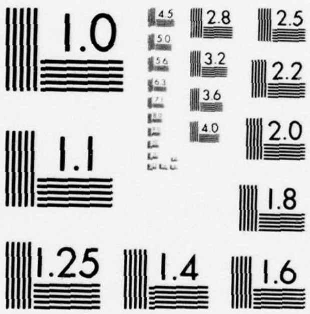
AD-A072 657 OKLAHOMA UNIV NORMAN DECISION PROCESSES LAB F/G 5/10
A MEMORY RETRIEVAL AID FOR HYPOTHESIS GENERATION. (U)
JUL 79 C GETTYS, T MEHLE, S BACA, S FISHER N00014-77-C-0615
UNCLASSIFIED TR-27-7-79 NL

| OF |
AD
A072657

END
DATE
FILMED
9-79
DDC


1.0
4.5
5.0
5.6
6.3
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.1
1.8
1.25
1.4
1.6

12

A Memory Retrieval Aid for Hypothesis Generation

Charles Gettys, Tom Mehle, Suzanne Baca

Stanley Fisher and Carol Manning

TR 27-7-79 July 1979

DECISION PROCESSES LABORATORY
UNIVERSITY OF OKLAHOMA

OFFICE OF NAVAL RESEARCH
CONTRACT NUMBER N00014-77-C-0615
WORK UNIT NUMBER NR197-040

12

A Memory Retrieval Aid for Hypothesis Generation

Charles Gettys, Tom Mehle, Suzanne Baca

Stanley Fisher and Carol Manning

TR 27-7-79 July 1979

prepared for

Office of Naval Research
Engineering Psychology Programs
Contract Number N00014-77-C-0615
Work Unit Number NR 197-040

Decision Processes Laboratory
Department of Psychology
University of Oklahoma
Norman, Oklahoma 73019

Approved for public release; distribution unlimited.

DISCLAIMER NOTICE

THIS DOCUMENT IS BEST QUALITY PRACTICABLE. THE COPY FURNISHED TO DDC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER (14) TR-27-7-79	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) (6) A MEMORY RETRIEVAL AID FOR HYPOTHESIS GENERATION.		5. TYPE OF REPORT & PERIOD COVERED (9) Rept. for Aug 1978-Aug 1979
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) (10) Charles Gettys, Tom Mehle, Suzanne Baca, Stanley Fisher and Carol Manning		8. CONTRACT OR GRANT NUMBER(s) (15) N00014-77-C-0615
9. PERFORMING ORGANIZATION NAME AND ADDRESS Decision Processes Laboratory Department of Psychology University of Oklahoma Norman, Oklahoma 73019 (12) 35p.		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR 197-040
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Engineering Psychology Programs-Code 455 800 N. Quincy Arlington, Virginia 22217		12. REPORT DATE (11) 27 Jul 1979
		13. NUMBER OF PAGES 27
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

1) decision theory
2) hypothesis generation
3) decision aiding
4) artificial memory
5) Expert

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Hypothesis generation consists of retrieving explanations for data from memory, and assessing these explanations for plausibility. Previous research has established that human hypothesis generation performance is deficient in both hypothesis retrieval and assessment. This study investigates an aid for the

over

410918

hypothesis retrieval process which is based on a model for hypothesis retrieval developed by Gettys, Fisher, and Mehle (1978). A computer simulates the human hypothesis retrieval process by searching an enriched associative memory which contains the associations of a number of individuals in the form a lists of hypotheses for each datum. When the data of a decision problem become known, the appropriate lists are searched by the computer. Hypotheses that are common to most or all of the lists are suggested to the user, who assesses them for plausibility. An experiment was performed to determine the utility of the aid for both expert and non-expert users. The aid produced a substantial gain in performance for both groups of users, suggesting that further development of the aid would be worthwhile in decision situations which are repeated often enough to warrant the creation of an enhanced artificial memory. Also discussed are several techniques for implementing the aid, and determining the maximum gain in performance that the aid can produce.

A memory retrieval aid to enhance hypothesis generation performance

The structuring of a decision problem is a vital precursor to the actual decision, since model specifications are used in all further analyses of the decision problem. If the decision maker fails to consider relevant hypotheses or acts in the decision model then the entire decision process can go awry because the model employed is incomplete or faulty.

Recently we have become convinced that the process of hypothesis generation, one of the vital constituents of problem structuring, is quite inefficient in non-routine tasks involving many possible hypotheses. For example, Gettys, Fisher and Mehle (note 1) found that hypothesis generation performance is quite impoverished; only about 50% of their subjects were able to generate two of the three relevant hypotheses in a hypothesis generation task. Curiously, while subjects were unable to retrieve complete hypothesis sets from memory, their assessment of the completeness of these sets was quite optimistic (Gettys, Mehle, and Fisher, submitted). Evidently the subjects believed that the hypothesis sets are more complete than they actually are because hypotheses that have not been generated are relatively unavailable in memory (Tversky and Kahneman, 1974). The inability to generate all relevant hypothesies coupled with the belief that more of the relevant hypotheses have been generated than is actually the case makes the subjects particularly vulnerable. Unaware of their deficiencies, they are, in effect, "fat and happy".

Since subjects fail to retrieve enough relevant hypotheses from memory, and are often unaware of their failure, it would be profitable to develop an aid for hypothesis generation. The hypothesis generation aid we propose is identical in logical structure to the hypotheses retrieval model developed by Gettys, Fisher and Mehle (1978) to describe the retrieval of hypotheses from human memory. However, it differs in that its associative memory is enhanced because it combines, or pools, the associations of a number of individuals by using a computer. Thus the aid is able to retrieve additional relevant hypotheses that were not retrieved by the user because the aid, in effect, searches the memory of many individuals, counteracting the inefficiencies in the memory of the individual user. The supplemental hypotheses provided by the aid should provide a larger, improved hypothesis set.

This aid is best suited for repetitive decision situations, or other situations where it is deemed worthwhile to go to the effort of constucting an artificial memory in advance. Some decision situations occur frequently such as automotive and electronic trouble shooting, or medical diagnosis; other situations may never have occurred, but are objects of much advance planning, and thought. These latter situations, such as planning for a possible melt-down of a nuclear reactor, have possible repercussions that are so profound that advance planning is conducted. In either of these types of situations the effort of constructing an artificial memory to aid hypothesis generation may be warranted. While the gain from each decision in a repetitive decision situation may be relatively small, there are many of these decisions that could be aided. In the case of the latter type of decision whose repercussions may be profound, the effort to construct an aid to hypothesis generation may be worthwhile, even if it is never used.

The aid is implemented by constructing an enhanced artificial memory in advance. While there are other techniques for constructing the artificial memory than the pooling process mentioned here, their discussion is deferred to a later point in the paper. The pooling process involves asking a number of individuals to search their memories for hypotheses that are associated with a datum. From these varied associations, a list that is rich in hypotheses is constructed by pooling the hypotheses of the contributors to the list. This process is repeated for each datum that is anticipated in the environment. The result is an enhanced associative memory where the lists for each datum consist of the associations between that datum and a number of hypotheses.

Once the artificial memory is stored in the computer it is ready for use. After the arrival of a datum, or a collection of data, a search is made of the list of possible hypotheses for each new datum, and hypotheses that appear on many of the lists are noted and added to a list of computer-generated potential hypotheses. Then this list of hypotheses is compared to the list of hypotheses that the user has generated. Hypotheses that the computer retrieved from its memory that were not retrieved by the user are displayed to the user. Finally, these additional hypotheses are assessed for plausibility by the user and added to the current hypothesis set if the user finds them plausible.

This aid is similar in logical structure to early medical diagnosis aids. These aids were unsuccessful because they employed deterministic inference in a probabilistic task. However, such an aid is viable when the task is aiding memory retrieval, as no attempt is made by the aid to engage in probabilistic inference. Since hypotheses are retrieved for subsequent evaluation by the user, the difficulties that were encountered in medical diagnosis are avoided.

The primary empirical question to be addressed in this study is the assessment of the actual gain in hypothesis generation performance that the aid provides when hypotheses suggested by the aid are assessed by the user. One variable that should have an effect on the improvement produced by the aid is the expertise of the user. Non-expert users should show the biggest gain in performance, and expert users should show more modest gains. This manipulation also addresses the question of whether non-expert subjects can serve as "surrogates" for scarce expert subjects.

A second variable of interest is the number of data in the hypothesis generation tasks. The aid does not contain a unique list of hypotheses for every possible combination of data that could occur. If, for example, there were 100 data that were possible, then the number of lists that would be neccessary would be two to the hundredth power. Instead, only 100 lists of hypotheses would be created. Hypotheses that are appropiate for multiple data are found by searching for hypotheses that are common the lists for the data of that problem. The advantage of this latter procedure is a tremendous reduction in the effort to construct the artificial memory of the aid, but the rule for finding hypotheses suggested by multiple data may be inefficient. Consequently, we decided to study the aid on the single, and in a multiple-data case.

Accordingly, the design of the aid study incorporates a between-subjects variable of expertise of the user, and within-subjects variables of number of data, and whether or not a particular hypothesis generation problem is aided. An additional non-aided control condition is incorporated in the design to assess irrelevant differences between the non-expert and the expert groups. This condition employs a hypothesis generation task in a domain where experts and nonexperts have comparable amounts of background and knowledge in order to assess differences in performance of the two groups due to nuisance variables unrelated to expertise, such as intelligence and motivation.

Method

Method and procedure for the aiding experiment.

Hypothesis generation tasks. The hypothesis generation tasks chosen for the aid experiment are the "Majors" task and the "Animals" task (Gettys, Fisher, and Mehle, 1978). In the Majors task the subject is given several courses that an University of Oklahoma (OU) student has taken, and is asked to generate a list of plausible majors for this student. The Animals task is similar except that the subject is asked to generate a list of plausible animals from several animal characteristics.

The Majors task was chosen because we have access to the veridical posterior probabilities of various majors given the classes that a OU student has taken. The presence of the veridical probabilites makes the evaluation of the aid possible as they provide the information necessary to create objective indices of performance. The veridical probabilities were not used in the construction, or the operation of the aid. The veridical probabilities are population values; 166,858 enrollment records at the University of Oklahoma covering a four-year period were tabulated to obtain these values.

Problems in the Majors task had either 1 or 3 data, and were aided on 50% of the trials. The Animals task where possible animal hypotheses were generated from animal characteristics served as a control task, and was not aided.

An example of a Majors problem is now described. This is a three data problem; the three data are: 1) Chemistry 1314-General Chemistry, 2) Chemistry 3053-Organic Chemistry, and 3) Mathematics 1513-College Algebra. The list of hypotheses generated by a randomly chosen expert subject included the following majors: Chemistry, Engineering, Pharmacy, and Physics. There were in fact 12 majors which had non-negligible probabilities for this problem. These majors and their veridical percentages are: Business (4.1%), Chemistry (6.2%), Education (2.1%), Engineering (4.1%), Laboratory Technology (9.3%), Liberal Studies (2.1%), Medical technology (4.1%), Microbiology (6.2%), Pharmacy (21.6%), Psychology (9.2%), University College (11.3%) and Zoology (10.3%). University College is an "undeclared" major for beginning students. The remaining majors all had percentages of less than 2%. The sum of the percentages of majors with percentages greater than 2% is 90.6%. If a subject achieved such a sum, it would have been optimal performance in this task as the subjects were instructed to respond with all hypotheses greater than 2%. As can be calculated from the percentages of the hypotheses that the subject generated, the subject's performance was 31.9%, and the subject failed to generate several important hypotheses such as Laboratory Technology, Microbiology, Psychology, University College, and Zoology. This performance is typical of the average subject for this problem. The number 31.9% has a direct theoretical interpretation; it is the probability that the subject's list of majors, or hypothesis set, contains the correct hypothesis. Thus, for this problem, this subject would have failed to consider the correct hypothesis with a probability of 68.1%.

An example of an animal problem was to name animals that have antlers. The responses of the same subject to this problem were: deer, moose, antelope, and reindeer. The reader is invited to generate additional hypotheses for this problem.

Apparatus. The experiment was controlled by a Compucolor computer which had a color graphics capability, and was programmed in extended Basic.

Subjects. The subjects in this experiment were drawn from two populations. Non-expert subjects were University of Oklahoma students who were required to have at least 60 hours of course work at the University and typing skills. These students were recruited from classes and newspaper advertisements, and were paid $5.00 for their participation. There were 16 subjects in this group.

The expert subjects were University of Oklahoma Curriculum Advisors. Various Colleges and Departments of the University maintain advising offices and employ individuals with a job title of "curriculum advisor" who are expert on University, College, and Departmental requirements, and the course offerings of the University. This group of experts are professional student advisors who work with student schedules on a daily basis. There are about 30 such advisors, and we recruited 16 for this experiment. These subjects were paid a $10.00 "honorarium" for their participation in addition to their usual salary.

Instructions to subjects. The instructions to subjects were elaborate. First, written instructions explaining the experiment and the aid were presented on the computer. It was explained that the possibilities suggested by the aid were to be carefully assessed; that the subjects should use their best judgment in deciding whether or not to include the aid's suggestions on their list of hypotheses. A particlarly pertinent section of these instructions is reproduced below:

> We are investigating a computer aid for memory in this study. We have found that people sometimes fail to remember relevant information in certain situations. The computer aid acts a a prompt for memory. We are interested in in learning how useful the aid is in helping people search their memories.
>
> As we are interested in the extent to which the aid helps you search your memory, it is vitally important that you understand everything about the experiment. For this reason, we want you to ask questions whenever you need clarification. We will be happy to explain any aspect of the experiment to you.
>
> One of the first things that a doctor does before making a diagnosis is to make a mental list of the possible diseases that the patient might have based on the patient's symptoms. If this list does not include the disease that the patient has, the doctor's diagnosis is bound to be wrong. So coming up with a complete list of possibilities is very important and we are studing an aid that should help people create a more complete list.
>
> Instead of investigating medical diagnosis which requires special expertise, we have chosen similar problems which have the same characteristics. Some of these problems involve generating a list of possible majors for an unknown OU student on the basis of courses that this student has taken. For example, if you knew that the unknown student had taken 9 hours of Zoology, you would probably include Biological Science majors on you list such as Zoology and Botany. The student could also be a Psychology major who took these courses as part of a Pre-Med program, or even an Art major who is fascinated by Zoology. Art, of course, is not nearly as likely, but it is possible. Many other possible majors exist. Can you think of

> any? How likely are they?
>
> To cut the task of generating this list down to manageable size, you need not add possible, but highly unlikely majors (such as Art) to the list you will generate. If the chances of a particular major are less than 2% you should not add it to your list, but all majors which are more likely than 2% should be included on your list.
>
> One way of making this clearer is to imagine that all the non-transfer students who had taken these Zoology courses for the last several years were assembled in a large auditorium. Students are seated by majors under large signs giving their majors. Some majors will have many students, others will have only a few, or none. Your task will be to list all the majors which include more than 2% of the total number of students in this room. If this isn't perfectly clear, now would be a good time to discuss this with the experimenter.
>
> Other problems will involve making lists of animals from their characteristics. Use the 2% rule here also. If the animal having the specified characteristics is quite rare, you need not add it to your list.

Following the written instructions the subjects worked three practice problems using the same procedure as in the main experiment. There was one of each type of problem used in the main experiment; an unaided "Majors" problem, an aided "Majors" problem, and an "Animal" problem were included in the practice set so that the subjects would have experience with all of the types of problems to be encountered in the main session.

Design of the study. The design is a 2 by 2 by 2 mixed factorial where expertise is a between-groups variable, and number of data and aiding are within-groups variables. There were four one-datum problems, and four three-data problems. Each subject was aided on 50% of the problems counterbalanced across number of data so that each problem was aided equally often for each group. The two "Animal" problems were included with the eight

"Majors" problems, so each subject worked a series of ten problems in the main part of the experiment. The order of the ten problems was randomized for each subject.

Procedure. Following the instructions the subjects worked ten problems at their own pace. The experimental session typically lasted between one and one and a half hours. The data of the problem was displayed on the video screen of the computer. The subject's answers were typed into the computer keyboard. For the "Majors" problems a spelling check was made by the computer. When the subject entered a major it was compared to a list of the 68 possible majors. If an exact letter-for-letter match was found, the major was added to the subject's list of plausible majors. If this match failed, then the computer executed a routine where it attempted to identify the entry. If a major closely approximating the subject's entry was found, the computer asked for confirmation that this major was in fact the one that the subject intended. The subject continued to enter majors until the subject believed that all the majors which included more than 2% of the students who had taken the specified courses had been identified. Then the subject entered "DONE" into the computer. If the problem was unaided for that subject, the program began the next problem. If that problem was aided then the aiding display was generated. Subjects were unaware that a particular problem was aided until this point to control the possibliity that they might rely on the aid if they knew that a particular problem was aided in advance.

If a problem was aided, the list of "Majors" that the subject generated, the data of that problem, and the "Majors" suggested by the aid were displayed. Any majors that the subject generated were removed from the aid before it was displayed to the subject. The subject had been told during the instructions and the practice problems that this aiding list was generated by the computer on the basis of other people's responses and that it was to be searched for majors that were greater than 2% to be added to the list. The majors suggested by the aid were numbered. Subjects indicated which majors were to be added to their list by typing the number of the major. They could adopt several majors by entering the numbers associated with the majors, separated by plus signs. Any majors so adopted were transferred from the aiding list to the list of responses adopted by the subject on the display. This process was repeated until the subject entered the number indicating that all the desired transfers to their list had been made. When all 10 problems had been worked, the session ended and the subjects completed a short questionaire concerning the aid.

All hypotheses that the subject generated were recorded, as were the hypotheses suggested by the aid and adopted by the subject. The basic index of performance for aided and unaided responses is the posterior probability associated with each of these majors, as was indicated previously in the example problem.

Creation of the aid.

The aid is created by generating a list of possible hypotheses for each datum, and storing these lists in a computer for future access. In principle, there are many ways that these lists can be generated. In situations which are relatively well understood, such as automotive and electronic trouble shooting, authoritative sources of information can be consulted to generate these lists. Alternatively, historical records can be consulted to provide this information.

In other situations, where such authoritative sources of information are unavailable, it is possible to generate these lists by pooling hypotheses generated by knowledgeable individuals. We chose the latter technique for this study because we believe it is likely to be the easiest to implement in an applied setting. The pooling process is essential because any individual may have lapses in hypothesis generation and generate an incomplete list of hypotheses. If, however, several individuals are used, then relevant hypotheses which one individual fails to retrieve are often retrieved by another individual because of differences in their experiences and because lapses in retrieval from memory by one individual may be compensated by successes of other individuals.

Once the lists of hypotheses for each datum are created, and stored in the computer, the aid is ready for use. As data arrive these lists are accessed and searched for hypotheses that are common to several of the lists. Hypotheses that occur more frequently than a threshold value are presented to the decision maker for assessment, and are adopted if they are sufficiently plausible.

Selection of the contributors to aid lists. The expertise and number of contributors to the lists are both important variables which affect the quality of the lists. First, it is desirable that the lists be generated by experts as the hypothesis set of an expert should be larger than that of a non-expert due to the expert's greater knowledge and experience. Second, as even experts have lapses in their memory, several experts should contribute to each list. An increase in the number of contributors should enhance the list to the extent that their experience and knowledge differ. The choice of the number of experts to consult is primarily governed by their expertise, and the importance of the problem. An increase in the number of contributors to the list should partially compensate for a lack of a high degree of expertise on the part of the contributors. On the other hand, if the contributors are experts of the highest quality, then a smaller number of contributors should be sufficient.

Contributors used in this experiment. In the present study, non-experts were used to generate the lists since there were not enough experts to do this task and also participate in the experiment. The non-experts used were students in Experimental Psychology at the Universty of Oklahoma. This course is typically taken by upperclassmen due to its prerequisites. Eighteen students generated a list in response to each datum, and their lists were pooled. A hypothesis was included on the list if it was given by any of the 18 contributors to the list.

The delta P metric. Once the aid has been generated it is important to assess its adequacy. We have developed a technique for studying the adequacy of the lists used in the hypothesis generation aid. This technique has a variety of potential applications. First, it is possible to characterize the maximum gain that could be expected using the aid if the hypothesis assessment of the user were perfect. Second, the performance of an unaided user can be roughly estimated using the memory retrieval model of Gettys, Fisher, and Mehle (1978). Third, it is possible to decide how many contributors to each list should be used. These ideas may have considerable practical importance if this aid proves to be successful in certain situations. Unfortunately these techniques require veridical posterior probabilities which may be difficult, or impossible to obtain in some situations.

This metric is named Delta P, and it is based on the following ideas. Each contributor to the lists is given a datum and is asked to generate an list of possible hypotheses. Each of these lists can be characterized by the probability, P, that the correct hypothesis is contained in the list. P is calculated by summing the posterior probabilities for each hypothesis on the list. This sum the probability that the list will contain the correct hypothesis. The value of P is less than 1.0 to the extent that the list is not exhaustive, or lacks relevant hypotheses. Various contributors to the list do not generate exactly the same hypotheses. Pooled lists should contain more plausible hypotheses than any individual's list and will have a greater value of P. This technique can readily be generalized to N individuals. The difference in P between an individual list and the pooled list resulting from N individuals is termed Delta P, which is the gain in P resulting from the pooling process. These elementary considerations yield several interesting results.

Estimating unaided performance. First, if P is calculated for each of the contributors, the average value of P is an estimate of the performance of an unaided user for a particular datum. The memory tagging model of Gettys, Fisher and Mehle can be used to make a rough estimate of unaided performance for multi-data problems by Monte Carlo techniques.

Estimating the maximum possible gain from the aid. The delta P value resulting from pooling the hypothesis lists of all contributors is an estimate of the maximum possible gain for users of the same level of expertise as the contributors. This gain may not be realized in practice if the aided user does not exploit the full potential of the aid, but if the aid shows a small value of Delta P in a given situation then the aid will be of little, or no use in that situation.

Estimating the number of contributors to the aid. by varying the number of contributors, N, from one to its maximum value, and calculating P for each possible value of N, a negatively-accelerated curve in P is traced out. This analysis can be performed by Monte Carlo techniques where the lists of the various contributors are randomly chosen, or by an exhaustive analysis where all possible combinations of contributors are assessed.

Setting the threshold values of the aid. There are two threshold values that impact on the performance of the aid. By adjusting these values the hypothesis set that the aid produces can be varied at will. First, the criterion for including the hypothesis on the list can be varied. In the present study this criterion was set so that if any of the 18 contributors to the aiding lists suggested a major for the aid it was included in the lists that the computer searched. Such a criterion admits many majors to the list that are quite unlikely, but maximizes the number of relevant hypotheses included on the list. We chose this criterion because it is possible to calculate what aided performance would have been if a more stringent criterion had been employed, and so are able to examine the performance of the aid with various criteria.

The second criterion that must be determined is the rule to be used by the computer when searching the list for hypotheses that are common to several of the lists. For the one-datum problems the choice is forced, as only one list is searched. For the three-data problems we picked a criterion that the major must appear on at least two of the three lists before it is suggested to the subject. We chose a value of two because previous research (Gettys, Fisher, and Mehle, 1978) suggests that subjects retrieve a hypothesis from memory when it is tagged for two out of three data.

By adjusting these two criteria it is possible to increase the number of relevant hypotheses that the aid retreives but at a cost of increasing the number of unlikely hypotheses that are retrieved. Each time the aid is implemented these decisions will have to be made. In effect, the mesh size of the net must be set to determine the minimum size of fish that will be caught.

Using these criteria the aid suggested 32 majors for the example problem discussed previously. (This happens to be the maximum number of majors suggested for any problem.) Of the 12 majors that were more likely than 2%, 9 were included on the aid list. The aid did not sucessfully retrieve Laboratory Technology (9.3%), Liberal Studies (2.1%), and University College (11.3), but it did retrieve hypotheses whose sum was 67.9%. Had the criterion for the inclusion of a major on the aid lists been at least two out of the 18 contributors to the aid, then the aid's performance would have been 63.9%, and the aid would have only suggested 9 hypotheses less than 2% rather than 23 as was actually the case.

Results and discussion

Performance on the control task.

As the experiment employed two distinct populations of subjects, we included a control condition to detect possible differences between our non-expert and expert subjects on a topic that was irrelevant to the expert's specialty. The task chosen was the "Animal" task which we felt tapped items of common knowlege which both groups should have. Thus differences in performance should be due to hypothesis generation ability.

The number of animal responses that were consistent with the data were tabulated for both groups. The mean number of appropriate responses for the non-experts was 5.16, while the experts achieved a mean of only 3.19 correct responses (F=6.90; df=1,30; p<.05).

It might be tempting to explain these results using some of the common prejudices connected with experts, but we believe that another explanation is more likely. We noticed that the expert's attitude toward these "Animal" problems was sometimes one of indifference. The experts were recruited with the idea that their expertise would contribute to the evaluation of the aid. We did not mention in our recruitment literature that another group of non-expert subjects would be a part of the experiment, or the purpose of the animal problems, nor did we volunteer this information unless asked. For these reasons, some of the experts probably regarded these problems as irrelevant trivia. The non-expert subjects were mostly advanced psychology majors who perhaps were hardened to the practices of experimental psychologists. While this comparision is perhaps flawed for the above reasons, the results suggest that the expert subjects are no better than the non-experts in general hypothesis generation ability, a result which will aid the interpretation of other results.

Unaided performance.

For problems where the expertise of the experts was relevant, one might expect that the experts would show superior performance to non-experts, and in fact this was the case. We summed the probabilities of all the "Majors" hypotheses generated by the subjects without the aid that were greater than 2% for both groups using the technique illustrated previously. The mean performance of the non-experts, expressed as a percentage, was 47.7%, while the mean for the experts was 50.6%. The difference in performance is 2.9% and this difference is statisically reliable (F=4.5; df=1,28; p<.05). This difference, while in the expected direction, is surprisingly low. We expected a larger difference.

This small difference between experts and non-experts raises some interesting questions about the role expertise plays in the hypothesis generation process. Our earlier reports of deficiences in hypothesis generation using non-expert subjects have been questioned due to our subject's lack of expertise. Our results for the expert and non-expert subjects suggest that subject-matter expertise is not a potent variable in hypothesis generation, and that non-expert subjects are satisfactory surrogates for expert subjects. It does not follow from these results that expertise is largely irrelevant in hypothesis generation. It may be that expertise in the subject matter of the task must also be coupled with daily performance of the task for the true advantage of expertise to become apparent. In any event, these results do indicate that the surprising deficiencies in hypothesis generation performance which we have observed previously, and replicate here, are not due to lack of subject matter expertise.

Perhaps the most interesting aspect of the unaided performance of both groups of subjects is its implication for practical decision making. Subjects, either expert or non-expert, are not capable of generating an adequate hypothesis set in this task. While the generality of this effect has not been completely established, this is cause for alarm. The percentages reported previously are not arbitrary scores, they reflect the probability that the subject's list will contain the true hypothesis. In other words, if a subject earns a score of of 50.6% this means that on the average the true hypothesis will not be considered on about 50% of the occasions when the subject generates hypotheses. We wonder whether decision analytic models are robust enough to tolerate such a high error rate?

There are several possible explanations for inadequate hypothesis generation. One is that the "majors" task is incredibly difficult. There is an element of truth in this argument; a modern university is in fact quite complex, and no single individual can be aware of the layers of College and Departmental requirements, recommendations, and student preferences. We believe, however, that this is a characteristic of many real-world situations that are not completely understood. In medical diagnosis, for example, a physician's knowlege is comparable to that of our experts in some sense. Both groups are dealing with an imperfectly understood enviornment. Both groups are capable of dealing with routine problems where standard procedures exist. These routine problems are not usually the subject of decision analysis; decision analysis is usually employed when our understanding of the problem is imperfect, and it is in these very problems that one would expect to find deficiencies in hypothesis generation. Furthermore, our earlier results were not obtained using difficult problems, instead we used problems which should have been easy for our subjects.

An alternative explanation to task difficulty may actually account for a larger percentage of the subject's deficiencies in hypothesis generation. In a study examining the process by which hypotheses are retrieved from memory, Gettys, Fisher, and Mehle (1978) found that retrieval processes are suprisingly inefficient. This result suggests that the memory search process by which subjects retrieve hypotheses misses many hypotheses that are in memory, but cannot be accessed from the data. If this is the case, then part of the deficiencies in hypothesis generation are due to failure to retrieve information that the decision maker possesses. This situation is, of course, exactly the situation with which the aid is designed to deal, and we would predict from this notion that the aid should prove effective in prompting the subject's memory.

Aided Performance

The performance of subjects on aided prblems was also calculated using the same procedures described previously, except that for aided problems the final list that the subject generated after using the aid was scored. Mean aided expert performance was 60.3% and non-expert performance was 57%. The difference between groups was not reliable (F=1.69; df=1,28; p>.2), but both groups were aided significantly by the aid. The experts showed an improvement of 13.3%, while the non-experts showed an improvement of 18.5% over their unaided performance. The difference in the improvement in performance was reliable (F=4.16; df=1,28; p< .05). There was also a reliable effect due to the number of data. Performance on one-datum problems was 62.2%, while performance on three data problems was 55% (F=23.94; df=1,28; p<.01).

The aid does produce the expected gain in performance that is consistent with the notion that decision makers can recognize, but not always retrieve, relevant hypotheses. It is interesting to note that the initial difference between non-experts and experts is reduced by the aid. The aid enhanced the performance of the non-experts to a greater extent, as might be expected.

The decision as to whether the aid is worthwhile to implement will depend on the importance of the gain in performance that it produces. The consequences of the gain will depend in a complicated way on the decision model which is appropriate in a given situation, and as we did not embed our hypothesis generation problems in a decision situation, we cannot calculate a gain in potential payoff from using the aid, nor can we estimate the costs of implementing the aid in a given situation, except to say that it should be relatively inexpensive. The aid does seem to be promising enough to warrant further development in other situations to further study its utility.

The results for number of data were as predicted. We hypothesized that when hypotheses must be retrieved that are consistent with several data that both the memory retrieval aid and the subjects would have more difficulty. However, the effect of number of data interacted significantly with the problem (F=5.46; df= 1,28; p<.05) and we employed only four problems at each level of number of data. These two considerations suggest that this result should be interpreted with caution, it may be an effect due to the particular problems chosen for the experiment.

Potential performance of the aid

Is it possible that most of the hypotheses the subjects generated were anticipated by the aid? Considerable insight into what actually happened in the experiment can be gained by "turning the tables" on the subjects and the aid. Suppose that all of the suggestions of the aid were adopted without assessment, and the subjects were invited to "aid" the aid. The aid that the subjects would provide in this situation would be those hypotheses that they retrieved from their memories that had not been generated by the aid. To perform this analysis, it is first neccessary to calculate how the aid would have performed without the help of the subjects. The result of this calculation is that the aid alone performed with a score of 76%! The absolute size of this number is not the major reason why it is impressive, if more, or better, contributors had been used it would have been higher. (In fact, had the historical technique that we used to ascertain the veridical probabilities had been used to generate the lists, the aid would have performed perfectly, achieving a score of 88.9%. This percent is just the sum of the probabilities that are greater than 2%.) The interesting result is the relative comparision of the aid alone compared to either the aidad, or the unaided subjects, and compared to the subjects aiding the aid. This latter result is the simplest, so it will be discussed first.

Table 1

A Comparision Between Unaided And Aided Human Performance
And The Aid By Itself

	Human		Aid only using a criterion of:		
	Unaided	Aided	1 per 18	2 per 18	2 per 18
Percent	50.6%	67.5%	76.0%	67.1%	62.9%
# Hyp>2%	3.04	4.71	6.12	4.75	4.25
# Hyp<2%	3.75	7.53	15.12	7.75	5.00

When the subjects "aided the aid", the gain in performance was less than 1%. This means that the subjects rarely retrieved hypotheses that were not retrieved by the aid, which is a powerful testament to the efficency of the pooling process.

The aid also performed better than the unaided subject and the aided subject, but at the cost of adding unlikely hypotheses to the list of majors. However, as mentioned previously, the criteria used by the aid can be adjusted to reduce the number of "false alarms". These calculations were performed for the aid as the sole hypothesis generator for various aid criteria. We manipulated the criterion used to include a major on the lists in the aid computer memory, using a criterion of either at least 1, 2, or 3 contributors out of the 18 contributors as the rule for inclusion on the aid lists. The results of these calculations, and the number of hypotheses that the aid "adopted" that were greater, or less than 2% are shown in table 1 with the results of aided and non-aided human performance.

(insert table 1 about here)

As can be seen from an inspection of table 1, the aid alone with a criterion of 2 out of 18 subjects is clearly superior to the unaided subject. It is also most interesting that it performs as well as a aided human, achieving about the same percentage performance with roughly equivalent false alarms.

The conclusion is inescapable. In this situation, at least, the aid could completely replace the human decision maker with no loss in performance. We do not seriously advocate such an extreme recommendation at this time for reasons of user acceptance, but these results suggest that the subjects would have contributed little to the performance of the aid had the tables been reversed.

Summary

The hypothesis generation aid was shown to enhance the hypothesis generation performance of both expert and non-expert subjects to a noticeable degree. These results also demonstrate the potential of creating an artificial computer memory based on human judgment, which in this situation at least, can achieve, by itself, better performance than an unaided human hypothesis generator.

Aknowledgments

This research was supported by the Office of Naval Research, Engineering Psychology Programs.

Several people contributed to the success of this study. Chuck Rice of the University Computing Services did the preliminary data reduction for the calculations of the veridical probabilities.

Vesta Gettys critically read an early draft of this report, and made many valuable suggestions. She also cheerfully delayed her vacation in the interests of comprehensible prose.

The help of Provost J. R. Morris and Ms. Penny Long were invaluable in recruiting the expert subjects.

Jeff Casey performed several data analyses, and the office staff of the Department of Psychology cheerfully helped in the reproduction of this technical report.

References

Gettys, C. F., Fisher, S. D., and Mehle, T. Hypothesis generation and plausiblity assessment. _Decision Processes Laboratory_, Unversity of Oklahoma, Norman, Oklahoma TR 15-10-78, October 1978. AD A060786.

Tversky, A. and Kahneman, D. Judgment under uncertainty: Heuristics and biases. _Science_, 1974, _85_, 1124-1131.

OFFICE OF NAVAL RESEARCH, CODE 455
TECHNICAL REPORTS DISTRIBUTION LIST

CDR Paul R. Chatelier
Military Assistant for Training and Personnel Technology
Office of the Deputy Under Secretary of Defense
OUSDRE (E&LS)
Pentagon, Room 3D129
Washington, D.C. 20301

Director
Engineering Psychology Programs
Code 455
Office of Naval Research
800 North Quincy Street
Arlington, Virginia 22217 (5 cys)

Director
Analysis and Support Division
Code 230
Office of Naval Research
800 North Quincy Street
Arlington, Virginia 22217

Director
Naval Analysis Programs
Code 431
Office of Naval Research
800 North Quincy Street
Arlington, Virginia 22217

Director
Operations Research Programs
Code 434
Office of Naval Research
800 North Quincy Street
Arlington, Virginia 22217

Director
Statistics and Probability Program
Code 436
Office of Naval Research
800 North Quincy Street
Arlington, Virginia 22217

Director
Information Systems Program
Code 437
Office of Naval Research
800 North Quincy Street
Arlington, Virginia 22217

Special Assistant for Marine Corps Matters
Code 100M
Office of Naval Research
800 North Quincy Street
Arlington, Virginia 22217

Commanding Officer
ONR Branch Office
ATTN: Dr. J. Lester
Building 114, Section D
666 Summer Street
Boston, Massachusetts 02210

Commanding Officer
ONR Branch Office
ATTN: Dr. C. Davis
536 South Clark Street
Chicago, Illinois 60605

Commanding Officer
ONR Branch Office
ATTN: Dr. E. Gloye
1030 East Green Street
Pasadena, California 91106

Office of Naval Research
Scientific Liaison Group
American Embassy, Room A-407
APO San Francisco, California 96503

Director
Naval Research Laboratory
Technical Information Division
Code 2627
Washington, D.C. 20375 (6 cys)

ONR, Code 455, Technical Reports Distribution List

Dr. Robert G. Smith
Office of the Chief of Naval
Operations, OP987H
Personnel Logistics Plans
Washington, D.C. 20350

Naval Training Equipment Center
ATTN: Technical Library
Orlando, Florida 32813

Human Factors Department
Code N215
Naval Training Equipment Center
Orlando, Florida 32813

Dr. Alfred F. Smode
Training Analysis and Evaluation Group
Naval Training Equipment Center
Code N-00T
Orlando, Florida 32813

Dr. Gary Poock
Operations Research Department
Naval Postgraduate School
Monterey, California 93940

Dean of Research Administration
Naval Postgraduate School
Monterey, California 93940

Dr. A. L. Slafkosky
Scientific Advisor
Commandant of the Marine Corps
Code RD-1
Washington, D.C. 20380

Mr. Arnold Rubinstein
Naval Material Command
NAVMAT 98T24
Washington, D.C. 20360

Bureau of Naval Personnel
Special Assistant for Research
Liaison
PERS-OR
Washington, D.C. 20370

Navy Personnel Research and
Development Center
Manned Systems Design, Code 311
San Diego, California 92152

Navy Personnel Research and
Development Center
Code 305
San Diego, California 92152

Navy Personnel Research and
Development Center
Management Support Department
Code 210
San Diego, California 92152

CDR P. M. Curran
Human Factors Engineering Division
Naval Air Development Center
Warminster, Pennsylvania 18974

Mr. Ronald A. Erickson
Human Factors Branch
Code 3175
Naval Weapons Center
China Lake, California 93555

Dean of the Academic Departments
U.S. Naval Academy
Annapolis, Maryland 21402

Mr. J. Barber
HQS, Department of the Army
DAPE-PBR
Washington, D.C. 20546

ONR Code 455, Technical Reports Distribution List

Dr. Joseph Zeidner
Technical Director
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, Virginia 22333

Director, Organizations and
Systems Research Laboratory
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, Virginia 22333

Dr. Edgar M. Johnson
Organizations and Systems Research
Laboratory
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, Virginia 22333

Technical Director
U.S. Army Human Engineering Labs
Aberdeen Proving Ground, Maryland 21005

U.S. Air Force Office of Scientific
Research
Life Sciences Directorate, NL
Bolling Air Force Base
Washington, D.C. 20332

Dr. Donald A. Topmiller
Chief, Systems Engineering Branch
Human Engineering Division
USAF AMRL/HES
Wright-Patterson AFB, Ohio 45433

Air University Library
Maxwell Air Force Base, Alabama 36112

Defense Documentation Center
Cameron Station, Bldg. 5
Alexandria, Virginia 22314 (12 cys)

Dr. Stephen J. Andriole
Director, Cybernetics Technology Office
Defense Advanced Research Projects Agenc
1400 Wilson Blvd.
Arlington, Virginia 22209

Dr. Judith Daly
Cybernetics Technology Office
Defense Advanced Research Projects Agenc
1400 Wilson Blvd.
Arlington, Virginia 22209

Professor Douglas E. Hunter
Defense Intelligence School
Washington, D.C. 20374

Dr. Richard S. Heuer, Jr.
ORPA/AMERS
Washington, D.C. 20505

Dr. Alphonse Chapanis
Department of Psychology
The Johns Hopkins University
Charles and 34th Streets
Baltimore, Maryland 21218

Dr. Meredith P. Crawford
Department of Engineering Administration
George Washington University
Suite 805
2101 L. Street, N.W.
Washington, D.C. 20037

Dr. Ward Edwards
Director, Social Science Research
Institute
University of Southern California
Los Angeles, California 90007

Dr. Kenneth Hammond
Institute of Behavioral Science
University of Colorado
Room 201
Boulder, Colorado 80309

ONR Code 455, Technical Reports Distribution List

Dr. Ronald Howard
Department of Engineering-Economic
Systems
Stanford University
Stanford, California 94305

Dr. William Howell
Department of Psychology
Rice University
Houston, Texas 77001

Journal Supplement Abstract Service
American Psychological Association
1200 17th Street, N.W.
Washington, D.C. 20036 (3 cys)

Dr. Clinton Kelly
Decisions and Designs, Inc.
8400 Westpark Drive, Suite 600
P.O. Box 907
McLean, Virginia 22101

Dr. Robert R. Mackie
Human Factors Research, Inc.
5775 Dawson Avenue
Goleta, California 93017

Dr. Gary McClelland
Institute of Behavioral Sciences
University of Colorado
Boulder, Colorado 80309

Human Resources Research Office
300 N. Washington Street
Alexandria, Virginia 22314

Dr. Miley Merkhofer
Stanford Research Institute
Decision Analysis Group
Menlo Park, California 94025

Dr. Terence R. Mitchell
University of Washington
Seattle, Washington 98195

Dr. Jesse Orlansky
Institute for Defense Analyses
400 Army-Navy Drive
Arlington, Virginia 22202

Professor Judea Pearl
Engineering Systems Department
University of California-Los Angeles
405 Hilgard Avenue
Los Angeles, California 90024

Professor Howard Raiffa
Graduate School of Business
Administration
Harvard University
Soldiers Field Road
Boston, Massachusetts 02163

Dr. Paul Slovic
Decision Research
1201 Oak Street
Eugene, Oregon 97401

Dr. J. A. Swets
Bolt, Beranek & Newman, Inc.
50 Moulton Street
Cambridge, California 94305

Dr. W. S. Vaughan
Oceanautics, Inc.
422 6th Street
Annapolis, Maryland 21403

Dr. Gershon Weltman
Perceptronics, Inc.
6271 Variel Avenue
Woodland Hills, California 91364

Dr. David Dianich
Chairman, Dept. of Business and
Economics
Salisbury State College
Salisbury, Maryland 21801

ONR Code 455, Technical Reports Distribution List

North East London Polytechnic
The Charles Myers Library
Livingstone Road
Stratford
London E15 2LJ
UNITED KINGDOM

Professor Dr. Carl Graf Hoyos
Institute for Psychology
Technical University
8000 Munich
Arcisstr 21
FEDERAL REPUBLIC OF GERMANY

Director, Human Factors Wing
Defense & Civil Institute of
Environment Medicine
Post Office Box 2000
Downsville, Toronto, Ontario
CANADA

Dr. A. D. Baddeley
Director, Applied Psychology Unit
Medical Research Council
15 Chaucer Road
Cambridge, CB2 2EF
UNITED KINGDOM

Dr. J. Baal Schem
Acting Director
Tel - Aviv University
Interdisciplinary Center for Technological
Analysis and Forecasting
Ramat-Aviv, Tel-Aviv
Israel

Dr. Amos Tversky
Department of Psychology
Stanford University
Stanford, California 94305