The second, less structured approach allows nonexperts to create their own categories by organizing elements as they see fit. The study reported here placed no constraints at all. More structured variants might specify the number of categories or the number of elements per category—at the risk of restricting subjects' ability to express their mental representations (Dumais & Landauer, 1984). The question of how many categories to use (sometimes known as the depth/breadth tradeoff) weighs the additional information provided by a finer partition against the compactness of a coarser partition. Studies with artificial databases (Kiger, 1984; Landauer & Nachbar, 1986; Snowberry, Parkinson & Sisson, 1983a) suggest that finer partitions are more transparent. Figure 1 echoes that conclusion, as well as showing greater metatransparency with the finer partitions.

Once a consensual categorization has been derived, meaningful labels are needed. The usefulness of categories can be enhanced by labels that convey their organizing principle (Baraslov, 1983; Murphy & Medin, 1985; Nakamura, 1985). Studies by both Furnas *et al.* (1983) and Lee *et al.* (1984) have found that lay users can describe their own perceptions well enough make systems more transparent to people like themselves.

HOW SHOULD THE CATEGORIES BE PRESENTED TO USERS?

Category systems have an inherent ambiguity, insofar as each category label necessarily represents items of some diversity (Homa, 1984; Kiel, 1981, Zadeh, 1965). Somehow, users need to learn what is meant by category labels that others have created. With computerized databases, users typically receive just the labels, although explanatory material may be available in online help menus or offscreen documentation. Any residual ambiguities must be resolved by trial and error. With hard-copy databases, users often receive the labels together with subsidiary information, such as the labels of subsections or even some exemplary items. Presumably, this additional information helps users discern what is meant by the general labels. However, it comes at the cost of cluttering the display. The more details that are provided, the harder it becomes to find those that are needed. It is, of course, the desire to avoid clutter that motivates the creation of categories.

Thus, the amount of explanatory detail is one design variable that is likely to affect categories' usability, possibly in a curvilinear way. Adding detail should help (Dumais & Landauer, 1984; Snowberry, Parkinson & Sisson, 1983b; 1985) until it becomes so voluminous that it constitutes clutter.

The studies

Three sets of studies are reported below. The first illustrates the general methodology, using the results in Fig. 1 as a case in point. The second set examines the effect of detail in presentation on the usability of categories. The final set examines two approaches to eliciting and labeling users' mental representations of a content area. Throughout, transparency and metatransparency are used as performance criteria. The resulting statistics could be used to assess the adequacy of an interface design or to predict performance with it. Further access to the research literature regarding these criteria, and the study of behavioral decision making from which they are drawn, may be found in Fischhoff (1986), Fischhoff and MacGregor (1987), Lichtenstein *et al.* (1982), Pitz and Sachs (1984), and Wallsten and Budescu (1983).

PERFORMANCE CRITERIA

As discussed earlier, two desirable features in a database are transparency and metatransparency. The former, ease of finding items in the database, can be measured by the proportion of correct category choices. The latter, an accurate appraisal of one's ability to use the database, can be measured by *calibration,* which contrasts expected success with actual success. Calibration can be seen graphically in a representation like Fig. 1 which presents the proportion of correct responses associated with each probability response (grouped into intervals, such as 0·50–0·59, 0·60–0·69, etc., so as to ensure stable estimates). Calibration is imperfect whenever the empirical curve deviates from the identity line. It is commonly measured by the mean squared vertical distance of the points on the curve from the identity line (weighted by the number of observations incorporated in each). A simple measure of the overall trend in miscalibration is *over/underconfidence,* equal to the difference between the overall mean probability and the overall proportion of correct category choices. A positive score indicates overconfidence, in the sense of expecting to be correct more often than is actually the case.

A final measure, providing some supplementary information, is *resolution,* equal to the variance in the proportions correct associated with different degrees of confidence (again, weighted by the number of responses involved). It reflects the ability to discriminate different levels of knowledge, even if one cannot assign them the probability value that would assure perfect calibration. One can be very well calibrated yet show poor resolution, for example, as a result of guessing blindly about two alternative choices and assigning 0·5 to each choice. And, one can show good resolution yet be poorly calibrated, for example, by always saying 0·51 when one is correct and 0·49 when one is incorrect.

METHOD

In day-to-day experience, with databases or other uncertain systems, people treat their uncertainties informally, seldom assigning an explicit probability to their state of incomplete knowledge (Beyth-Marom, 1982). In order to evaluate the appropriateness of their confidence, it is necessary to make those implicit feelings explicit. Fischhoff and MacGregor (1986) took this step in the context of databases by asking potential users of the *Abstract* to choose the first, second, and third most likely categories within which to find each of 11 informational items. After making their choices, subjects then indicated the probability that each was correct by distributing 1·00 of probability across the three choices and the complementary "All Other Categories". Participants in this study experienced little difficulty in providing probabilities. The potential locations were either the 33 *chapters* in the *Abstract* or eight superordinate *categories* of our creation. The items were facts such as "The percentage of female elementary-school teachers" and "The mean annual temperature in Bismark, ND". They are presented in full by Fischhoff and MacGregor (1986).

All subjects in these studies were recruited by advertisements for paid volunteers appearing in the University of Oregon student newspaper. They were divided roughly evenly between men and women, with the average age of the former being 24 and of the latter 21. Two thirds were students and the remainder somehow affiliated with the university community. As a group of educated young adults, they represent a population of potential users for the *Abstract* (and many other databases).

On the basis of our past experience, which indicated that special instruction about the meaning of probability is not needed, respondents were told no more about the response mode than "Estimate the probability that the item will be in each chapter [category] or under 'Some Other Chapter [Category]'. Those probabilities should be numbers from 0% to 100% and add up to 100%". A small number of subjects consistently failed to follow instructions and their responses were deleted.

Problems were presented and responses were recorded on pencil-and-paper questionnaires administered in a group setting. MacGregor, Fischhoff and Blackshaw (in press) replicated the task of Fig. 1, and several other like it, with individual subjects using a computer-interactive format. The change of format produced few notable differences.†

Demonstration study

RESULTS

Transparency

Table 2 shows the performance statistics accompanying Fig. 1. The "conditional proportion of correct category selections" (line 1) reflects the proportion of subjects selecting the correct choice among those who had yet to choose correctly. The "cumulative" proportion (line 2) shows how many subjects had chosen correctly by the end of a given round. On first choices, subjects receiving the chapters were almost 50% more likely to answer correctly (0·616 vs 0·437). On subsequent choices, performance of the two groups was more similar. Subjects who had yet to identify the correct location did so about one third of the time on their second and third choices. Indeed, by the end of the third round, category subjects were almost as likely to have located items as were chapter subjects.

For first choices, the greater precision of the fine partition apparently provided enough additional information to compensate for the larger number of options it

† The largest of those differences was an increased tendency to choose just one or two possible locations for items with one version of the computer interactive task. Dividing 1·00 of probability over a smaller number of alternatives means expressing greater mean confidence in those selections that are made. There was, however, no corresponding increase in the proportion of correct selections. The result was considerable overconfidence and poor calibration. This deterioration in performance seems to have been due to arbitrary features of the interface design which reduced how hard users thought about alternative locations.

TABLE 2
Selected performance statistics

	Categories			Chapters		
	1st	2nd	3rd	1st	2nd	3rd
Transparency						
(proportion of correct category selections)						
Conditional	0·437	0·379	0·346	0·616	0·395	0·338
Cumulative	0·437	0·651	0·772	0·616	0·761	0·824
Metatransparency						
Simultaneous search						
Proportion correct	0·437	0·237	0·155	0·616	0·249	0·112
Mean confidence	0·599	0·225	0·112	0·676	0·201	0·093
Over/underconfidence	0·162	−0·012	−0·043	0·060	−0·048	−0·019
Calibration	0·056	0·007	0·005	0·009	0·005	0·002
Resolution	0·003	0·002	0·000	0·024	0·007	0·003
Number of responses	670	666	652	427	406	365
Sequential search†						
Mean confidence	0·599	0·569	0·607	0·676	0·617	0·728
Over/underconfidence	0·162	0·190	0·251	0·060	0·222	0·391
Calibration	0·056	0·065	0·110	0·009	0·094	0·227
Resolution	0·003	0·003	0·001	0·024	0·001	0·000
Number of responses	670	377	227	427	157	80

† Proportion correct is equal to the conditional proportion of correct category selections (line 1 of table).

Source: Fischhoff and MacGregor (1986).

forced subjects to examine. On subsequent choices, however, chapter subjects were either less able to examine all those options thoroughly or else unable to extract additional information from them. Perhaps the best-guess chapter seemed so right that alternatives did not seem very credible, whereas the categories were sufficiently ambiguous that alternatives stayed alive longer as possibilities.

The third line of the table, under simultaneous search, shows the proportions of all choices that were correct, including, for second and third choices, cases in which a preceding choice was correct. If subjects waited for the outcome of each choice before proceeding to additional ones, then these subsequent choices would not get made. They would be made, however, in a system that elicited several requests at once for batch processing—hence the name *simultaneous search,* as compared with *sequential search,* in which each option is examined in turn.

Metatransparency

The remainder of Table 2 describes the metatransparency of these partitions. The fourth line shows subjects' mean confidence in their selections. For first choices, both groups had greater mean confidence than proportion correct, a discrepancy reflected in the positive overconfidence scores of the following line (which equal, simply, the difference between these two statistics). Given subjects' great confidence in their initial choices, very little probability is "left over" for second and third choices. Probabilities for those choices were, in fact, sufficiently low as to bring them in line with subjects' level of knowledge, leaving a weak overall tendency to underconfidence.

The poor calibration curve (in Fig. 1) for the first choices of subjects using the coarse partition is reflected in their calibration and resolution statistics. The former is very large (0·056), in keeping with the large distance between the calibration curve and identity line. The latter is very small (0·003), in keeping with the negligible variability in the proportion of correct responses associated with different levels of confidence.

The metatransparency of the fine partition (the chapters) is markedly better. The first choice curve has an upward trend, indicating that knowledge increases with confidence. The calibration statistic is much smaller (0·009), reflecting the greater proximity of the curve to the identity line. The resolution statistic is much larger (0·024), reflecting the sensitivity of confidence to knowledge. One summary of this curve is that subjects can distinguish (or "resolve") three levels of knowledge, corresponding roughly to 40%, 50%, and 70% correct, to which they assign probabilities of under 0·50, between 0·50 and 0·65, and 0·70 and above, respectively. The adequacy of such metatransparency for particular users and search problems is an empirical and analytical question.

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 for the third than for the second. From this (simultaneous search) perspective, subjects' expectations were about as realistic here as with the first choices. The curves with the open circle are the result

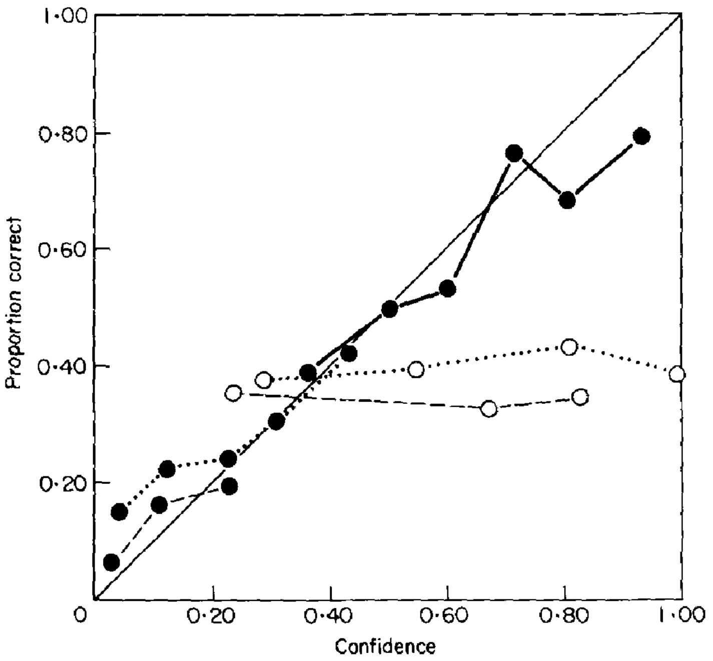

FIG. 2. Calibration of simultaneous probabilities (closed circles) and of sequential probabilities obtained by conditionalizing simultaneous probabilities for second and third choices (open circles). Fine partition subjects. (Source: Fischhoff & MacGregor, 1986.) ——, first choice; . . . ; second choice; – – –, third choice.

of asking 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 (0·60, 0·30, 0·05, 0·05), then the implicit conditional probability for the second choice is 0·75 (= 0·30/(1·0 – 0·60)), while for the third it is 0·50. From this (sequential choice) perspective, subjects' performance was much poorer than with the other (simultaneous choice) perspective. They seem poorly attuned to the details of just how confident to be that "now they had the right one".†

As mentioned, coarse partition subjects' overall confidence in their second and third choices was quite appropriate. Within those choice sets, however, there was little relationship between confidence and knowledge. This is seen in flat calibration curves (not shown) and near-zero resolution scores.

DISCUSSION

The performance statistics of Table 2 provide a way of describing both users' ability to exploit a database and their understanding of its exploitability. They are used here to characterize two alternative entry-level menus for a given database. Although *a priori* reasons could be raised for predicting the superiority of either menu, the present performance measures showed the fine partition to be uniformly better. Despite presenting four times as many alternatives to consider, that partition gave people a better chance of choosing correctly and a better feel for their level of knowledge. As a result, they should be better able to take failure in stride and move on to refine their mental model of the system. Such metatransparency also leaves users better equipped to decide whether any of the options are attractive gambles, or if they are better off withdrawing from the menu and learning more about the system, in order to sharpen their probability distribution over the options (Bates, 1977).

What probability of being correct is high enough to justify selecting a location should depend on the value of the answer being sought, the cost of the search, the cost of being wrong, and the opportunities for learning from one's mistakes (Blair, 1980). The first of these factors depends upon the practical purpose of the search. The second depends upon properties of the system. The third depends upon properties of both the system (e.g. how long an error can go undetected) and the decision problem (e.g. what happens if the wrong information is used). The fourth depends upon what feedback the system provides and how well users can exploit it. Poor calibration suggests a poorly differentiated mental model of the system, meaning that even the reasons for correct answers are not always well understood.

† In point of fact, however, these subjects performed a simultaneous choice task. A separate study (in Fischhoff & MacGregor, 1986) 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) conditionally 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 subsequently considered choices seem particularly unlikely. The recommendation implied by these results is to encourage searchers to consider several alternatives prior to beginning their search, which fits with other psychological results (e.g. Koriat, Lichtenstein & Fischhoff, 1980; Slovic & Fischhoff, 1977).

If forced to choose between these two menus, the chapters clearly dominate. In some situations, however, that choice might not be on offer. For example, technical constraints might limit the number of options appearing on a menu, forcing some categorization of the chapters. The following sections consider ways of creating more viable categories and of improving the usefulness of the categories that one has.

How should items be assigned to categories?

In the preceding studies, the editors of the *Abstract* provided the fine partition, whereas we provided the coarse one. Thus, in neither case, did users control either which categories were offered or how elements were assigned to them. A modest strategy for relying on potential users to make a partition more meaningful is to keep a fixed set of categories but have users assign elements to them. A more ambitious strategy, described later, is to have users create the categories themselves. Both are undertaken here in an attempt to improve the coarse partition.

METHOD

Subjects in the *fixed category group* received a questionnaire akin to those used above. Their items were the 33 chapters, and their task was to locate them in our eight categories. Subjects were told that the *Abstract*

> "... has 33 chapters, each containing different kinds of information. Currently, they are just presented in a list, one after another. The following set of categories has been proposed to give the *Abstract* an internal organization. [Category labels appear here in list form.] In order to help evaluate the usefulness of this organization, we would like you to judge where you would expect to find each chapter. Please do so in the following way: (1) read the list of categories; (2) read the list of chapters; (3) for each chapter, first choose the category that it seems most likely to be in. Write that in the space indicated. Then choose the second most likely category and the third most likely category, writing them in as well. Finally, estimate the probability that the chapter will be in each category or under 'Some Other Category.' Those probabilities should be numbers from 0% to 100% and add up to 100%".

RESULTS AND DISCUSSION

Table 3 presents the percentage of subjects choosing each category as being the most likely place to look for each chapter. The underlined percentage corresponds to the category to which we ourselves assigned each chapter. For the first 12 chapters, our choice and that of the modal subject coincided. Those choices differed, however, for Chapters 13, 15, 18–20, 24–26, and 32–33, that is, for 10 of the 33 cases. In addition, even when the modal subject agreed with us, many others often did not. Overall, subjects' first choice agreed with our choice 58·0% of the time. This rate is presented as "proportion correct" in Table 4, which presents performance statistics for this task, treating subjects' responses as attempts to locate the chapters in the categories. The left-hand side of the table treats subjects' location selections according to where we (the experimenters) had placed the chapters. Of subjects who chose wrong initially, 37·2% would have found the correct category on their second choice, with 31·6% of the remainder finding it on the third choice, by which time 80% would have found the correct category. This

TABLE 3
Assignment of chapters to categories

	Categories							
Chapters	Census information	Health Welfare	Environment	Government	Education Science	Commerce	Personal finance	Industry
(1) Population	98	2						
(2) Vital Statistics	43	39	5	5	5	5		
(3) Immigration and Naturalization	61	11		20		7		
(4) Health and Nutrition	2	98						
(5) Education		2			98			
(6) Law Enforcement, Courts and Prisons	7	9		73		7	2	2
(7) Geography and Environment			98			2		
(8) Public Lands, Parks, and Travel		7	66	16		11		
(9) Federal Government Finances				95		2	2	
(10) State and Local Government Finances				86		14		
(11) Social Insurance and Human Services	2	93		5				
(12) National Defense and Veterans	2	2		95				
(13) Labor Force, Employment and Earnings	23	2		11		16	5	43
(14) Income, Expenditures and Wealth	14	2		7		9	61	7
(15) Prices	5	2		7		55	14	18
(16) Elections	9			86	2	2		
(17) Banking, Finance and Insurance				7		61	23	9
(18) Business Enterprise				7	2	25	2	64
(19) Communications	5	5	2	21	35	23		9
(20) Energy		5	52	9	18	2		14
(21) Science			2		95			2
(22) Transportation—Land		7	18	16	2	43		14
(23) Transportation—Air and Water		7	20	16	2	39		16
(24) Agriculture		5	43	5	5	18		25
(25) Forest and Forest Products			52	2	2	11		32
(26) Fisheries			52	5	2	18		23
(27) Mining and Mineral Products			32	5		11		52
(28) Construction and Housing	2	11	5	7	2	11	2	59
(29) Manufactures				2		7		91
(30) Domestic Trade and Services		5		9		61	5	20
(31) Foreign Commerce and Aid				48		50		2
(32) Outlying Areas under U.S. Jurisdiction	14		5	74	2	2	2	
(33) Comparative International Statistics	35			53		9		2

seems like a relatively low percentage, considering that there are only eight options, several of which seem completely inappropriate for each chapter.

The remaining rows under "Experimenter Key" reveal overconfidence similar to that observed with the coarse partition group in Fig. 1 and Table 2. Unlike that group, however, subjects here were fairly sensitive to their relative degree of knowledge, as can be seen in the fairly low calibration score and fairly high resolution score. The corresponding calibration curve (not shown) had a pronounced

TABLE 4

Performance statistics for assigning chapters to categories

	Experimenter key			Subject key		
	1st	2nd	3rd	1st	2nd	3rd
Transparency						
(Proportion of correct category selections)						
Conditional	0·580	0·372	0·316	0·676	0·465	0·335
Cumulative	0·580	0·732	0·804	0·676	0·823	0·876
Metatransparency						
Proportion correct	0·580	0·160	0·085	0·676	0·156	0·063
Mean confidence	0·744	0·165	0·071	0·744	0·165	0·071
Over/underconfidence	0·164	0·006	−0·014	0·067	0·009	0·008
Calibration	0·055	0·005	0·001	0·031	0·003	0·002
Resolution	0·020	0·004	0·001	0·023	0·005	0·000
Number of responses	1449	1372	1245	1449	1372	1245

upward slope, falling just below the fine partition curve in Fig. 1, absent the irregular "bump" at 0.7. Similar calibration curves for tasks with similar proportions correct is a common result (Lichtenstein *et al.*, 1982). From a simultaneous search perspective, subjects' overall confidence in their second and third choices was highly appropriate. Those choices were not often correct and were not viewed with much confidence. The corresponding calibration curves (not shown) and statistics show modest sensitivity to relative degrees of knowledge within these choices. From a sequential search perspective (not shown), however, subjects were (again) considerably overconfident in both cases. Had their first choice failed, then they would have been almost as confident of finding the chapter in their second choice (0·645), but much less likely to do so (0·372). The same applies for the third choice (mean confidence = 0·780; knowledge = 0·316).

The right side of Table 4 scores individual subjects' location selections according to a key based on all subjects' modal first choice. As mentioned, this involves changing the categorization of 10 chapters. Clearly, the percentage of correct first choices must increase and the degree of overconfidence decrease (since confidence levels are unchanged). However, the amount of change is of interest, as is the effect on performance with second and third choices. First choice transparency increased by 0·096, while calibration went from 0·055 to 0·031. These improvements would seem to justify the present modest investment in discovering what these category labels mean to subjects. Second- and third-choice performance was relatively unchanged.†

Before endorsing this subject-created classification system, one should examine its other properties. A rather technical one is the distribution of chapters across

† Here, as in the studies reported below, similar patterns emerged for the metatransparency of second and third choices. From the simultaneous search perspective, the calibration curves clustered around the lower end of the identity line, sometimes slanting upward, sometimes lying flat. From the sequential search perspective, the curves were flat and below the identity line, as in Fig. 2. Only results for the former perspective will be reported.

categories. High transparency could have come from lumping most chapters into a few categories. However, this was not the case. Our classification system had two categories with 7 members, two with 5, one with 3, and three with 2. Subjects' key had one category with 7 members, two with 6, one with 5, two with 3, and one each with 1 and 2. These are quite similar distributions, each of which might be faulted for not having a more even spread of chapters across categories (although that is a function of the chapters as well as the categories).

A more significant test is whether the subject key is useful for locating items as well as chapters. Changing from the experimenter key to the subject key shifted four of the 11 items from one category to another (along with the chapters containing them). Scoring responses of the coarse partition group with the new key *decreased* the proportion of correct first choices modestly overall (from 0·437 to 0·351) and dramatically for three of the four moved items. The drop in percentage correct, coupled with the same confidence assessments, led to an increase in overconfidence, with calibration worsening (from 0·055 to 0·091). The accompanying curve (not shown) was flat and lay beneath that of Fig. 1.†

Figure 3 shows further details for two of the items where performance deteriorated with the subject key. With Item 1, almost all fine partition subjects (i.e. 87%) thought that Energy was the most likely chapter for a question about nuclear power; most (i.e. 52%) of the present subjects thought that the Energy chapter

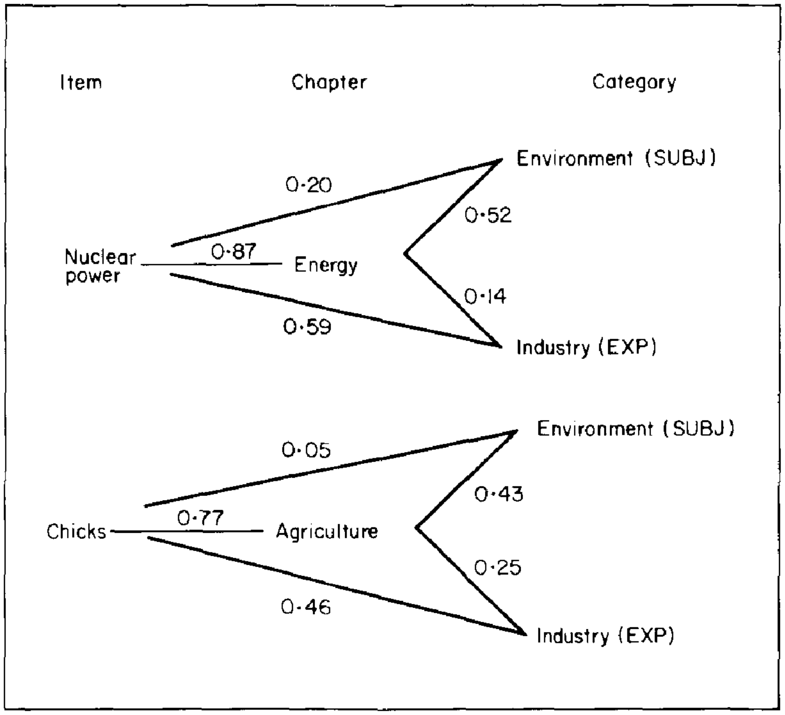

FIG. 3. Proportions of subjects placing items in chapters, chapters in categories, and items in categories for those chapters chosen either by experimenters (EXP) or by the modal subject in the fixed-category method group (SUBJ). Full wording of the first item was "The number of operative nuclear power plants". Full wording for the second item was "The number of chicks hatched in the U.S. yearly".

† Performance on second and third choices was similar with the two keys.

belonged in Environment category; however, when faced with Energy and Environment as category options (among others), coarse partition subjects picked environment by a 3:1 margin (0·59 to 0·20). Thus, the terms considered here (nuclear power, energy, environment, industry) seem to have been sufficiently ambiguous to evoke different associations in different contexts. By like token, in the second example, Chicks fit best in Agriculture. However, whereas Agriculture belonged more to Environment than to Industry, Chicks went the other way.

If the meanings of terms depend on their context, then categorization becomes problematic. The subjects who created this key saw both the chapters and the categories. The (coarse partition) subjects whose responses were evaluated according to the subject-created key saw the categories and items, but not the chapters. Presumably, their interpretations of these terms would have matched better had they had similar exposures. By like token, our familiarity with the *Abstract* limited our ability to create categories that would be meaningful to less-informed users. Similar differences in perspective could account for the limited transparency of the 33 chapters created by the actual editors of the *Abstract* (with subjects' first choices identifying the correct chapters only 60% of the time). The next pair of studies attempts to improve the match between the perspectives of category users and category creators, by offering more detail on category contents. In interactive systems, such elaborations could be presented online or in hard-copy adjuncts.

How should categories be presented to users?

Assuming that categories are used most effectively when interpreted similarly by users and creators, one should attempt to equate the experiences of those two groups. That could be done by having users browse extensively before starting any specific search. Where that is impractical (e.g. because users are unwilling to invest the time, because search costs are prohibitive, because the database is very large), more directed exposure is needed.

The following studies explore the effect of offering users the next level of detail for the menu they are asked to use. *Elaborated coarse partition* subjects received the categories along with the chapters that they contain. *Elaborated fine partition* subjects received the entire six-page Table of Contents from the *Abstract,* showing the 348 subsections in the 33 chapters. In both cases, the additional detail might improve performance (e.g. by clarifying the meaning of the categories, or by encouraging users to think about them more), leave performance unchanged (e.g. because there was too much information to absorb, or because the underlying scheme did not make that much sense), or even reduce performance (e.g. because the information overloaded the user, or because its details caused confusion).

After completing this task, subjects located the items in the subordinate categories which had just been used in a clarifying role. Conflicting predictions are also possible regarding the effect on performance of having already sought the items within superordinate categories and of seeing a set of superordinate categories that attempt to organize them. Using an artificial database, Snowberry *et al.* (1983b) found that presenting subordinate categories was helpful whereas presenting superordinate ones was not.

METHOD

Each subject attempted to place the 11 items in two separate category schemes. For the elaborated coarse partition group, the first attempt involved the eight coarse categories and the second attempt involved the 33 chapters. Their tasks were identical to those of the previous coarse and fine partition groups except that they saw the chapters organized by category. Their second task was introduced with, "Now, we would like you to go back over the same set of facts and indicate for each the *chapter* in which it is most likely included". Nothing was said about maintaining consistency across their two tasks. For the elaborated fine partition group, the first attempt involved the 33 chapters and the second attempt the 348 subsections of those chapters, all of which was shown to subjects throughout both tasks. Subjects indicated their choice of subsection by writing the page on which the subsection began.

RESULTS AND DISCUSSION

Adding the defining chapters dramatically improved subjects' performance with the categories. As can be seen in the upper left quarter of Table 5, the proportion of correct first choices increased from 0·437 to 0·639, slightly higher than that for the fine categories alone (0·616). Combined with a smaller increase in confidence (from 0·60 to 0·68), this increase in transparency was accompanied by a marked improvement in metatransparency. Overconfidence dropped from 0·162 to 0·309; the calibration score dropped by 60%, whereas the resolution score quadrupled. There were also improvements on second and third choices. Conditional proportions correct were 0·438 and 0·410 (compared with 0·379 and 0·346 before and 0·395 and 0·338 for the fine partition group), so that by the third choice correct locations were identified in 86·5% of all cases (compared with 77·2% and 82·5% for the coarse and fine groups, respectively). Thus, the improvement was even better than one would have expected had these subjects gone straight to the chapters for their search. In some way, the categories helped to define the chapters.

As seen in the lower left corner of Table 5, elaborating the fine categories had much less effect. The proportion of correct first choices did rise from 0·616 to 0·669. However, confidence rose by the same amount, leaving metatransparency untouched. Overconfidence and resolution stayed at the same (modest) levels, while calibration deteriorated somewhat, and the calibration curve (not shown) was visually less impressive (than its counterpart in Fig. 1). This suggests that the subsections improved the transparency of the categories, however, the display was too large to be exploited systematically or to give a comprehensive overview of the database's organization (at least with the present level of study). On the second and third choices, most aspects of performance were actually somewhat inferior to that of the original fine partition group. Conditional proportions correct decreased to 0·320 and 0·200, reducing the cumulative proportion correct to 0·808 (from 0·824). Perhaps by this point in the search, the mass of detail was more of a hindrance than a help.

After placing items in the coarse categories, subjects in the elaborated coarse partition group placed them again, this time in the fine categories (see upper-right corner of Table 5). Performance here was essentially indistinguishable from that

TABLE 5
Performance statistics for assigning items under different category presentations

	Choice					
	1st	2nd	3rd	1st	2nd	3rd
	Coarse categories presented with chapters			Fine categories presented after coarse without subsections		
Transparency						
(Proportion of correct category selections)						
Conditional	0·639	0·438	0·410	0·602	0·421	0·342
Cumulative	0·639	0·796	0·865	0·602	0·766	0·836
Metatransparency						
Proportion correct	0·639	0·265	0·126	0·602	0·242	0·102
Mean confidence	0·678	0·218	0·094	0·684	0·215	0·098
Over/underconfidence	0·039	−0·048	−0·032	0·084	−0·028	−0·005
Calibration	0·021	0·010	0·002	0·016	0·005	0·007
Resolution	0·013	0·005	0·001	0·024	0·014	0·006
Number of responses	363	343	302	359	326	293
	1st	2nd	3rd	1st	2nd	3rd
	Fine categories presented first with subsections			Subsections presented after chapter using page key		
Transparency						
(Proportion of correct category selections						
Conditional	0·669	0·320	0·200	0·353	0·194	0·090
Cumulative	0·669	0·770	0·808	0·353	0·468	0·504
Metatransparency						
Proportion correct	0·669	0·214	0·062	0·353	0·133	0·053
Mean confidence	0·729	0·188	0·083	0·712	0·196	0·091
Over/underconfidence	0·059	−0·025	0·021	0·359	0·064	0·038
Calibration	0·015	0·008	0·001	0·158	0·019	0·016
Resolution	0·012	0·007	0·002	0·012	0·002	0·001
Number of responses	396	365	307	391	346	284

when the fine categories were presented alone, without the "aid" of being organized into the coarse categories. There was no practice effect from having worked with the items once. Depending upon the measure, performance was either slightly better or slightly worse here than with the coarse categories. Apparently, once the meaning of the categories has been clarified, subjects can exploit the fact that there are only eight of them for more sophisticated guessing.

After placing items in the fine categories, subjects in the elaborated fine partition group placed them again, this time in the subsections. As might be expected, this was quite a difficult task, with the overall proportion of correct choices being 0.353. This performance is poorer than that obtained with the initial coarse partition presentation (0·437). Nonetheless, this might still be preferred to that display. Although subsection subjects were correct on 8% fewer occasions, when correct they were within a few pages of the actual location of the needed information (rather than within a few chapters, for the coarse category subjects). Other aspects of performance with the subsections were less attractive. Subjects were highly confident in their highly imperfect category selections (mean = 0·712). The corresponding calibration curve (score = 0·158) showed only a slight upward tendency over the observed range; moreover, it was far below the identity line, as a reflection of the great overconfidence (= 0·359). The proportions of correct responses for second and third choices were very low (0·194, 0·090). Over three choices, the cumulative proportion correct was 0·504; the mean cumulative judged probability of being correct was 0·95. Thus, subjects were almost certain that they had found the right place by the end of their third choice, yet did so only in half of all cases. Apparently, the subsection titles seem so specific that they inspire great confidence that one can tell what they contain. Yet, often they do not.

Despite these limitations, approaching the *Abstract* through its subsections might still be a good idea if users who were wrong were at least “close” to the right answer. For a hard-copy database, like the *Abstract,* “close” might be defined in terms of physical proximity. For example, one might be able to flip forward and backward through neighboring pages if the chosen subsection were not productive. To examine this possibility, we defined “close” as “within the same chapter” and rescored the subsection selections on that basis. This more lenient scoring doubled the proportion of correct first choices (to 0·673). As confidence was unchanged, overconfidence was greatly reduced, although the calibration curve (not shown) remained relatively flat. The calibration score improved from 0·158 to 0·029, but that was because the curve now intersected the identity line, rather than being far below it. Resolution scores remained low (0·010 vs 0·012), in keeping with the flatness. By the end of their third choices, subjects had chosen a subsection in the correct chapter 77·7% of the time, compared with 80·8% to 83·6% of subjects locating the correct chapter directly in other conditions.

Overall, the fact that subjects often select the correct subsection suggests that they might as well be asked to do so, if such a display is possible. About 35% of the time their subsection selection is correct, leaving them very close to the target. In an additional 32% of the cases, their subsection is in the correct chapter, leaving them fairly close. The risk of using the subsections as an entry-level menu lies in subjects' poor calibration. The weak relationship between confidence and correctness may create a feeling of frustration and difficulties in directing their search. Perhaps the best advice to users is to “go directly to a subsection, but don't expect it to be right (however right it seems)”.

How can consensual categories be created?

METHOD

Subjects in the fixed category groups created categories in the constrained sense of determining where chapters would go under category labels that we created. The following studies greatly reduce these constraints. Subjects using the *sorting* method for category creation received a packet of 33 library file cards, each bearing the name of one chapter. They were told that:

> "We'd like you to sort these chapter titles into categories, such as might be used for sections in the *Abstract.* No such sections exist today and we are interested in developing an internal organization that might be suggested to the editors of the *Abstract.* To be useful, such an organization should make sense to potential readers of the *Abstract,* such as yourselves.
>
> In organizing the chapter titles, we would like you to use the following procedure:
> (1) read all of the cards;
> (2) tentatively sort them into a set of categories;
> (3) think of a name for each category;
> (4) see if that name applies to each chapter in the category. If not, try changing the label;
> (5) check to see if each chapter fits best with the category that it currently is in, or with some other category. If not, try moving the chapter;
> (6) *repeat steps 3 through 5 until you have a solution that you are happy with.*

They were given some blank cards for writing "tentative section labels . . . to help organize your work". In completing the task, subjects were explicitly allowed to use more than one label for a category and to put ungrouped chapters into a pile of "All Other Categories". They were told that it was a place to put "chapters that do not seem to fit anywhere else. Because it is not a very useful section heading, do not use it unless you have to. On the other hand, do not put chapters in sections where you think that it will be hard to find them afterward, because you would not think of looking there". Their goal was further specified as "The system of categories you develop should be one that helps you (or others like you) find information that you need. Ideally, four to nine categories might be best. However, it may be that you feel most comfortable with a large number of categories with a few items in each. Conceivably, you may see no opportunity for categorization at all—meaning that each chapter has its own category—although this is somewhat unlikely".

The obvious strength of such an open-ended procedure is allowing greater opportunity for subjects to express themselves. The price it exacts from subjects is working harder. The price it exacts from the investigator is having to combine the idiosyncratic views of different subjects into a consensual category set. The following analysis offers one approach to that task.

RESULTS AND DISCUSSION

Despite the open-ended nature of the task, subjects used fairly similar numbers of categories. Although the number ranged from four to 12, 81% of subjects produced from four to seven categories, with 12 producing four, 18 producing five, 13 producing six, two producing seven, five each producing eight and nine, and one each producing 10, 11, or 12. Only 26% of subjects availed themselves of the opportunity to use an "All Other" category. On the basis of these results, the consensual category set should have about six members and not include an "All Other" option.

In order to decide what those categories might be, a similarity matrix was created. Its rows and columns were the chapters. Each entry represented the number of times that a subject included two chapters in the same category. This matrix was then subjected to a hierarchical clustering routine in which "Initially, each variable is considered as a separate cluster. The amalgamating process continues in a stepwise fashion (joining variables or clusters of variables) until a single cluster is formed that contains all the variables". (Hartigan, 1981, p. 448) Because this process proceeds incrementally, identifying a set of categories still requires the exercise of judgment, particularly with regard to items falling close to the borders of more than one cluster (suggesting that there is more than one way to think about its identity). Table 6 presents a set of categories derived from our examination of the clustering results. Testing these categories requires them to be labeled, the topic addressed next.

TABLE 6

Proposed categories and chapters for the Statistical Abstract of the United States

Category name	Chapters
(I) Demography (Demographic information)	Population Vital Statistics
(II) Health, Education and Social Services (Social Services)	Health and Nutrition Education Social Insurance and Human Services
(III) Foreign Affairs and Immigration (Foreign Relations)	Immigration and Naturalization Foreign Commerce and Aid Comparative International Statistics
(IV) Government (Government)	Law Enforcement, Courts and Prisons Federal Government Finances State & Local Government Finances National Defense and Veterans Elections Outlying Areas under U.S. Jurisdiction
(V) Natural Resources and Science (Natural Resources)	Geography and Environment Public Lands, Parks, and Travel Energy Science Agriculture Forest and Forest Products Fisheries Mining and Mineral Products
(VI) Economy (Business)	Labor Force, Employment and Earnings Income, Expenditures, and Wealth Prices Banking; Finance and Insurance Business Enterprise Construction and Housing Manufactures Domestic Trade and Services
(VII) Communications and Transportation (Communications)	Communications Transportation—Land Transportation—Air and Water

Note: Initial label was produced by subjects. Label in parentheses was provided by investigators.

An example of an ambiguous chapter (in this particular context) is "Science". It was highly clustered with the National Resources group, but also had a moderate association with the first two members of the Social Services group (whose common thread would have to be redefined were Science included there). Another reflection of the problematic nature of Science is its being included in "All Other Chapters" by 11 of the 15 subjects who used that response option.

Labeling categories

Once categories are created, they need labels. Like the categories themselves, labels could come from either subjects or investigators, with similar tradeoffs in terms of potential convenience, coherence, and ethnocentrism. Both approaches are used here interactively: The *label production* group received just the categories of Table 6 and was asked to produce suitable labels. The *fixed subject category* group assigned the chapters to categories whose labels were produced by the label production group (much as the fixed category group did with our labels). These judgments were then used to produce a final set of subject labels. As a direct test of the usefulness of these new categories, the *coarse subject partition* group and *elaborated coarse subject partition* group attempted to locate the 11 items of information in them. The former received just the labels, while the latter received the label plus the category display of Table 6. They paralleled the coarse partition and elaborated coarse partition groups which attempted to use our original categories. Finally, as a test of how well we might have been able to do without this arduous label-production process, the *experimenter label* group had subjects judge the adequacy of and attempt to use labels that we created after the clustering analysis.

METHOD

The *label production* group received a form that opened with the categories of Table 6, about which they were told, "we [the experimenters] presented a number of individuals like yourself with a list of all the chapters in the *Abstract* and asked them to sort the chapters into categories that might be used to give the abstract an internal organization. [These are] the results of their efforts. . . . We would like you to suggest appropriate names for each of these categories, much as would appear in a table of contents. . . . Carefully read over the chapters comprising each category and give the category label you think best describes its contents". After doing so, they wrote their labels on a chart like Table 6 and then looked for the 11 items in those categories, using the usual simultaneous search procedure. Having done so, they were given the opportunity to review the set of categories and offer new labels for them (should they so wish). Finally, they judged the appropriateness of the category labels that we had created on a 10-point scale, anchored at 0 = not appropriate and 9 = highly appropriate ($n = 42$).

The *fixed subject category* group attempted to locate each of the 33 chapters in the seven categories as described by the modal label provided by the label production group. They made three choices and assigned probabilities in the same way as the fixed category group described earlier ($n = 38$).

The *coarse subject partition* group attempted to place the 11 items in the categories described by the seven subject-produced labels ($n = 42$).

The *elaborated coarse subject partition* group placed the 11 items in the seven categories, as described by both the subject-produced labels and the constituent chapters ($n = 42$).

The *elaborated experimenter label* group initially judged the appropriateness of the investigator-produced labels of Table 6 (in parentheses), providing alternatives where they saw fit. Then they attempted to place the 11 items in these categories. The first half of their task was used as supplementary information for developing the canonical subject-produced labels. The second half was used as a baseline to evaluate the contribution of letting subjects produce labels (by assessing performance with labels that we had produced prior to asking subjects for suggestions) ($n = 39$).

RESULTS AND DISCUSSION

Developing labels

The left half of Table 7 shows subjects' evaluations of our labels. "Done First" refers to the elaborated experimenter label group; "Done Last" refers to the label production group. The time of rating had no effect on ratings. In both cases, subjects gave our efforts reasonably high marks overall (means of 6·8 and 6·9 on the 0–9 scale). However, few labels were altogether satisfactory and at least one, Communications (VII) was judged to be quite poor. Where subjects were dissatisfied with our label (and even where they were not), they typically made the effort of providing an alternative. This tendency was more pronounced (74% vs 54%) with the label production group which had already invested in creating its own

TABLE 7
Acceptance of proposed labels

Subject-produced category number	Judging our labels				Judging own labels‡	Judging subject labels§	
	Done first†		Done last‡				
	Mean rating	Offer alt	Mean rating	Offer alt	Offer alt	Mean rating	Offer alt
I	7·1	36	7·4	79	79	7·2	50
II	7·6	60	7·3	74	67	7·8	30
III	6·4	55	7·0	57	48	7·4	50
IV	7·7	43	7·6	43	43	7·7	53
V	7·2	50	7·2	86	74	7·6	45
VI	7·1	48	6·9	81	83	7·9	33
VII	4·8	86	5·3	100	100	7·9	38
Mean	6·8	54	6·9	74	71	7·6	43

† Elaborated experimenter label group.
‡ Label production group.
§ Coarse subject partition group.

labels. Unfortunately, however, there was relatively little agreement among their suggestions—except that the great majority (79%) wanted the last category to be called Communication and Transportion (apparently belonging to a generation, unlike the investigators, in which the former does not imply the latter).

For the elaborated experimenter label group, the most common suggestion was to append the name of one category member to our label. For example, most of their suggested changes to our proposed "social services" involved adding something about health; the combination of high ratings for appropriateness and a high rate of suggested alternatives (Table 7) seems to reflect the perception that our label was on target, but only on part of it. The conjunction of the two terms seemed to capture the concept that underlay the sorting group's assignment of these chapters to a common category. Taken to the extreme (i.e. listing all members), using compound labels trivializes the notion of a category. However, as long as the display allows several words, adding terms to a label (e.g. Natural Resources and Science) may help ensure that secondary themes in a category are recognized.

Subjects in the label production group produced highly consensual labels for categories I, IV, VI, and VII (those appearing in Table 6). Their suggestions for Category II involved terms like health and social services, alone or in combination. Their initial suggestions for Category III usually began with "foreign" and seldom changed after doing the item-location task. Of the nouns following "foreign", "affairs" was most common and seems closest to the content of the category, except for the immigration chapter which we added to the title. Although few subjects made this suggestion, we thought that its location would be less obvious to users who, unlike present subjects, had not seen the category's contents (or been told to worry about categorization). These subjects' suggestions for Category V were distributed over natural resources (11), environment (11), science (6), ecology (4), and several unique labels. We settled on "natural resources", as it seemed more likely to connote the notion of exploitation than "environment" (whose alternative interpretations seemed to underlie the results depicted in Fig. 3). "Science" was added to the label because of its marginal role in the categorization, because it seemed further from "natural resources" than from "environment", and because of the label production group's suggestions.

Judging the labels

Thus, the "Subject Labels" were based on subjects' proposals, but adjusted in response to subjects' evaluations of our own proposals, to the disagreements among subjects' proposals, and to our own reading of potential problems for other subjects who had not seen the chapters underlying the categories. One test of these labels' adequacy is how they were rated by subjects who had tried to use them (in the coarse subject partition group). As shown on the right of Table 7, these labels were given high ratings for appropriateness. Although many subjects (43%) again offered alternatives, there was little consensus among their suggestions, beyond a recurring desire to add direct reference to items that they had sought. That desire suggests some limit to the labels' adequacy. However, satisfying it risks "overfitting" the label to particular exemplars.†

† However, representing categories by deliberately chosen examples holds some promise as a way of communicating category meaning (Baraslov, 1983; Dumais & Landauer, 1984; Kiel, 1981).

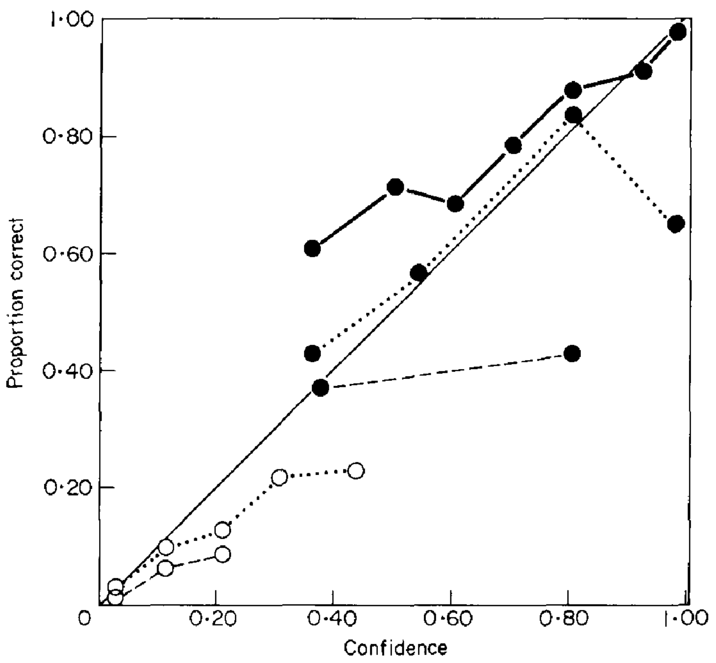

FIG. 4. Calibration for subjects locating chapters in subject-produced categories identified by subject-produced labels. Open circles indicate actual responses; closed circles indicate sequentialized responses. ——, first choice; . . . , second choice; – – –, third choice.

A more direct test of the labels is the fixed subject category group's ability to assign chapters to them, using the subject key in Table 6 as a criterion. Here, performance was outstanding. The conditional proportions correct for the three choices were 0·827, 0·934 and 0·958. These proportions are high in an absolute sense and high relative to the comparable results for assigning chapters to our category labels (Table 4). Only two chapters (2, 32) were not assigned correctly by at least two-thirds of subjects on their first attempt. For both, the correct choice was the second most popular selection, whereas the most popular first choice was inconsistent with their actual content; as a result, the labels were left unchanged.† One obvious contributor to this success was mentioning part or all of 10 chapter titles in the corresponding category label. Subjects almost unanimously identified the location of these chapters. Although only one of the other chapters elicited such near-perfect performance, by the end of the third choice 27 of 33 chapters had been correctly located by at least 95% of subjects. Figure 4 shows the corresponding calibration curves. All but the sequentialized third-choice probabilities (the curve for which was based on the relatively few occasions in which subjects had been wrong twice) are relatively good. The first choice curve, lying mostly above the identity line exemplifies the underconfidence typically observed with particularly easy tasks (Lichtenstein *et al.*, 1982).‡ It was, in fact, almost as far from the line as

† Specifically, the most common first choice for "Outlying Areas under U.S. Jurisdiction" was in "Foreign Affairs and Immigration;" for "Vital Statistics," it was "Health, Education and Social Services".

‡ The drop at the right extreme of the sequentialized second-choice curve shows that subjects' second choice was less likely to be right when they were certain that it was (and made no third choice) than when they left some probability for other choices.

the corresponding curve for subjects who attempted to locate the chapters under our category labels (Tables 3 and 4).† Thus, despite the improvement in transparency with subject labels, there was no change in calibration or resolution. Perhaps the metatransparency observed here reflects a general limit to people's ability to distinguish levels of knowledge with this kind of material—at least without some additional aids or training (Lichtenstein & Fischhoff, 1980).

Using the categories

Where people place chapters can help predict their performance in looking for the general kinds of information appearing as chapter titles. The statistics in Table 8 summarize subjects' performance in locating the 11 items in various representations of these categories.

The first set shows the cumulative result of all these efforts to produce a usable set of coarse category labels. It is directly comparable to the first group in Table 2, except that both categories and labels were produced by subjects. The contribution to transparency is marked. The conditional proportion correct is higher at each choice [(0·563, 0·546, 0·658) vs (0·437, 0·379, 0·346)]. By the third choice, the cumulative proportion correct is 0·907 (vs 0·772), a difference that seems only partially due to the smaller number of categories among which to choose. There was an improvement in metatransparency as well, as evidenced by the superior calibration and resolution scores. Rather than being flat, the calibration curves (not shown) for all three choices sloped upward. For first choices, the curve moved upward from about (0·4, 0·4) to about (0·6, 0·6), then levelled off to show no greater proportion correct with higher confidence. Having modest improvements in transparency prompt modest† improvements in metatransparency is a common result in calibration studies. Seeing them here indicates that there is nothing unique about these confidence assessments.

Results for the elaborated subject partition group show the effect of adding the contents of the categories to the display. The contribution is neither large nor consistent, to either transparency or metatransparency. This comparison suggests that a consensually defined set of labels may not need (or benefit from) elaborating its categories' contents.‡

Results in the lower left corner of Table 8 show how good performance would have been without any labels at all, beyond what users themselves provide. It is the best coarse category performance observed yet in terms of most measures (except resolution). The first-choice calibration curve (not shown) slopes upward from about (0·4, 0·6) to (0·95, 0·7) (the sequentialized second-choice curve looks quite similar). These results, too, suggest that there is relatively little value to the subject-produced labels—if the categories themselves can be presented.

The final set of results in Table 8 shows performance with our own labels atop the

† Although metatransparency is no greater than elsewhere, underconfidence at least means that surprises are more likely to be pleasant ones, with users more often finding material than users' confidence would lead them to expect. Instead of the frustration that may follow overconfidence, the risk with underconfidence is failing to exploit fully a database whose potential is underestimated.

‡ Displaying the categories substantially improved success for only one item, "The number of chicks hatched in the U.S. yearly", with the proportion of correct first choices increasing from 0·500 to 0·846. Along the lines of Fig. 3, chicks fit into Agriculture and Agriculture into Natural Resources and Science. It was less obvious that chicks went into that category when its contents were not laid out.

TABLE 8

Performance statistics for locating items in subject-produced categories

	Coarse subject partition (subject labels)			Elaborated subject partition		
	1st	2nd	3rd	1st	2nd	3rd
Transparency						
Conditional	0·563	0·546	0·658	0·604	0·647	0·500
Cumulative	0·563	0·794	0·907	0·604	0·844	0·905
Metatransparency						
Proportion correct	0·563	0·239	0·136	0·604	0·256	0·078
Mean confidence	0·654	0·235	0·089	0·727	0·187	0·073
Over/underconfidence	0·091	−0·004	−0·047	0·123	−0·069	−0·005
Calibration	0·030	0·005	0·013	0·042	0·008	0·006
Resolution	0·010	0·006	0·006	0·006	0·002	0·000
Number of responses	462	450	408	422	401	376
	Label production (own labels)			Elaborated experimenter label		
	1st	2nd	3rd	1st	2nd	3rd
Transparency						
Conditional	0·652	0·613	0·418	0·533	0·567	0·420
Cumulative	0·652	0·856	0·907	0·533	0·790	0·865
Metatransparency						
Proportion correct	0·652	0·213	0·060	0·533	0·265	0·088
Mean confidence	0·696	0·208	0·087	0·709	0·208	0·079
Over/underconfidence	0·044	−0·005	0·027	0·177	0·057	−0·009
Calibration	0·026	0·009	0·002	0·050	0·012	0·005
Resolution	0·003	0·003	0·001	0·008	0·004	0·004
Number of responses	451	434	379	458	442	394

subject-produced categories. They are slightly inferior in almost all respects to the other modes of presenting these categories, most notably the undue confidence that they inspire. That confidence came even though these subjects' first task had been to judge the appropriateness of each label and provide alternatives.

General discussion

Searching most databases involves a series of gambles, as users attempt to pick locations with a high probability of containing sought information. Whenever they cannot always find the right location, it is important that users know what their chances are. With realistic expectations, they can avoid premature frustration, seek help when needed, properly scrutinize the products of their search, and be ready to retrace their steps. Thus, ease of location and realism of expectations are separate performance criteria, for designing and evaluating databases and their interfaces. Using these criteria, we conducted a series of 11 experimental studies involving some 500 individuals, exploring different ways of creating, labeling, and displaying categories for organizing a natural database, *The Statistical Abstract of the United States*.

Our point of departure was a plausible partition of the *Abstract*'s 33 chapters into eight categories. The labels that we gave to these categories proved to have limited transparency and metatransparency. Organizing the chapters under these category labels according to subjects' (rather than our) judgments of where they belong *reduced* other subjects' ability to locate items of information in them. By contrast, allowing subjects to determine the structure of the categories produced a marked improvement in usability. Moreover, other subjects were able to create labels for these categories that stood on their own, attaining a level of usability that was obtained with the experimenter-produced labels only when their full contents were listed.

Listing the contents of categories may help users by clarifying the meaning of category labels (e.g. Environment) that could be construed in more than one way. Or, it may just allow them to circumvent a poor set of categories and go directly to the chapters. Listing needed contents is possible, of course, only where the display can accommodate the information and where users can absorb it. For example, the full list of chapters might be too much for some VDT displays; it might be an encumbrance for users needing only category-level information. In these studies, providing the full set of 348 subsections certainly would have gone beyond the limits of single-screen computer displays. The fact that it did little to help subjects looking for chapters suggests that it may have strained subjects' ability to use hard copy. Where full categories were displayed, performance was not improved further by adding labels generated by other subjects or by the investigators. Given the large number of names that people may append to objects (Furnas *et al.*, 1983) and to collections of objects, no one set of labels is going to appeal to everyone. However, the combination of the subject-generated labels for subject-generated categories communicated well enough to need no elaboration.

When given the unstructured task of organizing the chapters, subjects typically used an intermediate number of categories (four to seven), and did so similarly enough to allow a reasonably clear-cut group clustering. Performance was markedly better with those categories than with seemingly reasonable ones of our own creation. Subjects also proved to be a useful source of category labels. However, caution was needed in cases where subjects' exposure to the full set of chapters might cause them to interpret summary terms (e.g. Environment) differently than would people who see them solely as category labels. In that way, seeing a broader context (and perhaps even the act of thinking about categorization) may reduce the naivete that makes subjects a unique source of insight regarding the perspectives of lay users. Thus, when soliciting the views of people like the ultimate users, it is important to equate familiarity with the system. Of course, naive users' own perspectives may change with experience—if they persist with the system. The greatest value of a user-centered system may be in making the initial encounter with a system sufficiently satisfying that users will persist long enough to learn its idiosyncracies.†

† One data point for the rate of change with experience is our finding (MacGregor *et al.*, in press) that outcome feedback on whether one had correctly identified the location of these items on up to three choices had no appreciable effect on either transparency or metatransparency.

Here, as in other tasks, transparency and metatransparency were related. With the poorest performance (40% correct first choices), calibration curves lay well below the identity line, indicating great overall overconfidence. These curves were flat as well, indicating complete insensitivity to when one is more or less likely to have identified the correct location. Across tasks, as transparency increased, so did confidence, however at a lesser rate, so that the disparity between confidence and success was reduced. With the greatest transparency (80% correct; Fig. 4), subjects showed moderate underconfidence. The overconfidence typically observed on first choices "left" relatively little probability over for the second and third choices, which typically showed mild underconfidence. The sequentialized curves for these choices, which looked at subjects' (implicit) conditional probability that "now they had the right answer", showed patterns like the first choices. Except where transparency was poorest, calibration curves had an upward slope, indicating that subjects had some sensitivity to the extent of their own knowledge. The slope was, however, fairly shallow, indicating moderate insensitivity as well.

Users who are this confident in their ability to use an interface should be relatively willing to give it a try. Users who are this overconfident should feel recurrent frustration, especially in cases where they are most confident of being correct. It has proven possible to improve calibration through training (Lichtenstein & Fischhoff, 1980; Murphy & Winkler, 1984). However, it requires making explicit probability assessments and receiving intensive, organized feedback. With a computerized system, providing such feedback would be straightforward for users willing to provide probabilities. Users unwilling to be bothered during actual searches might still be willing to endure the inconvenience. Barring that, the instructions to a system should give users some notion of how well they should expect to do (e.g. about 50% correct on first choices, with some limited ability to tell how likely individual items are to be found). How to convey such expectations is an open question.

Such performance statistics themselves may be useful for those managing a system, as well as for those using it. One could develop models predicting system usage patterns for individuals with varying degrees of persistence and tolerance for frustration. As long as searching is an uncertain process, both users and providers should be better off with accurate estimates of those uncertainties.

This research was supported by the National Science Foundation under Grant 1ST 8312482 to Perceptronics, Inc. Any options, 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. The support of the Foundation is gratefully acknowledged. We also thank Mark Layman for his continued diligent computing help, as well as Leisha Sanders, Geri Hanson, and Sonny Eberts for technical support.

References

Adelson, B. (1984). When novices surpass experts: the difficulty of a task can increase with experts. *Journal of Experimental Psychology: Learning, Memory and Cognition,* **10,** 483–495.

Baraslov, L. W. (1983). Ad hoc categories. *Memory and Cognition,* **11,** 211–227.

Bates, M. J. (1977). Factors affecting subject catalog search success. *Journal of the American Society of Information Science,* **28,** 161–169.

Beyth-Marom, R. (1982). How probable is probable? Numerical translation of verbal probability expressions. *Journal of Forecasting,* **1,** 257–269.

Blair, D. C. (1980). Searching biases in large interactive document retrieval systems. *Journal of the American Society for Information Science,* **31,** 271–277.

Bookstein, A. (1985). Probability and fuzzy-set applications to information retrieval. *Annual Review of Information Science and Technology,* **20,** 117–151.

Broadbent, D. E., Fitzgerald, P. & Broadbent, M. H. P. (1986). Implicit and explicit knowledge in the control of complex systems. *British Journal of Psychology,* **77,** 33–50.

Chi, M. T., Feltovich, P. J. & Glaser, R. (1981). Categorization in representation of physics problems by experts and novices. *Cognitive Science,* **5,** 121–152.

Cooper, W. S. (1978). Indexing documents by Gedanken experiments. *Journal of the American Society for Information Science,* **29,** 107–119.

Dumais, S. T. & Landauer, T. K. (1984). Describing categories of objects for menu retrieval systems. *Behavioral Research Methods, Instruments, and Computers,* **16,** 242–248.

Ericsson, H. A. & Simon, H. A. (1984). *Verbal Protocols as Data.* Cambridge, Massachusetts: MIT Press.

Fidel, R. (1983). Factors affecting online bibiographic retrieval: a conceptual framework for research. *Journal of the American Society for Information Science,* **34,** 163–180.

Fischhoff, B. (1987). Judgment and decision making. In Sternberg, R. J. & Smith, E. E. Eds., *The Psychology of Human Thought.* New York: Cambridge University Press.

Fischhoff, B. & MacGregor, D. (1986). Calibrating databases. *Journal of the American Society for Information Science,* **37,** 222–233.

Fischhoff, B., Slovic, P. & Lichtenstein, S. (1980). Knowing what you want: measuring labile values. In Wallsten, T. Ed., *Cognitive Processes in Choice and Decision Behavior.* Hillsdale, New Jersey: Erlbaum.

Furnas, G. W., Landauer, T. K., Gomez, L. M. & Dumais, S. T. (1983). Statistical semantics: analysis of the potential performance of key-work information systems. *Bell System Technical Journal,* **62,** 1753–1806.

Hartigan, J. (1981). Cluster analysis of variables. In Dixon, W. J. *et al.* Eds, *BMDP statistical software: 1981 edition* (448–455). Berkeley: University of California Press.

Homa, D. (1984). On the nature of categories. *The Psychology of Learning and Motivation,* Vol. 18. New York: Academic Press.

Katz, R. W., Murphy, A. H. & Winkler, R. L. (1982). Assessing the value of frost forecasts to orchardists: a dynamic decision-making approach. *Journal of Applied Meteorology,* **21,** 518–531.

Kiel, F. C. (1981). Constraints on knowledge and cognitive development. *Psychology Review,* **88,** 197–227.

Kiger, J. I. (1984). The depth/breadth tradeoff in the design of menu-driven user interfaces. *International Journal of Man–Machine Studies,* **20,** 201–213.

Koriat, A., Lichtenstein, S. & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory,* **6,** 107–118.

Krzysztofowicz, R. (1983). Why should a forecaster and a decision maker use Bayes Theorem. *Water Resources Research,* **19,** 327–336.

Landauer, T. K. & Nachbar, D. W. (1986). *Test of a model of menu-traversal time.* Murray Hill, New Jersey: Bell Communications Memorandum.

Lee, E., Whalen, T., McEwen, J. & Latremouille, S. (1984). Optimizing the design of menu pages for information retrieval. *Ergonomics,* **27.**

Lichtenstein, S. & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance,* **26,** 149–171.

Lichtenstein, S., Fischhoff, B. & Phillips, L. D. (1982). Calibration of probabilities: state of the art to 1980. In Kahneman, D., Slovic, P. & Tversky, A., Eds, *Judgement under Uncertainty: Heuristics and Biases.* New York: Cambridge University Press.

MacGregor, D., Fischhoff, B. & Blackshaw, L. (1987). Search success and expectations with a computer interface. *Information Processing and Management.* In press.

March, J. G. (1978). Bounded rationality, ambiguity, and the engineering of choice. *Bell Journal of Economics,* **9,** 587–608.

McDonald, J. E., Stone, J. D., Liebelt, L. S. & Karat, J. (1982). Evaluating a method for structuring the user–system interface. *Human Factors Society Proceedings*, **26**, 551–555.

Murphy, A. H. (1972). Scalar and vector partitions of the probability score. *Journal of Applied Meteorology*, **11**, 273–282.

Murphy, A. H. & Winkler, R. W. (1984). Probability of precipitation forecasts. *Journal of the American Statistical Association*, **79**, 391–400.

Murphy, G. L. & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychology Review*, **92**, 289–316.

Murphy, G. L. & Wright, J. C. (1984). Changes in conceptual structure with expertise: differences between real-world experts and novices. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **10**, 144–155.

Nakamura, G. V. (1985). Knowledge-based classification of ill-defined categories. *Memory and Cognition*, **13**, 377–380.

Nakamura, K., Sage, A. P. & Iwai, S. (1983). An intelligent database interface using psychological similarity between data. *IEEE Transactions on Systems, Man, and Cybernetics*, **SMC-13**, 558–561.

Nisbett, R. E. & Wilson, T. D. (1977). Telling more than we can know: verbal reports on mental processes. *Psychological Review*, **84**, 231–259.

Pejterson, A. M. (1980). Design of a classification system for fiction based on analysis of actual user–librarian communication and use of the scheme for control of librarians' search strategies. In Harbo, O. & Kajberg, L. Eds, *Theory and Application of Information Research*. London: Mansell.

Pitz, C. G. & Sachs, N. J. (1984). Judgment and decision: theory and application. *Annual Review of Psychology*, **35**, 139–163.

Raiffa, H. (1968). *Decision analysis*. Reading, Massachusetts: Addison–Wesley.

Roth, D. L. (1985). The role of subject expertise in searching the chemical literature and pitfalls that await the inexperienced searcher. *Database*, **8**, 43–46.

Savage, R. E., & Habinek, J. K. (1984). A multilevel user interface: design and evaluation through simulation. In Thomas, J. C. & Schneider, M. L. Eds, *Human Factors in Computer Systems*. Norwood, New Jersey: Ablex.

Slovic, P. & Fischhoff, B. (1977). On the psychology of experimental surprises. *Journal of Experimental Psychology: Human Perception and Performance*, **3**, 544–551.

Snowberry, K., Parkinson, S. & Sisson, N. (1983a). Computer display menus. *Ergonomics*, **26**, 609–712.

Snowberry, K., Parkinson, S. & Sisson, N. (1983b). Effects of help fields on hierarchical menu search. *Ergonomics*, **26**, 552–556.

Snowberry, K., Parkinson, S. & Sisson, N. (1985). The effects of help fields on navigating through hierarchical menu structures. *International Journal of Man–Machine Studies*, **22**, 479–491.

Snyder, K. M., Happ, A. J., Malcus, L., Paap, K. R. & Lewis, J. R. (1985). Using cognitive models to create menus. *Human Factors Society Proceedings*, **29**, 655–658.

Tulving, E. & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior*, **5**, 381–391.

U.S. Department of Commerce. (1983). *Statistical Abstract of the United States*. Washington, DC: Author.

von Winterfeldt, D. & Edwards, W. (1986). *Decision analysis and behavioral research*. New York: Cambridge University Press.

Wallsten, T. & Budescu, D. (1983). Encoding subjective probabilities: a psychological and psychometric review. *Management Science*, **29**, 151–173.

Watson, S. R. & Buede, D. M. (1987). *Decision synthesis*. Cambridge, England: Cambridge University Press. In press.

Witten, I. H., Cleary, J. G. & Greenberg, S. (1984). On frequency-based menu-splitting algorithms. *International Journal of Man–Machine Systems*, **21**, 135–140.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, **8**, 338–353.