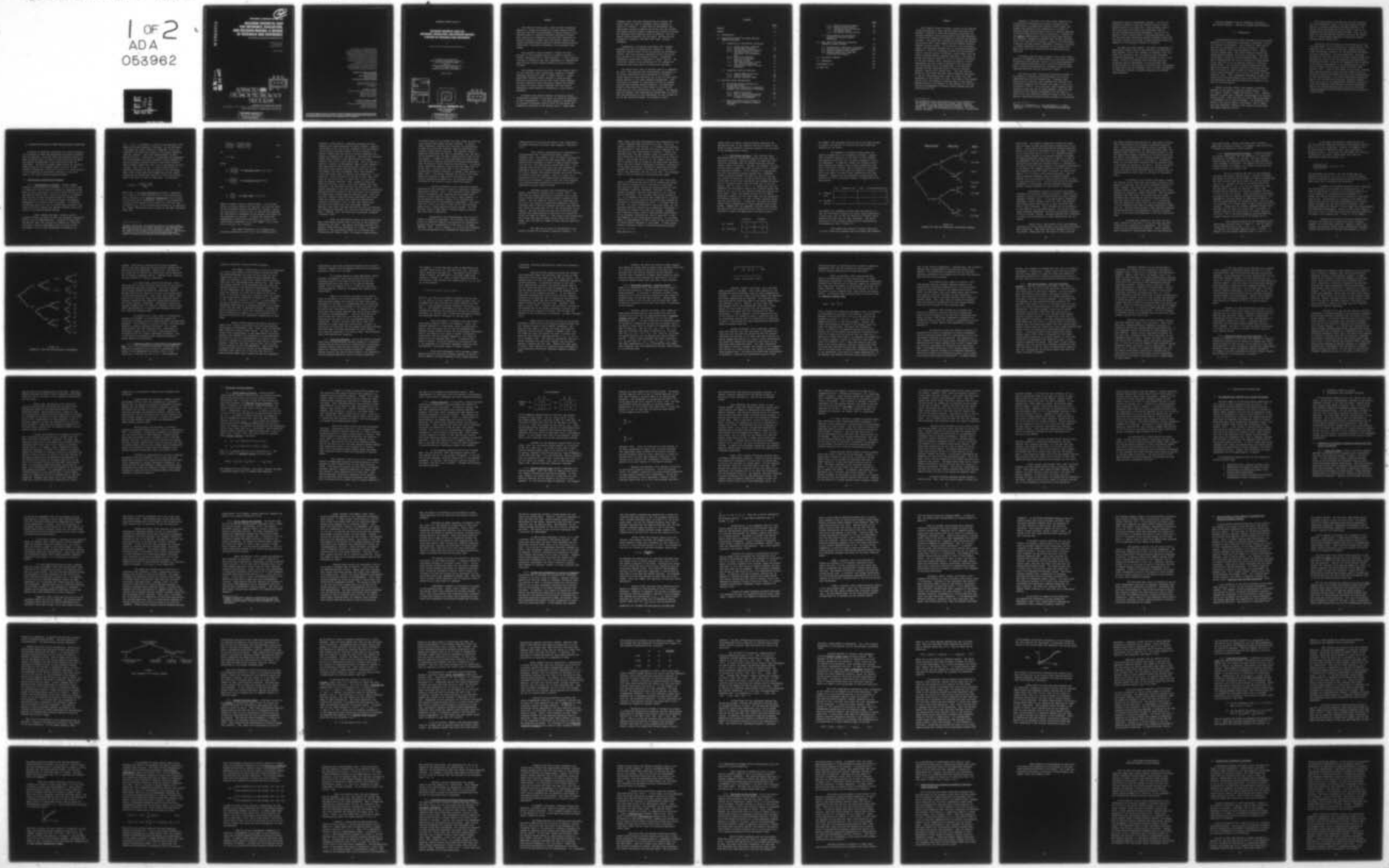
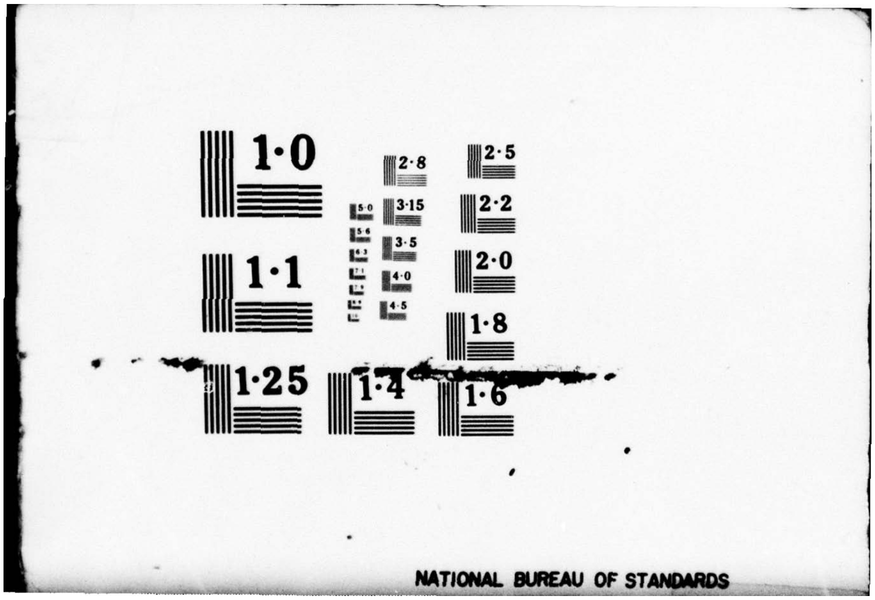
AD-A053 962 DECISIONS AND DESIGNS INC MCLEAN VA F/G 5/10
DECISION THEORETIC AIDS FOR INFERENCE, EVALUATIONS, AND DECISIO--ETC(U)
FEB 78 G W FISCHER, C W KELLY, W EDWARDS N00014-76-C-0074
UNCLASSIFIED TR-78-1-30 NL

1 OF 2
ADA
053962





1·0
2·8
2·5
5·0
3·15
2·2
5·6
3·5
1·1
2·0
4·0
4·5
1·8
1·25
1·4
1·6
NATIONAL BUREAU OF STANDARDS

12

TECHNICAL REPORT TR 78-1-30

DECISION THEORETIC AIDS FOR INFERENCE, EVALUATION, AND DECISION MAKING: A REVIEW OF RESEARCH AND EXPERIENCE

DECISIONS AND DESIGNS INCORPORATED

Gregory W. Fischer
Ward Edwards
Clinton W. Kelly, III

February 1978

ADVANCED ARPA DECISION TECHNOLOGY PROGRAM

CYBERNETICS TECHNOLOGY OFFICE
DEFENSE ADVANCED RESEARCH PROJECTS AGENCY
Office of Naval Research • Engineering Psychology Programs

The objective of the Advanced Decision
Technology Program is to develop and transfer
to users in the Department of Defense advanced
management technologies for decision making.
These technologies are based upon research
in the areas of decision analysis, the behavioral
sciences and interactive computer graphics.
The program is sponsored by the Cybernetics
Technology Office of the Defense
Advanced Research Projects Agency and
technical progress is monitored by the Office
of Naval Research – Engineering Psychology
Programs. Participants in the program are:

Decisions and Designs, Incorporated
Harvard University
Perceptronics, Incorporated
Stanford University
The University of Southern California

Inquiries and comments with
regard to this report should be
addressed to:

Dr. Martin A. Tolcott
Director, Engineering Psychology Programs
Office of Naval Research
800 North Quincy Street
Arlington, Virginia 22217

or

LT COL Roy M. Gulick, USMC
Cybernetics Technology Office
Defense Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, Virginia 22209

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the 6570th Aerospace Medical Research Laboratory, the Defense Advanced Research Projects Agency, or of any other agency of the U.S. Government.

TECHNICAL REPORT TR 78-1-30

DECISION THEORETIC AIDS FOR INFERENCE, EVALUATION, AND DECISION MAKING: A REVIEW OF RESEARCH AND EXPERIENCE

by

Gregory W. Fischer, Ward Edwards and Clinton W. Kelly, III

This research was supported by the
6570th Aerospace Medical Research Laboratory
under Contract F33615-73-C-4056 — new
and by the
Defense Advanced Research Projects Agency
under Contract N00014-76-C-0074
Monitored by the Office of Naval Research

390 664

February 1978

DECISIONS and DESIGNS, INC.

Suite 600, 8400 Westpark Drive
McLean, Virginia 22101
(703) 821-2828

DISTRIBUTION STATEMENT A

Approved for public release;
Distribution Unlimited

SUMMARY

Over the past twenty years, there has been increasing emphasis on research concerned with human decision making abilities and with the development of formal methods to aid decision makers in reaching logically consistent choices. This broad area of research is of particular importance in national security contexts where key decision makers must resolve extremely complex decision problems characterized by uncertainty, conflicting information, and enormously high stakes.

This technical report presents a summary of major portions of the literature bearing on people's ability to process information and to reach decisions. It also contains a review of laboratory and field assessments of judgmentally-based decision aiding systems embodying decision analytic concepts.

The evidence reviewed provides a strong research basis for the conclusion that unaided human judgment in complex inference and decision tasks is highly fallible. Formal algorithms (decision models) applied in these contexts typically yield better results than global human judgment. The data supporting these conclusions suggest that people are better at making simple judgments than they are at aggregating large amounts of information to form overall decisions.

Consistent with these findings, the decision aiding technologies reviewed in this report are based on principles of task disaggregation. A decision problem is divided into its relevant attributes, each of which is well within the judgmental capacities of the decision maker. People make

judgments about attribute probabilities and values, and formal models are used to aggregate these judgments to arrive at a decision. A large body of experimental and experiential evidence supports the notion that this divide-and-conquer approach leads to substantially better inferences and decisions than otherwise would be obtained. This research divides naturally into two parts, one dealing with probability judgments, the other with value (utility) judgments.

Probabilistic Information Processing (PIP) systems decompose the task of probabilistic inference. People identify relevant states of the environment and information sources; they also estimate likelihood ratios linking the data sources to the environmental states. Aggregating information across data is assigned to Bayes' theorem. The literature leading to the formulation of PIP systems and evaluating their application is extensively reviewed.

The second major input to decision making is judgments of value. This requires that each possible consequence of the action alternatives being considered be assigned a single numerical value reflecting the utility of that consequence relative to all other possible consequences. Both the theory and methods for assigning utilities to complex outcomes have recently become available. The technology based on multi-attribute utlity theory is exciting and promising, but still relatively in its infancy. The growing body of evidence, both published and unpublished, on development and application of this technology, as well as some of its as yet unsolved problems, is reviewed in depth.

CONTENTS

PREFACE

This technical report contains a review of the research literature bearing on human decision making abilities and presents a summary of field and laboratory experience with decision-theoretic aiding systems which have been developed to improve the coherence of judgments in operational decision settings. The review was initiated by Decisions and Designs, Incorporated (DDI) under support from the Defense Advanced Research Projects Agency. In the course of the literature search, DDI personnel happened upon an unpublished manuscript prepared in late 1973 by Gregory Fischer and Ward Edwards[1] (then at the University of Michigan), under support from the 6570th Aerospace Medical Research Laboratory. The paper by Fischer and Edwards turned out to be an excellent treatment of the same subject matter targeted for review by the DDI researchers and only slightly outdated by the four years that had elapsed since its preparation. The manuscript captured and well-reflected about 14 years of relevant research (roughly 1960 through 1973). With the kind permission of the 6570th Aerospace Medical Research Laboratory to use the unpublished report (and, of course, with the permission of the authors, as well), the scope of the planned review effort was changed to updating much of the Fischer-Edwards material with emphasis on the addition of field and laboratory assessments of decision aiding systems that had occurred since 1973.

[1] Dr. Fischer is now at the Institute of Policy Sciences and Public Affairs, Duke University, Durham, North Carolina. Dr. Edwards is at the Social Sciences Research Institute, University of Southern California, Los Angeles. The third author, Dr. Kelly, is with Decisions and Designs, Incorporated, McLean, Virginia.

Sections 1.0 through 3.0 of this report summarize the now-extensive literature which points to profound human fallibilities in probabilistic decision making tasks, and present data bearing on the usefulness of a variety of decision aiding concepts which stem from a Bayesian approach to decision making in the face of uncertainty. Inasmuch as the general implications to be derived from these sections are generally unchanged by research that has been conducted since 1973, and since an excellent review of the intervening literature has been recently completed,[2] Sections 1.0 through 3.0 remain essentially as presented in the original 1973 paper with only minor modifications and additions having been made.

In keeping with the major purpose of this review paper, which was to summarize evidence bearing on the worth (utility) of decision aiding methods, Section 4.0 has been expanded to reflect applied experience with a variety of decision aids, many of which were developed and introduced under the ARPA-supported Advanced Decision Technology Program.

For those who seek a definitive quantitative answer to the question of decision aid effectiveness, the results summarized here will be disappointing. Although field applications of decision aiding methods have occurred with increasing frequency over the past five years, no one is in a position to say if, or by how much, actual decisions were improved through the use of aiding methods. There are many reasons for this apparent shortcoming, not the least of which is the very practical problem of having a busy, often harried, operational decision channel operate in a structured

[2] Slovic, P., Fischhoff, B., and Lichtenstein, S. (1977). Behavioral Decision Theory. Annual Review of Psychology. 28:1-39.

experimental mode for measurement purposes. Any such requirement is bound to bring the delicate matter of technology transfer to an early end. Other problems, some of which are unique to decision aid evaluation, include: the good decision-bad outcome phenomenon, the lack of objective criteria against which to evaluate decisions, and the fact that most significant decision problems (those that would merit aiding) are unique, single-occurrence events not amenable to the systematic accrual of evidence.

For these and other reasons, the state of assessment of operational decision aids is less objective and definitive than desired. Much of the evidence is anecdotal; more of it comes from laboratory work via construct and convergent validation approaches, a literature reviewed in part in this report. While the validation work continues and until occasional natural experiments in field settings offer firmer evidence of the efficacy of decision aids, the question of their value will be answerable only in terms of extensions of laboratory evidence, axiomatic reasoning, and through user acceptance.

DECISION THEORETIC AIDS FOR INFERENCE, EVALUATION, AND DECISION-MAKING: A REVIEW OF RESEARCH AND EXPERIENCE

1.0 INTRODUCTION

Computer-based decision aiding systems have been oriented primarily to the problems of collecting, displaying, storing, and retrieving information. This reflects a common belief that lack of information is the major obstacle to good decision making (Slovic and Lichtenstein, 1971). Modern military command posts, with their elaborate displays and communications systems, typify this philosophy. But after two decades of research on human decision making and judgments, psychologists are now in a position to argue persuasively that this is not an adequate approach to decision system design. The available research demonstrates that people are severely limited information processors who base their decisions on a small number of items of information (Slovic and Lichtenstein, 1971; Slovic, Fischhoff, and Lichtenstein, 1977). Thus, management information systems that merely provide decision makers with large amounts of data are misdirected; the decision makers will be unable to use most of this information.

Edwards (and many others) has proposed a different approach to the design of decision-aiding systems. (See, for example, Edwards, Lindman and Phillips, 1965). Edwards' approach is based upon a Bayesian formulation of the decision-making process. From this point of view a decision problem is decomposed into two major subtasks: (a) diagnosing the state of the decision maker's environment, and (b) based upon this diagnosis, selecting that course of action with the highest expected utility.

This review paper is divided into two major sections. The first summarizes a body of literature about the human ability to process information and make decisions. The second section discusses the current state of the art in judgmentally based decision-making systems. Here, we consider not only the Bayesian approach, but also the regression or "bootstrapping" approach to augmenting people's decision-making capabilities.

Throughout this discussion, we will often use the term "decision maker" as if it referred to a single person. This usage partially reflects the fact that research has been almost exclusively focused on the decision-making processes of single individuals. But more importantly, the usage reflects the fact that normative decision theory assumes the existence of an actor with an internally consistent set of beliefs and values. While such an assumption may be dubious in the case of individuals, it is obviously false at the level of complex organizations (March and Simon, 1958; Allison, 1971). Later in this paper, we will argue, however, that normative decision theory can and should be used to reconcile conflicts of belief and value within organizations so that organizations, or at least organizational subunits, may act in a rational and internally consistent fashion. Thus, while most of the experimental studies we review will treat the inferences and decisions of single individuals, most of the applied decision-aiding work we discuss has involved multiple decision makers operating in an organizational setting.

2.0 DESCRIPTIVE STUDIES OF HUMAN DECISION MAKING CAPACITIES

Following the Bayesian formulation of a choice situation, our discussion of human decision-making abilities is divided into two sections. The first deals with people's ability to make inferences about the state of the environment and about the possible consequences of the actions that might be undertaken. The second considers human ability to integrate these inferences with judgments about the value of the possible consequences of actions to arrive at a final decision. In this framework, probabilistic inference is the first stage in the overall decision-making process.

2.1 Probabilistic Information Processing

2.1.1 Simple Bayesian inference - Decision makers typically find themselves in possession of one or more items of information each of which is relevant to diagnosis of the state of their environment, but none of which are definitive. The decision maker's goal in such a situation is to extract from the data at hand the maximum amount of information about the state of the environment and about the possible consequences of actions in that environment. In addition, the decision maker will want to revise these opinions in the light of any new evidence that becomes available during the course of the decision-making process.

Bayes' theorem provides a formally optimal procedure whereby the decision maker can aggregate information across a set of independent data. Consider the case where the decision maker can construct an exhaustive, finite set of mutually exclusive hypotheses about the state of the environment, and let these hypotheses be denoted by H_1,

H_2, ..., H_n.[1] For example, H_1 might be the hypothesis "The enemy will attack objective X tomorrow," H_2 the hypothesis "The enemy will attack objective Y tomorrow," and H_3 the hypothesis "The enemy will remain in his present defensive position tomorrow." Note that exhaustiveness is often obtained by fiat. Definition of a meaningful and manageable set of hypotheses will often represent a major obstacle to applying Bayesian decision theory in real world contexts. Here we simply assume that such a set of hypotheses has been created. Next, let D_1, D_2, ... D_m be a set of data or observations bearing on the true state of the environment. Then, when certain additional assumptions are met, Bayes' theorem may be sequentially applied to revise the decision maker's opinions in light of these data. In particular

$$P(H_i | D_k) = \frac{P(D_k | H_i)\ P(H_i)}{P(D_k)}. \qquad (1)$$

Here $P(H_i)$ is the *prior probability* assigned to hypothesis H_i in light of all past data but prior to the consideration of D_k; $P(H_i | D_k)$ is the *posterior probability* of H_i in light of D_k and all prior information; $P(D_k | H_i)$ is the probability that D_k would be observed given that H_i is true; and $P(D_k) = \sum_i P(D_k | H_i)\ P(H_i)$ is the unconditional probability of observing D_k. Bayes' theorem may also be written in the odds likelihood ratio form

[1] In many situations, the possible states of the environment depend on the course of action selected by the decision maker. This complicates the inference process but does not alter the logic of the process or the rational solution. Henceforth, we generally ignore the distribution between predicting states of nature and predicting states conditional on acts.

$$\frac{P(H_i|D_k)}{P(H_j|D_k)} = \frac{P(D_k|H_i)\ P(H_i)}{P(D_k|H_j)\ P(H_j)}, \tag{2a}$$

or

$$\Omega_1 = L\Omega_0, \tag{2b}$$

where

$$\Omega_1 = \frac{P(H_i|D_k)}{P(H_j|D_k)}, \text{ the } \underline{\text{posterior}}\ \underline{\text{odds}} \text{ of } H_i \text{ to } H_j,$$

$$L = \frac{P(D_k|H_i)}{P(D_k|H_j)}, \text{ the } \underline{\text{likelihood}}\ \underline{\text{ratio}} \text{ for } D_k,$$

and

$$\Omega_0 = \frac{P(H_i)}{P(H_j)}, \text{ the } \underline{\text{prior}}\ \underline{\text{odds}} \text{ of } H_i \text{ to } H_j.$$

These two forms of Bayes' theorem assume: (a) that each datum is reliably observed and reported, (b) that the state of the environment remains stationary (or constant) during the time period of interest, and (c) that the data observed are conditionally independent of one another with respect to the hypotheses of interest. To define the conditional independence condition more precisely, D_k is said to be conditionally independent of D_l with respect to H_i if and only if $P(D_k|H_i, D_l) = P(D_k|H_i)$.

This simple formulation of a probabilistic inference problem has given rise to over fifteen years of research in which people's inference processes have been compared with those prescribed by Bayes' theorem. Rapoport and Wallsten (1972) and Slovic, Fischhoff, and Lichtenstein (1977) provide excellent reviews of this literature. In general, humans have been found to be very suboptimal processors of probabilistic information. Although they typically revise their opinions in the same direction as Bayes' theorem, they do not revise them enough. This conclusion is based on a variety of experimental tasks in which subjects have been asked to make inferences about which of two or more statistical models has generated a given set of data. For example, Phillips and Edwards (1966) used bookbags filled with poker chips as binomial data generators. In one condition, one bag contained 60 red and 40 blue chips, the other 40 red and 60 blue. On each trial, one of the two bags was randomly selected. Then a sample of chips was drawn from the bag, one at a time, each sampled chip being returned to the bag before the next one was selected. The primary advantage of this and other tasks involving statistical data-generating processes is that they permit the calculation of "objectively" optimal Bayesian posterior odds. Since the Phillips & Edwards study many other experiments have replicated the original finding that people are conservative information processors, that is, that they extract less certainty from sample data than does Bayes' theorem. Moreover, this result has been obtained not only with binomial data-generating processes, but also with multinomial (Phillips, Hays, & Edwards, 1966) and normal data generators (DuCharme & Peterson, 1968).

After the conservatism effect was well established, many investigators shifted their attention to the problem of determining its cause. One popular hypothesis has been that subjects misperceive the diagnosticity of data. In terms of Equation 2, they incorrectly assess the likelihood ratio. Lichtenstein and Feeney (1968), for example, asked subjects to predict which of two targets was under attack by observing the locations where bombs actually impacted. In making these inferences, subjects were told that the distribution of bomb hits around the intended target was described by a circular normal density function. The subjects, however, apparently simplified the task by considering only the ratio of the distances from the bomb's impact point to each of the two targets. A number of other experiments have also found that subjective sampling distributions deviate substantially from the formally correct sampling distributions. In several cases, subjective sampling distributions have been found to be too flat, a result which leads to the prediction of conservatism (Peterson, DuCharme, & Edwards, 1968; Wheeler & Beach, 1968). Even worse, Kahneman and Tversky (1972) have found that subjective sampling distributions are almost totally unaffected by sample size, an extremely severe violation of the formally correct statistical models.

A second explanation of the conservatism effect hypothesizes that subjects are unable to aggregate information across data (Edwards, 1968). This hypothesis arose naturally from the observation that conservatism was substantially greater in large samples of data than in small samples. In her dissertation research, Wheeler (1972) obtained very strong support for the misaggregation hypothesis. She found that over a very wide range of odds levels, odds revisions based on but a single datum were nearly optimal. But over the same range of odds levels, inference based on several data were quite conservative.

Misperception and misaggregation were originally viewed as competing hypotheses. For instance, the suboptimality of subjective sampling distributions was explained as arising from the misaggregation of the data in the sample (Edwards, 1968). By now, however, it is clear that misperception and misaggregation are complementary explanations of conservatism and that both contribute to the suboptimality of human inferences (Phillips, 1966; Rapoport & Wallsten, 1972).

Still another interpretation of conservatism is the hypothesis that a response bias effect causes it. DuCharme (1967), for example, using two normal distributions differing only in mean as his data generators, found that subjects were reasonably accurate in odds estimates between 1:10 and 10:1, and conservative outside that range. This combines with the fact that almost any data aggregation experiment requires very extreme posterior odds or probabilities to suggest that perhaps subjects are simply reluctant to estimate such large numbers. This, of course, is a particular version of the misaggregation interpretation--a version quite different in spirit from that which supposes that something goes systematically wrong with the data aggregation process inside the head.

DuCharme's finding and interpretation are sharply challenged by Wheeler's Ph.D. thesis (1972). She also used two normal distributions differing only in mean, using visually displayed extents instead of numbers as her stimuli. The most important feature of her experiment was its use of very carefully constructed sequences of stimuli, including some in which the Bayesian posterior odds remained well within the range from 10:1 to 1:10 for sequences of ten stimuli. Her main finding was that even within that range subjects were systematically conservative--as conservative within the range as outside it. She found no more conservatism for extreme odds than for odds falling within the central range, once the effect of number of data items is taken into account.

The issue of the locus of conservatism is not totally resolved as an abstract scientific question. It seems likely that both misaggregation and misperception play a role; response biases may well do so also. While the issue is important for science, it is perhaps less important for application. From the point of view of application, the main point is that conservatism is wasteful. Indeed, the accuracy ratio, the most frequent index of conservatism in this literature, can be directly interpreted as an index of waste.[1] An accuracy ratio of .20, for example, means that 80% of the data (and so perhaps of the money spent gathering data) have been wasted on suboptimal information processing--in the sense that optimal information processing would have reached the same degree of certainty on the basis of only 20% of the information that was in fact used. The challenge, then, is that of reducing or eliminating this waste.

Some scientists interested in this topic have taken to questioning the whole line of research and thought that bears on conservatism as irrelevant to psychology. They argue that Bayes' theorem and Bayesian posterior odds are external, arbitrary standards, irrelevant to human thinking processes, and that conservatism is not a psychological phenomenon. This argument seems quite strange to us. It is somewhat like saying that comparison of a scale's reading with the true weight of the object being weighed is irrelevant to understanding the scale. In a sense, that may be true. Yet anyone interested in using that scale for its intended practical purpose will be especially interested in comparing its reading with known true weights. Moreover, such comparison easily leads to more profound understanding of the processes underlying the scale's operation, as has also happened in the conservatism literature. It is always useful in psychophysics to have a model of the stimulus, and to relate sensory phenomena to that model. In human probabilistic inference, models of the data-generating process

[1] See Section 2.1.8.

(where known) and Bayes' theorem together constitute the appropriate model of the stimulus--and are as useful to the study of human inference as Fourier analysis is to the study of the psychophysics of sound.

2.1.2 Multi-stage inference - The inference tasks described above are considerably simpler than many faced in real-world contexts. Here we consider the additional complexities introduced when the inference process involves several levels of analysis. Consider, for example, the case of data reports subject to error (as they will be in almost all realistic contexts). In the context of a hypothetical military intelligence problem, suppose that a commander wishes to predict whether the enemy will (H_1) or will not (H_2) attack his position X in the following week, and that on the basis of all prior information the commander believes that it is equally likely that the enemy will or will not attack. That is, $P(H_1) = P(H_2) = .5$. Assume also that the commander has intelligence reports that indicate whether the enemy is (D_1) or is not (D_2) massing troops and supplies for this attack. These intelligence reports are not wholly diagnostic, however, for the enemy might give the appearance of attacking objective X in order to divert friendly forces from the true objective. In addition, the enemy can attack without making preparations by using reserves which he has stored. Suppose that the commander's beliefs about whether the enemy will make preparations given that he is or is not planning to attack are summarized by the matrix below.

	$P(D_1 \mid H_i)$	$P(D_2 \mid H_i)$
H_1: Attack	.9	.1
H_2: No Attack	.1	.9

For example, the commander believes that if the enemy decides to attack, the odds are 9:1 that he will mass troops and supplies for the attack.

The commander's inference problem about the enemy's preparations is not completely reliable. Suppose that the enemy's security is very good, so that friendly intelligence sources may fail to detect preparations which are undertaken. In addition, assume that friendly intelligence sources may falsely report attack preparations when none in fact have occurred. Let R_1 denote the intelligence report "It looks as if the enemy is preparing to attack," and R_2 the report "No enemy attack preparations have been observed." Suppose that the commander's beliefs about the reliability of his intelligence reports is summarized by the matrix below.

	$P(R_1$: "Preparing"$\|D_i)$	$P(R_2$: "No Preparations"$\|D_i)$
D_i: Preparation	.6	.4
D_i: No Preparation	.2	.8

For example, the commander believes that the odds are 3:2 that his intelligence sources will detect enemy preparations if they occur, but 1:4 that his sources will falsely report preparations which have not been undertaken. The entire structure of this inference problem can be represented by the probability tree in Figure 2-1.

Even though this problem is highly simplified, it is not intuitively apparent what conclusions the commander

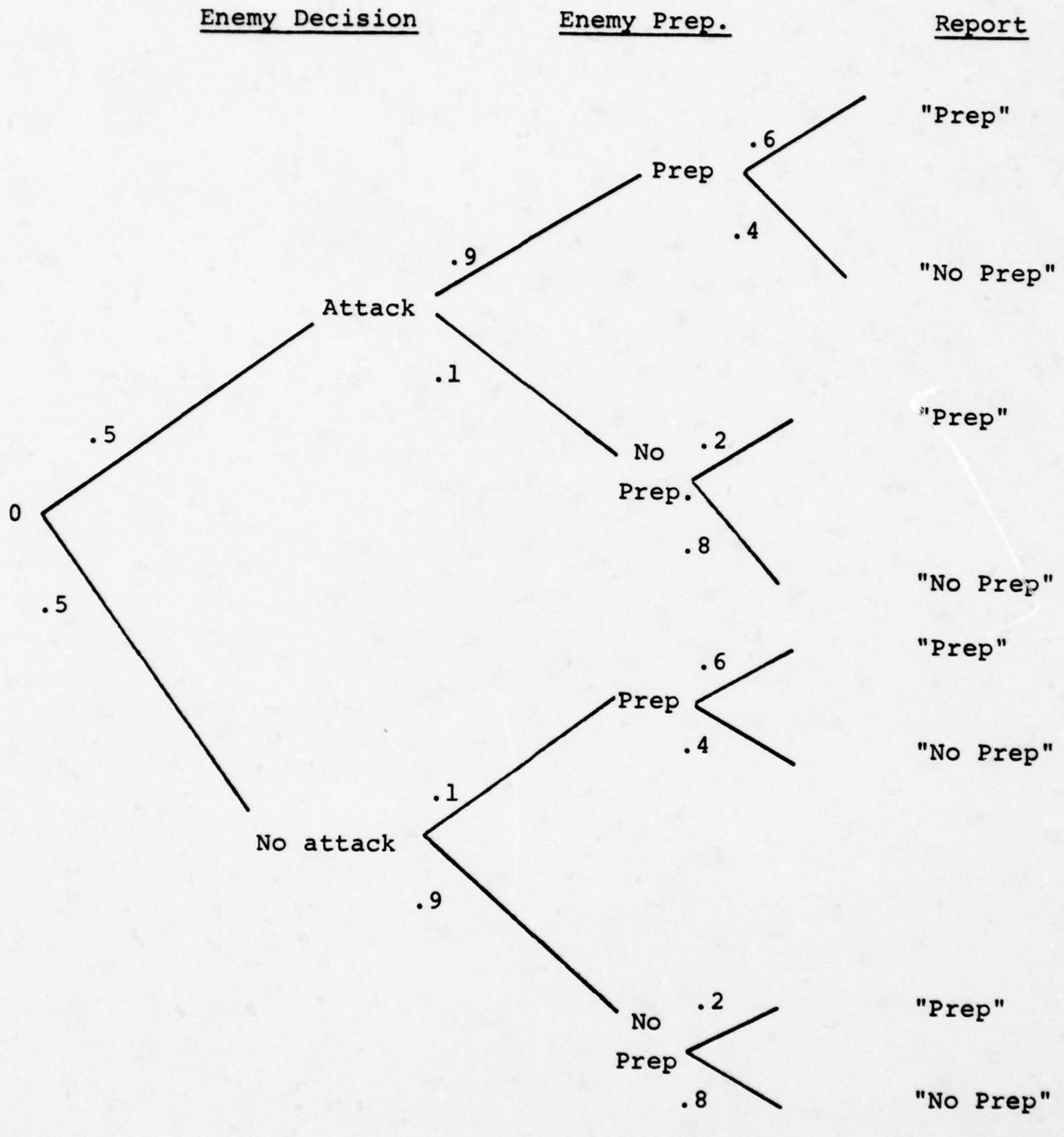


FIGURE 2-1
PROBABILITY TREE FOR UNRELIABLE INTELLIGENCE PROBLEM

should draw. For example, suppose he receives the report that the enemy has been observed making preparations for an attack. One strategy which experimental subjects sometimes adopt is to ignore the unreliability of the report, treating it as a true datum (Kelly, 1972). If a commander adopted this inference strategy in the present example, the report of enemy preparations would lead him to assign 9:1 odds to the hypothesis that the enemy would attack. A second and intuitively more appealing strategy would be to degrade the 9:1 odds of attack to take account of the possible unreliability of the report; for example, multiplying the 9:1 odds by .6, the probability that the report is correct, to obtain final odds of 5.4:1 in favor of the attack hypothesis. Both of these inference strategies lead to extreme estimates, however, for the objectively correct posterior odds in favor of attack are only 2.33:1. In fact, both of the two heuristic approaches to multi-stage inference discussed above will produce extreme estimates in a wide variety of contexts.

Experimental studies of intuitive multi-stage inference have generally found that subjects' inferences are excessive relative to multi-stage versions of Bayes' theorem (Snapper & Fryback, 1971). In one particularly interesting experiment, Schum, DuCharme, and Pitts (1971) had subjects make tachistoscopic observations of data from which they were to make inferences. Here unreliability was introduced by the subjects' own perceptual errors. As in other multi-stage inference experiments, estimates are generally excessive relative to Bayes' theorem. Subjects adjusted their inferences to reflect the unreliability of their observations, but not enough.

Gettys, Kelly, and Peterson (1973) also obtained extremism in a multi-stage inference task involving multinomial data generators. They found that their subjects fall into two groups. Median data for the larger of the two groups was almost perfectly described by the "best-guess" strategy. This is the second of the heuristic strategies described above. Median data for the smaller group of subjects was more nearly approximated by the optimal Bayesian odds. The Bayesian model differs from the best-guess strategy in that the best-guess strategy considers only the datum favored by the report, then degrades the inference based on this datum to reflect the unreliability of the report. The Bayesian model, on the other hand, considers also the possible truth of the datum not favored by the report. Steiger and Gettys (1972) have obtained further support for the best-guess model in a task in which subjects merely indicated which of two hypotheses was more likely on the basis of unreliable data.

In contrast to the studies cited above, Youssef and Peterson (1973) found that multi-stage inferences, while generally suboptimal, are not always excessive relative to Bayes' theorem. They found that, for a given cumulative Bayesian posterior odds, multi-stage inferences were always excessive as compared to simple inferences. When the optimal posterior odds were large, both simple and multi-stage inferences were conservative; when the optimal posterior odds were small, both were excessive. It is significant that this experiment differs from the ones discussed above in that inferences were based on several observations rather than one. The best-guess strategy, which generally produces extreme estimates for one-observation inference problems, may not be as intuitively appealing in the multi-observation case.

Although more research on the topic of multi-stage inference is required, it is by now clear that such inferences are frequently very suboptimal. And in many cases they are substantially excessive. From a decision-making standpoint, excessiveness seems even more dangerous than conservatism. Falsely concluding that a potential enemy had initiated an attack, for example, might precipitate a disastrous exchange of nuclear weapons.

2.1.3 Non-stationary environments - Inference problems are further complicated when the assumption of a stationary (unchanging) environment is violated. Yet many decision-making systems are designed explicitly to detect changes in a non-stationary environment. The primary function of early warning systems, for example, is to detect transitions from a state of peace to a state of war.

We will illustrate the logic of non-stationary inference with another simple example. Suppose a commander has reason to believe that the enemy will attack his positions, and that his prior odds favor attack (H_1) over non-attack (H_2) by 3:2. Suppose also that he can obtain data as to whether the enemy is (d_1) or is not (d_2) marshalling tanks for the attack, and that the odds are 7:3 that a tank buildup will be observed if an attack is planned, but only 3:7 if no attack is planned. (Here the problem is simplified by assuming that all reports are completely reliable.) The commander's inference problem is complicated by the fact that the enemy commander can change his mind. That is, even if he decides not to attack today, he may reverse his decision tomorrow. Assume that the friendly commander believes that the odds that the enemy will switch from a "no attack" to "attack" decision, given an initial "no attack" decision, are 50:50. But if the enemy commander initially makes the "attack" decision, he will have to go through with it.

This inference problem can be represented by the multi-stage inference tree of Figure 2-2. Here H_1^i denotes "attack decision made on day i," for i = 1, 2. Similarly, d_1^i denotes "tank buildup observed on day i," and d_2^i denotes "no tank buildup observed on day i," for i = 1, 2.

A few simple calculations reveal that one can get into serious trouble by treating a non-stationary environment as if it were stationary. Suppose, for example, that no tank buildup is observed on day 1, but a buildup is observed on day 2. If we treat the environment as stationary, an application of Bayes' theorem shows that these two observations offset one another, and

$$\frac{P(H_1^2 | d_2^1, d_1^2)}{P(H_2^2 | d_2^1, d_1^2)} = 60{:}40 \text{ or } 1.5{:}1,$$

the prior odds of an attack. But the correct odds, as calculated from Figure 2-2, are 5.33:1 in favor of attack. Failure to consider non-stationarity in the problem outlined above would lead to extremely bad inferences.

Actually, the preceding discussion also illustrates the slipperiness of the concept of non-stationarity. In order to draw Figure 2-2 and do the calculation based on it, it was necessary to conceptualize the environment in a manner that treated the possibility of change as a stationary fact of life. Stationarity is a property of any model; it simply means that the character and parameters of the model remain unchanged over the time period of interest. Non-stationary models are useless -- and so models of non-stationarity are sometimes indispensable. The way to deal with a changing world is to find a conceptualization that treats the possibility of change in an unchanging way. Figure 2-2 is an example.

Experimental studies of intuitive inference have revealed that people show a surprising ability to detect changes in non-stationary processes. Rapoport (1964), for example, had subjects observe events generated by a binomial process and asked them to estimate the parameter p of that

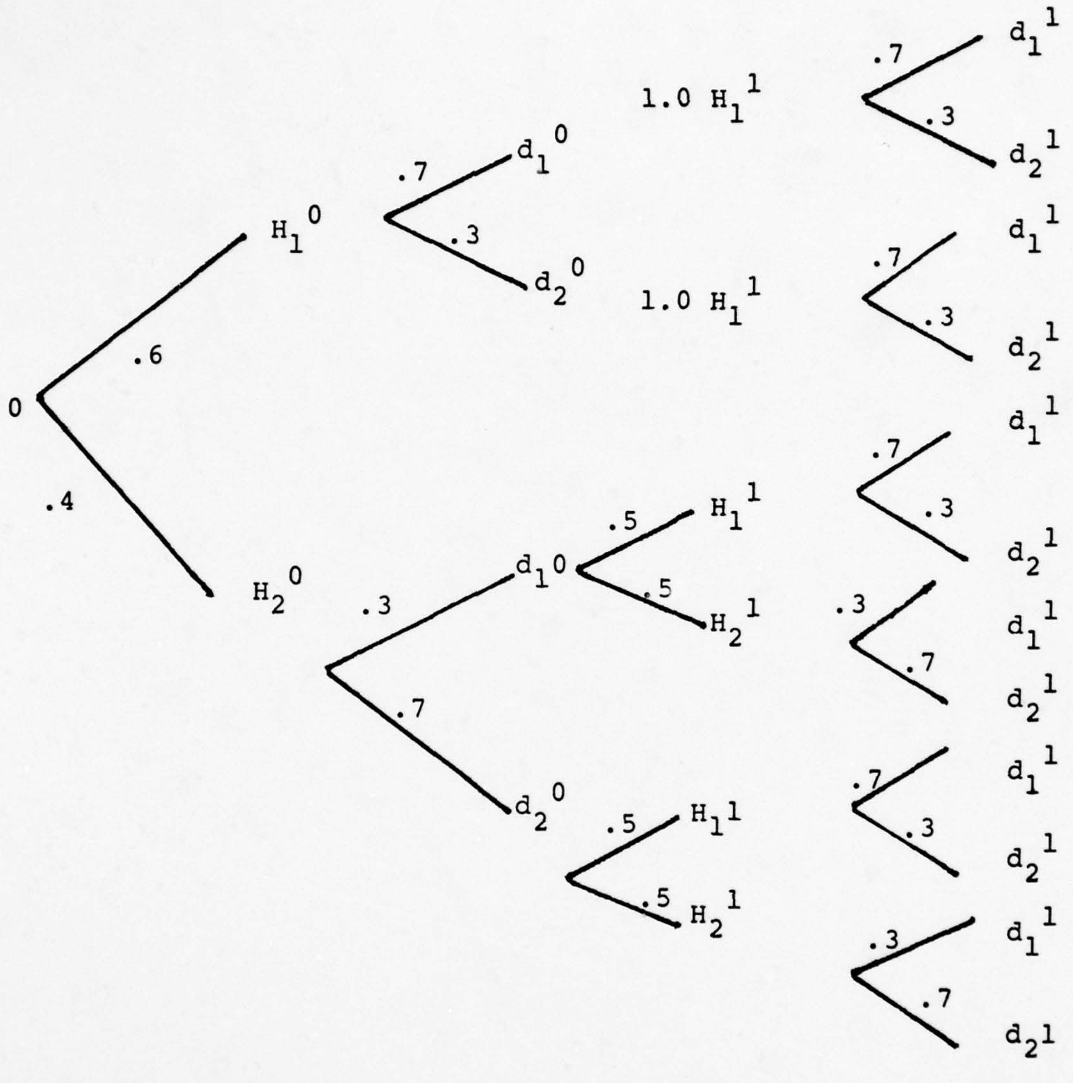


FIGURE 2-2

PROBABILITY TREE FOR NON-STATIONARY ENVIRONMENTS process. They were not warned that the actual parameter value of the data generating process would be changed during the course of the experiment. Despite this, the subjects did a good job of tracking the first shift, and an excellent job of tracking subsequent shifts. Robinson (1964) obtained the same results in a similar experiment.

In a later study, Chinnis and Peterson (1970) asked subjects to make inferences about which of two binomial processes had generated a sequence. Their task differed from standard bookbag-and-poker-chip paradigms, however, because there was one chance in ten that a new bookbag would be selected between trials. Thus, their task closely resembled the non-stationary attack scenarios discussed above. Subjects' estimates were generally conservative relative to the optimal Bayesian inferences. Nonetheless, statistical analyses revealed that the subjects' judgments could be much better approximated by a non-stationary Bayesian model than by a stationary Bayesian model.

So although only a small amount of experimental evidence is available, it would appear that people are rather good at making intuitive inferences about non-stationary processes. More studies are required, however, with particular emphasis on complex inference situations. For it seems quite possible that when burdened with a large amount of information, people will be forced to adopt simplifying strategies which will produce suboptimal inferences. Even in the Chinnis and Peterson study, subjects exhibited a marked degree of conservatism.

2.1.4 <u>Inferences based on conditionally nonindependent data</u> - In all of our previous discussion we have made the simplifying assumption that all data are independent with respect to the hypotheses of interest, that is, that $P(D_i|D_j,H_k) = P(D_i|H_k)$ for all i, j, k. This assumption cannot be justified in many real-world contexts.

For example, suppose that D_1 and D_2 are photographs of a given point taken by two different aircraft and that the second photograph was taken one hour after the first. Suppose also that an intelligence analyst examines both photos and observes a large fuzzy object which he believes might be a well camouflaged enemy missile site. Both photos are clearly relevant to the hypothesis that the enemy has deployed missiles in the area being photographed. But they are not independent with respect to this hypothesis. For given the presence of a large fuzzy object on the first photo, it is highly likely that a similar object will appear on the second photo, regardless of whether it is a missile or not. That is, $P(D_2|D_1,H) > P(D_2|H)$. In this example, D_1 and D_2 are redundant, and failure to compensate for this redundancy would lead to an excessive estimate of the probability that the enemy is deploying missiles. Examples can also be constructed in which the occurrence of two data is more informative with respect to a given hypothesis than would be indicated by the independent consideration of each datum.

Several studies suggest that people are able to adjust their intuitive inferences quite well when faced by conditionally nonindependent data. Schum (1966) had subjects make inferences using six data sources. Two pairs of data sources were nonindependent in some experimental conditions. Subjects were alerted to the possibility of nonindependence, told where to look for nonindependence, and asked to tally the joint frequency of occurrences of data from the designated sources. Given all this help, subjects did a fairly good job of adjusting for the nonindependence. Schum found that a Bayesian model which took account of nonindependence provided a much better fit to the mean data than did a simple Bayesian model which treated the data as independent.

Nevertheless, rank order correlations between the individual subjects' responses and the Bayesian model were only moderate, typically in the .5 to .65 range.

In a similar series of three experiments, Schum, Southard, and Wombolt (1968) again found that intuitive inferences based on six items of data with two-way nonindependencies were quite close to the optimal Bayesian values. But with samples of 12 or 18 data with pairwise nonindependencies, intuitive inferences were substantially conservative.

Again it is difficult to come to firm conclusions on the basis of so little experimental evidence. The data available suggest that subjects can make fairly good inferences on the basis of nonindependent data provided that the number of data being processed is small, but that they do poorly with large samples of data. It should also be noted that these studies bypassed the very difficult problem of locating nonindependencies. Subjects were told which data sources were correlated, and the use of frequency matrices made the nature of the nonindependencies fairly transparent. Further, subjects never had to cope with three-way or higher order data interactions. Finally, it should also be noted that the two studies cited above differed from most simple inference experiments by using subjects with extensive experience in inference tasks. Naive subjects might have behaved less optimally.

2.1.5 Multi-cue inference - All of the studies discussed above have utilized Bayesian inference models as a normative standard against which intuitive inferences may be evaluated. A second research tradition has utilized linear regression models as a standard for evaluating behavior. Regression studies typically use several data sources or cue dimensions, and one response dimension, which is usually continuous.

For example, a college applicant might be described by the cues SAT verbal score (X_1), SAT math score (X_2), and high school grade point average (X_3). The subject could then be asked to predict the applicant's first-year grade point average in college (Y). Using a linear regression model a statistician could estimate the coefficients b_1, b_2, b_3, and b_4 in the equation

$$\hat{Y} = b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 + \varepsilon$$

where b_1, b_2, and b_3 are weighting constants for the cues X_1, X_2, and X_3, b_4 is an additive scaling constant, and ε is a random error term with mean zero which reflects the fact that Y is only probabilistically related to the cues X_1, X_2, and X_3. The subjects' performance in such a task is usually evaluated by comparing the correlation between the subjects' estimates (Y_e) and the true Y scores with correlation between the regression model's estimates ($\hat{Y}$) and the true Y.

From the subjects' standpoint this task differs from typical Bayesian tasks in two respects. First, the response dimension Y is continuous, which is equivalent to stating that the subject has an infinite hypothesis set. Bayesian studies seldom use more than five discrete hypotheses. Second, the subject is not asked to assess a probability distribution over the response dimension Y, but rather to specify only the expected value of Y. If subjects were asked to assess a continuous probability distribution over Y, the responses in a regression experiment could be evaluated against a continuous version of Bayes' theorem.

Slovic and Lichtenstein (1971) provide a comprehensive review of multi-cue inference studies, so here we merely summarize the main findings of this rather extensive literature, providing representative rather than exhaustive references.

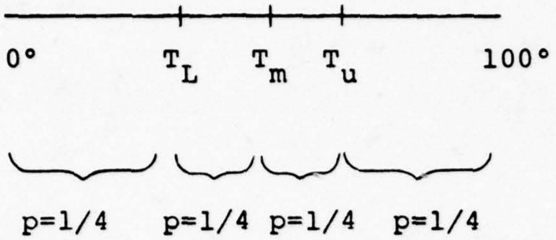
Bayesian studies typically provide the subjects with information about how each data class relates to each hypothesis by providing them with the $P(D_i|H_j)$ matrix. The emphasis here is on how the subject will aggregate information across hypotheses. In regression studies, on the other hand, subjects must discover the relation between each of the cue dimensions and the criterion dimension Y. They do this by making estimates and then receiving feedback on the true state of the criterion variable Y. In the simplest regression paradigm where there is only one cue, subjects typically do an excellent job of learning the cue-criterion relationship, though they have more difficulty learning negative relations than positive ones. (Bjorkman, 1965; Naylor & Clark, 1968). Subjects also do better, relative to the standard set by the optimal regression equation, when the cue-criterion relation is strong, that is, when the variance of the random error distribution ε is small (Bjorkman, 1968).

People also do very well relative to the optimal regression policy in multi-cue learning situations when cues are linearly related to the criterion (Smedslund, 1965). They can also learn to utilize non-linear cue-criterion relations, but not as well (Hammond & Summers, 1965), and one study has found that people can learn to predict non-additive cue-criterion relations such as $Y = X_1 - X_2$ or $Y = X_1 X_2$ (Brehmer, 1969). The study of non-additive cue utilization considered two cues only, however, and even here subjects' learning was very slow. Subjects can also learn to correct their implicit cue weightings when confronted by a non-stationary environment (Peterson, Hammond & Summers, 1965).

Together, the multi-cue learning studies suggest that people are rather good at making point estimate predictions of a criterion variable on the basis of multiple cues, provided that they receive extensive outcome feedback. It would be quite interesting to extend these studies by having subjects assess a whole probability distribution over the criterion variable rather than just make a point estimate.

2.1.6 Real-world inference: Bayesian studies - A number of investigators have gone beyond the laboratory to see how well people do in making inferences about real-world processes. In contrast to laboratory studies, these studies tend to involve decision makers with some degree of expertise about the process being studied. Because these studies provide us with the only firm evidence about human inferential capacities in the real world, each is considered in some detail.

Peterson, Snapper and Murphy (1972) asked two experienced meteorologists to make predictions about the following day's high and low temperatures using a credible interval procedure. (We illustrate this for the high temperature only.) For each prediction the forecaster began by specifying a temperature T_m such that he thought that it was equally likely that the actual high temperature would be above or below T_m. That is, $P(T > T_m) = P(T < T_m) = 1/2$. Next the forecaster was asked to estimate a $T_u > T_m$ such that $P(T_m \leq T \leq T_u) = P(T > T_u) = 1/4$. That is, T_u divided the region above T_m into two subjectively equally likely regions. Finally, each forecaster was asked to specify a $T_L < T_m$ such that $P(T < T_L) = P(T_L \leq T \leq T_m) = 1/4$. Thus, the three temperatures T_L, T_m, T_u divided the high temperature dimension into four subjectively equally likely regions as illustrated below



Peterson, Snapper, and Murphy (1972) evaluated these estimates in two ways. First, they compared T_m, the forecaster's median estimates, with the actual temperatures recorded. A very strong linear relation between the predicted and actual temperatures was observed. The forecasts were also evaluated by noting how many of the actual temperatures fell between T_L and T_u. Ideally, one would hope that approximately 50% of the observations would fall in this region, for this would indicate that the forecaster's subjective likelihoods were well calibrated with their environment. In fact, however, too many observations fell in this central 50% credible interval, indicating that the two forecasters studied were too conservative in estimating T_L and T_u. They should have made estimates which were more tightly clustered around T_m.

Murphy and Winkler (1973) conducted a similar study and found not only that the median T_m estimates were very good, but also that the central 50% and 75% credible intervals were accurate in the sense that the proportion of actual temperatures falling in a region closely approximated the subjective probability assigned to that region by the bisection method described above. But, when forecasters were asked to directly estimate the probability of a temperature falling in a fixed region, they tended to overestimate these probabilities. The authors concluded that the procedure of dividing a continuum into subjectively equally likely regions may result in objectively more veridical probability assessments than does the procedure of having subjects directly assign probabilities to fixed intervals.

Staël von Holstein (1971) also studied the ability of meteorologists to make predictions. He asked experienced forecasters, university meteorologists, meteorology students, and statisticians to directly assess the probability that temperature and amount of precipitation would fall into fixed intervals. These intervals were selected so that, on the basis of prior frequency data, each interval was equally likely. To evaluate these forecasts Staël von Holstein used the quadratic scoring rules

$$S_k(p) = 2p_k - \sum_i^h p_i^2.$$

Here p_i denotes the probability assigned to the i-th category, p_k the probability assigned to the category in which the observed event actually fell, and $S_k(p)$ the score for the probability vector $p \equiv (p_1, p_2, \ldots, p_n)$ given that the observed event fell in the k-th category. Clearly, the higher the probability assigned to the category which contains the observed event, the higher the forecaster's score. In addition, the quadratic scoring rule has the property that the forecaster can maximize his expected score only if he states his true opinion; that is, there is no incentive to hedge one's bets or to overstate the probability of a likely category in hopes of receiving a high score (Staël von Holstein, 1970). Using this scoring rule, only seven of the thirty subjects received a higher score than one would have obtained using the past frequency data. In view of the large amount of additional information available to the forecasters, this result is rather disappointing. On the average, the subjects received a score which was 95.7%

that of the relative frequencies. Interestingly, the university meteorologist outperformed the professional weather forecasters. The professional forecasters were too confident, assigning excessively high probabilities to what they believed were the most likely outcomes.

In a similar study, Staël von Holstein (1972) asked market analysts, bankers, business students, and statisticians to assign probabilities to five price-change ranges for selected issues on the Swedish stock market. Again, for each issue, the price-change ranges were selected to be equally likely on the basis of past performance. Forecasts were made for two-week intervals over a period of several months. Here, only three of seventy-two subjects beat the historical frequencies, with the average subject receiving 94% the score of the historical model.

Together, these studies do not support the belief that people are skilled intuitive statisticians. Only the point estimates of the next day's temperature were highly accurate. And 24-hour temperature forecasting is relatively trivial given the inertia in temperature; one can do very well by using today's high temperature to predict tomorrow's.

The failure of Staël von Holstein's subjects to outperform the historic frequencies is quite discouraging. For the historical models were not conditioned on any of the information available to the subjects. One possibility is that, despite the use of scoring rules, the method of having subjects directly assign probabilities to event classes does not provide a good means for extracting what they really know. The Murphy and Winkler (1973) results suggest that the credible interval procedure of successively dividing a continuous variable into subjectively equally likely ranges will produce more veridical predictions. It seems unlikely, however, that changes of response mode will lead to a dramatic improvement in predictions. Rather, these real world studies provide strong support for those who argue that people are suboptimal probabilistic information processors who need all the help they can get from statisticians and decision theorists.

2.1.7 Real-world studies: multi-cue inference paradigm - With his classical review Meehl (1954) began a controversy that has not yet been completely resolved after over twenty years of additional research. Clinical psychologists and medical doctors spend a substantial portion of their time classifying patients into disease categories on the basis of multiple signs and symptoms. Since these signs and symptoms are only probabilistically related to the patients' true states, one can, if sufficient data is available, use linear statistical models to make these diagnostic predictions. The goal of Meehl's review was to compare subjective or clinical predictions with statistical or actuarial predictions to see which approach would produce the more veridical results. To those with great faith in human powers of reasoning, the answer seemed quite apparent. Since people may (and claim to) consider subtle interactions between cues, and statistical models rely upon fairly simple-minded (usually additive) forecasting rules, human diagnosticians should be able to substantially outperform statistical models. In reviewing almost twenty comparative studies, however, Meehl (1954) found no instances in which people outperformed statistical or actuarial models; in every case either the two approaches were essentially equal, or the actuarial approach was better. This result is particularly damaging since the statistical procedures available at that time were fairly crude (in terms of computational capacity), and, as a consequence, the actuarial models used were very simple by present standards.

Subsequent research has not altered Meehl's conclusion that people are not particularly impressive as diagnosticians, at least not in comparison with statistical methods. Oskamp (1965) found, for example, that as the amount of data available to clinical psychologists increased, their confidence in their own assessments increased substantially while the accuracy of their assessments remained relatively constant. Hoffman, Slovic, and Rorer (1968) asked radiologists to assess the likelihood that a gastric tumor was malignant on the basis of seven clinical signs. The median interjudge correlation was only .38. Slovic (1969) obtained a similar result when he asked two experienced stock market analysts, known for their similar philosophies, to assess the growth potential of 128 hypothetical market issues described in terms of attributes such as price/earnings ratios. Here the correlation between the two sets of assessments was only .32. In neither of these experiments was an external validating criterion present. But the low degree of agreement between the supposed experts implies that their average validity will also be low.

Additional studies for which external criteria have been available support this conclusion. Dawes (1971) found that graduate admissions committee evaluations of prospective psychology graduate students correlated only .19 with subsequent faculty evaluations of those students who were admitted. Simple multiple regression procedures, on the other hand, generated scores which correlated .38 with subsequent faculty ratings. Wiggins and Kohen (1971) found that evaluations of graduate school applicants correlated only .33 with the first year grades of those admitted whereas regression models correlated .57 with these grades. And Slovic (1971) reviewed ten studies of investment services and leading stock market analysts and concluded that in every case one would have done better to have disregarded the advice of the experts and simply select a random sample of stock issues.

In the final study to be discussed here, Goldberg (1968) attempted to determine whether the quality of multi-cue predictions could be enhanced by extensive training and feedback. He had experienced clinical psychologists, clinical psychology graduate students, and naive subjects make psychotic vs. neurotic predictions on the basis of MMPI personality inventory scales. Over the course of the experiment subjects made thousands of predictions with immediate feedback on the final diagnosis of each case. As one might hope, the naive subjects improved considerably over the 17-week period. But the experienced clinicians and clinical psychology graduate students showed little improvement. Despite their extensive training, they accurately diagnosed only 65% of the test-case patients as compared to 70% for a simple actuarial model.

Together the real-world studies reviewed in the last two sections reveal that people's abilities as intuitive statisticians are quite modest. Whenever a substantial data base has been available, simple statistical models have outperformed skilled human judges. Thus, one implication of this research seems quite clear: if adequate data are available, use statistical models rather than intuitive human judgment. But what if no large data base is available, or if data are too costly to collect? We shall return to this question shortly.

2.1.8 Descriptive models of human inference - As Rapoport and Wallsten (1972) noted, the emphasis of inference studies has shifted from the question of whether such inferences are optimal--they are not--to the question of why they are suboptimal. Initial attempts to explain this suboptimality generally viewed people as degraded Bayesian processors. For example, the odds likelihood ratio form of Bayes' theorem may be modified by incorporating an exponent α:

$$\Omega_1 = L^{\alpha}\ \Omega_0.$$

This exponent has generally been referred to as an "accuracy ratio" (Phillips & Edwards, 1966) and has served as the dependent variable in many Bayesian studies. In their review of the inference literature, Rapoport and Wallsten (1972) found that this accuracy ratio adjusted form of Bayes' theorem did a good job of explaining the mean data from the most simple probability revision experiments. Nevertheless, they criticized this approach for a number of reasons. First, the actual value of the accuracy ratio is highly task dependent, varying with the type of data generator, the diagnosticity of the data, and the number of data observed. Second, the model does not always provide a good fit to the data of individual subjects. And third, the model provides no insights as to why intuitive inferences deviate from Bayesian optimality.

Wallsten (1971) has used the theory of conjoint measurement to determine whether or not intuitive inferences are even qualitatively consistent with the product rule of Bayes' theorem. His analysis indicated that fully one-third of the subjects in his study used a processing strategy which was not even ordinally consistent with the product rule of Bayes' theorem. In a similar view, Shanteau (1971) used a standard bookbag-and-poker-chip paradigm and had his subjects give their responses on a probability scale. Using functional measurement tests (which are based on the analysis of variance), Shanteau found that his subjects' responses were based on an additive rather than multiplicative combination rule. Pitts, Downing, and Reinhold (1969) also found that subjects employed an additive strategy in a sequential revision task, incrementing their estimate by a constant amount with each new datum. In this study the size of the increment depended much more on the number of data to be observed than on the diagnosticity of the data. Apparently, subjects adjusted the size of their increments so that they would not "run out of room" on the probability scale before the last datum.

Beach, Wise, and Barclay (1970) found that subjects used different strategies when presented with a simultaneous sample of data than when presented with a sequential sample. When presented with a sample of, say five observations, subjects used the proportion of red balls in the sample as the basis for their estimates of the probability of the predominantly red data generator. Such a strategy completely ignores the diagnosticity of the data being processed. When data were presented sequentially, on the other hand, estimates did not depend so heavily on the observed sample proportion.

Tversky and Kahneman have attempted to develop a set of unifying principles which will explain these and other deviations from optimal inference. The basic idea underlying their work is that people invoke simple heuristic strategies which often have little relation to formally optimal models. For example, their "representativeness" hypothesis (Kahneman & Tversky, 1972) asserts that people assess the likelihood of uncertain events by considering the degree to which an event is similar to the main features of its parent population or the process which generated it. The strategy of focusing on the sample proportion of red balls in a binomial revision task provides a good example here. For the sample proportion seems, intuitively at least, to be representative of the parameter of the binomial data generating process. One consequence of the representativeness heuristic is that people tend to ignore sample sizes, for sample size is not a characteristic of the data generator. Kahneman and Tversky (1971) have repeatedly shown that intuitive statistical inferences are grossly insensitive to the effects of sample size on sampling distributions.

A second heuristic principle, termed by Tversky and Kahneman (1971b, 1972) the "availability" hypothesis, asserts that the subjective probability assigned to an event depends upon the number of favorable instances retrieved from memory and the ease of their retrieval. These memory retrieval processes are affected by recency, salience, and imaginability, all of which may or may not be related to the event's past frequency of occurrence. Tversky and Kahneman have obtained extensive support for this hypothesis in simple verbal learning tasks.

"Starting point and adjustment" strategies represent yet another simple heuristic (Slovic & Lichtenstein, 1971). For example, in assessing a probability distribution over the future price of some commodity, a market analyst might take the present price, increment it by a fixed percentage, then get a rough fix on the distribution's spread. Reliance on simple computational rules of this type can lead to serious biases when the rule includes implicit assumptions about the occurrence of events which are themselves only probabilistically determined (Stael von Holstein, 1972).

Although our understanding of intuitive inference processes is far from complete, it is by now clear that people make extensive use of heuristic strategies which bear little resemblance to optimal statistical strategies. In later sections we shall discuss a number of decision-aiding procedures which have been devised to augment limited human inferential capacities and to eliminate some of the biases inherent in intuitive inference.

2.2 Suboptimal Decision Behavior

2.2.1 *Simple gambling behavior* - Choices between simple gambles, usually involving only two possible outcomes, provide the prototypical setting for studying decision making under risk. Although a number of other strategies have been discussed, the *expected utility principle* is now widely accepted as the appropriate normative standard for decision making under risk (Luce & Raiffa, 1957; de Groot, 1970). Notationally, let A_1, A_2, ... A_r denote a finite set of actions which are available to the decision maker and let X_1, X_2, ... X_s denote a mutually exclusive and exhaustive set of consequences which might arise from these actions. Finally, let $(p_{1i}X_1, p_{2i}X_2, \ldots p_{si}X_s)$ denote the probability distribution of consequences associated with action A_i, where, for example, p_{2i} denotes the probability that consequence X_2 will occur given that act A_i is selected. When preferences satisfy certain normatively appealing properties, it can be shown (Luce & Raiffa, 1957) that there exists an interval scale *utility function* U such that:

a) $X_i \overset{\cdot}{>} X_j$ if and only if $U(X_i) \geqq U(X_j)$

b) $A_i \overset{\cdot}{>} A_j$ if and only if $EU(A_i) \geqq EU(A_j)$.

Here $X_i \overset{\cdot}{>} X_j$ denotes that X_j is not preferred to X_i, and $EU(A_i)$ denotes the *expected utility* of act A_i where

$$EU(A_i) = p_{1i}U(X_1) + p_{2i}U(X_2) + \ldots + p_{si}\, U(X_s).$$

The expected utility principle, like Bayes' theorem, provides a benchmark against which to compare human behavior.

A number of studies provide strong support for the contention that people evaluate simple gambles in an expected utility maximizing fashion. Tversky (1967) asked prison inmates to state minimum selling prices for simple gambles whose consequences consisted of cigarettes, money, and candy. Here, expected monetary value provided a good approximation to the subjects' bids, and expected utility models a near perfect fit. Goodman, Saltzman, Krantz, and Edwards (1973) conducted a study in a Las Vegas casino. Here considerable sums of money were at stake. Their primary finding was that expected monetary value came so close to predicting what subjects did that it was not worthwhile to consider more sophisticated models.

Both of the above studies inferred preferences from bids. Lichtenstein and Slovic (1971) compared the preference orderings inferred from bids with those obtained by asking subjects to choose between pairs of gambles. They found systematic discrepancies between the two orderings, a result which is inconsistent with the expected utility principle. When stating selling prices subjects focused more on the amount to win, but when choosing between gambles, they placed more weight on the probability of winning. Thus, even in evaluating simple gambles people sometimes adopt strategies which violate the principles of rational choice.

More recently, Slovic, et al (1977) have experimentally investigated insurance purchasing behavior in the presence of small probabilities of very large financial losses. They found that people are more likely to purchase insurance against a moderate probability of a small loss than against a small probability of a large loss. This finding seriously violates the expected utility model's usual predictions about insurance purchases, and suggests that the expected utility model fails to explain behavior in the face of low probability catastrophic events. This interpretation is consistent with the findings of Kunreuther's (1976) field study of earthquake and flood insurance purchases.

2.2.2 Complex decisions - In principle, complex decisions could also be evaluated by seeing whether they conform to the expected utility principle. In practice, however, they have been compared with models which maximize expected monetary value (EMV) using a probability model based on the objective parameters of the environment. This approach can be rationalized on two grounds. First, in those studies in which subjects' pay depends upon the quality of their decisions, the actual money payoffs involved are typically quite small. Other research has shown that utility functions for money are essentially linear for small cash values (Tversky, 1967), so it is not unreasonable to assume that subjects do in fact attempt to maximize EMV. Other studies in which subjects are asked to make hypothetical choices with large monetary outcomes are usually presented as business scenarios. Since EMV maximization is commonly used as a goal in business and industry, it does not seem unreasonable to evaluate subjects' performance against this standard.

In the simplest paradigm which we will consider here, Pitz and Reinhold (1968) had subjects observe five data sampled from one of two binomial processes. At the end the the data sequence, subjects selected one of two binomial processes. If correct, they received a positive reward; if incorrect, they had to pay a penalty. Consider the following two payoff matrices.

True Hypothesis

		H_1	H_2
Subject says	H_1	a	-a
	H_2	-a	a

	H_1	H_2
H_1	a	-2a
H_2	-a	2a

In the symmetric payoff matrix on the left, the optimal strategy is very simple. Let $p_1 = p(H_1)$ and $p_2 = p(H_2)$. One should predict H_1 if and only if EMV $(H_1) \geqq$ EMV (H_2). It can easily be shown that this will be the case if and only if $p_1 \geqq p_2$. Thus, one selects H_1 only when $p_1 \geqq p_2$. Since the data-generating processes used by Pitz and Reinhold were symmetric about .5, subjects could maximize EMV simply by selecting the color which occurred most often in the observed sample. In this simple case, most of the subjects' choices were in fact consistent with the EMV maximization strategy.

Consider now the asymmetric payoff matrix on the right. Here too, the subject should predict H_1 if and only if EMV $(H_1) \geqq$ EMV (H_2). It is easy to show that in this case the subject should select H_1 if and only if $p_1/p_2 \geq 2$. In general, asymmetric payoff matrices will lead to asymmetric cutoff odds like these. Pitz and Reinhold's results indicated that the subjects did shift their cutoff odds away from 1:1; but not nearly far enough. Thus, even in this extremely simple task, subjects made many suboptimal responses.

2.2.3 Optimal stopping tasks involve a somewhat more complex paradigm. Here, too, data are sampled from one of two randomly selected data-generating processes. In this case, however, the subject must pay a fixed cost for each datum he observes, and may decide at any point to stop sampling. At this point he makes his decision. The optimal strategy for this problem can be simply stated. One should continue sampling until the EMV of further sampling is less than that of making a decision based on the data already available. The mathematics of determining the EMV of continued sampling are quite complex (de Groot, 1970), but when the decision maker may purchase an unlimited number of observations, it can be shown that there exist critical odds cutoffs $\Omega_1{}^*$ and $\Omega_2{}^*$ such that the decision maker should continue sampling until either

$$\frac{P_1}{P_2} \geq \Omega_1{}^*$$

or

$$\frac{P_1}{P_2} \leq \Omega_2{}^*$$

(Edwards, 1965). Given the difficulty of this problem, it may seem unreasonable to expect that intuitive decisions will approximate those dictated by the optimal strategy. Optional stopping tasks are quite insensitive to errors in the selection of cutoff odds (Rapoport & Wallsten, 1972), however, so if subjects use "reasonable" strategies they should do very well relative to the optimal strategy.

Rapoport and Wallsten (1972) provide an overview of the optional stopping literature. Fried and Peterson (1969), for example, conducted an optional stopping experiment in which subjects behaved quite suboptimally. Across the various conditions of their experiment, subjects' actual earnings ranged from 10% to 72% of the earnings they would have received had they employed the optimal strategy. In general, subjects deviated from the optimal strategy by buying too little information. This is not an uncommon result.

Pitz, Reinhold, and Geller (1969), to cite another example, found that subjects generally relaxed their cutoff odds as the number of observations previously purchased increased. One heuristic strategy which will produce such an effect has been termed the World Series strategy. A subject might decide to select the first hypothesis favored by four data. This is equivalent to using a "best out of seven" criterion, but stopping as soon as one hypothesis is clearly the victor. This strategy can only be applied to binomial tasks. Pitz, Geller, and Reinhold compared the implications of the World Series model with those of the fixed-cutoff and fixed-sample-size models. Their test was very general, for it allowed for the possibility that the subject might shift his odds or sample size cutoffs from trial to trial. Although none of the models provided a perfect fit to the data, the World Series model was strongly favored. So again we find evidence suggestive that subjects simplify complex tasks by adopting intuitively appealing but suboptimal heuristics.

Wendt (1969) studied information purchase using a different experimental paradigm. Instead of having subjects make sequential purchases at a fixed cost, he allowed them to bid for the purchase of a simple datum. He then compared the subjects' bids with the actual EMV of the data. The optimal bids varied as a function of the prior odds for two hypotheses, the diagnosticity of the data, and the payoffs for the subsequent decision. The subjects' bids were generally ordinally consistent with the EMV maximization model. But subjects had an extremely suboptimal tendency to make substantial bids for data with little diagnostic value. In some conditions the subjects overbid by 200-300% for low diagnosticity data. Shanteau and Anderson (1971) replicated Wendt's findings. Their analysis showed that subjects' bids were a simple multiplicative function of the priors, diagnosticity, and payoffs. This strategy is generally ordinally consistent with the optimal strategy, but leads to positive bids for data which have *no value*, a prediction which was supported by the data. Here, too, subjects consider the appropriate variables, but combine them in a suboptimal fashion.

In a quite realistic information-purchase experiment, Moskowitz (1972b) asked experimental aerospace R & D managers to make hypothetical decisions about the development strategy of a laser system. They were asked to consider two competing designs and were allowed the option of developing both systems in parallel. They also had the option of buying experiments which could help them choose between the two designs. Despite their extensive experience, very few of the managers adopted the optimal alternative. They spent far too much money on relatively uninformative experiments. On the average, the EMV of the strategies adopted by the subjects was only 74% of that for the optimal strategy.

Search problems define another research paradigm with close links to many real world settings. Formally, a target may be located in any one of *r* locations, and the subject has a prior probability associated with each of these locations. He may then search locations one at a time. The search procedure is not perfectly reliable, however, so that one may search the correct location but fail to detect the target. In addition, each search has an associated cost, which may vary across locations. The optimal strategy, at each stage of the process, is to search the location with the highest P(Detect)/Cost ratio. Throughout the process, an optimal subject would use Bayes' theorem to revise his location probabilities following each unsuccessful search. Rapoport (1969) found that subjects generally do not search the best location. This result was replicated by Rapoport, Lissitz, and McAllister (1970). When the search cost varies across locations, subjects tend to place too much weight on probability of detection and not enough on the cost of search. Despite the fact that they seldom choose the optimal location, they do quite well at this task, seldom incurring a cost more than 25% greater than that of the optimal strategy. This result is due in part to the fact that search tasks are relatively insensitive to error. But it probably also reflects the fact that subjects make reasonably good inferences about the target's locations.

Decision making becomes even more difficult when the decision maker's actions affect the state of his environment. Optimal policies for such problems can generally be obtained only through the application of the mathematical technique known as dynamic programming. Rapoport (1967) provides a general discussion of dynamic decision-making tasks and dynamic programming models. He also illustrates the approach with a simulated stock market problem. Subjects were confronted with the task of buying or selling stock with unknown demand. The demand distribution for future time periods varied as a linear function of the subjects' present decisions. The subjects' performance was compared with that of two EMV maximizing models. The first model used a uniform distribution over demand throughout all stages of the problem. The second model began with a uniform demand distribution, but revised that over the course of a scenario using Bayes' theorem. The subjects' performance on the task was quite good, falling between that of the optimal adaptive model and that of the optimal non-adaptive model.

Inventory problems represent another dynamic decision task. Here, too, the decision maker is faced with uncertain demand. If he buys too much inventory, he will have to pay storage costs; if he buys too little, he pays a penalty cost for failing to meet demand. Because unused inventories accrue over time, the task is dynamic. If the problem has a fixed number of stages, the decision maker should attempt to attain a zero stock level by the end of the last stage. Because this task is representative of complex real-world tasks, it is of considerable interest to see how well subjects do on it. Rapoport and Calder (1970) found that college students fare poorly. Mean earnings in the various conditions of their experiment ranged from 53% to 61% of those for the optimal strategy. Subjects generally did not respond at all to the fact that they were coming to the last stage of a problem, thus ending with excessive inventory levels. They responded appropriately to storage costs, but generally ignored the costs associated with failing to meet demand.

Moskowitz (1972a) obtained similar results using graduate students in industrial administration who were familiar with inventory problems. His task was even more complex, with subjects being asked to set both work force and production levels. Over all conditions, costs incurred by subjects were 57.4% greater than those for the optimal model. This difference ranged from 26.5% when demand estimates were available only one stage into the future, to 94%, when demand estimates were available three stages into the future.

Miller, Kaplan, and Edwards (1967, 1969) conducted probably the most complex decision-making simulation experiments. Their subjects, USAF pilots and ROTC students, were asked to allocate tactical fighter strikes on the basis of mission requests. In making such decisions, one must consider the value of the target, the probability of destroying the target (which varies with the number of aircraft assigned to it), the probability distribution of the values for future mission requests, and the number of aircraft remaining. Scarcity entered the problem because only a fixed number of strikes could be made in each time period. Subjects in one condition of these experiments assigned values to each target request as it came in, then decided how many (if any) aircraft to allocate to it. Subjects in a second condition merely assigned values to targets. A complex mathematical programming model made the actual allocation decisions using the maximization of subjective expected utility as a criterion. The performance of the two systems was compared with the decisions which would have been made by an optimization model which knew beforehand what request would be received in each time period. This model provided a "perfect hindsight" standard of performance. The results of these experiments generally indicated that the intuitive judgment attained 40% to 50% the score of the perfect hindsight model, whereas the dynamic programming model attained 85% to 90%.

The implications of these studies are rather clear. As decision tasks become increasingly complex, the quality of intuitive decisions declines steadily when compared with the decisions generated by formal optimization models. This result provides strong support for those who have argued that people are severely limited information processors. It also suggests that formal decision theory might be applied to overcome these limitations. The next section of this paper discusses the decision-aiding technologies that have been devised to accomplish this task.

3.0 DECISION-AIDING TECHNOLOGIES

3.1 The Decomposition Approach--or, Divide and Conquer

The basic theme of the decision-aiding technologies we will discuss is quite simple--divide and conquer. The experimental studies discussed in the previous section generally support the contention that people are better at making simple judgments than they are at aggregating large amounts of information to form an overall decision. In the context of probabilistic inference, for example, people are much better at assessing the diagnostic value of a single datum than they are at aggregating information across a set of data (Edwards, 1968; Slovic and Lichtenstein, 1972; Wheeler, 1972). In the context of decision making, human evaluations of simple gambles are much more nearly optimal than are human decisions in complex tasks. These findings have led to the conclusion that complex decision-making tasks should be decomposed into a set of component subtasks, each of which is well within the judgmental capacities of the decision makers involved. In such a system, people make judgments about values and probabilities, and formal models aggregate the implications of these judgments to arrive at a recommended decision. Usually, but not always, the formal aggregation will be accomplished by a computer.

The decomposition approach which we will consider involves six major tasks:

1. Recognizing that a decision problem exists.
2. Identifying the possible courses of action.
3. Constructing a probabilistic model of the decision-making environment.
4. Constructing a model to evaluate the possible consequences of each available action.

5. Selecting a course of action.
6. Implementing the alternative selected.

Our discussion focuses on the third, fourth, and fifth tasks. We are still, unfortunately, ignorant of the creative processes involved in the first two steps. Although progress has been made (Simon and Newell, 1972), we do not yet have anything resembling an adequate normative theory of creative problem solving. For the time being, at least, we must rely upon the creative abilities of those who make decisions. The final task, decision implementation, is also important. We do not consider here the organizational processes which intervene between a stated decision and the actual behavior produced by the organization. When the correspondence between intended behavior and actual behavior is low, of course, the benefits of a careful decision analysis may well be lost. Allison's (1971) discussion of the Cuban missile crisis provides a fascinating example of the discrepancy between decision makers' intentions and the actions of their subordinates.

3.2 Probabilistic Information Processing Systems--The Technology of Inference

3.2.1 Simple PIP System - Early in the 1960's Edwards proposed that probabilistic inference tasks be decomposed into four major subtasks (Edwards, Lindman, & Phillips, 1965). In this formulation people identified relevant states of the environment and information sources which could discriminate between these states. They also estimated likelihood ratios linking individual data with hypotheses about the environment. The tasks of aggregating information across data was assigned to Bayes theorem. These inference systems came to be called PIP--an acronym for Probabilistic Information Processing. Implicit in the original formulation of PIP was the assumption that the environment could be described by a stationary, single-stage inference model in which all data were conditionally independent with respect to the hypotheses of interest. When this assumption is satisfied, data may be considered one at a time; people can assess single-datum likelihood ratios; and the simple odds-likelihood ratio form of Bayes theorem can be used to update the posterior odds distribution.

In one early attempt to validate the PIP approach, Kaplan and Newman (1966) asked subjects to make inferences about the probable target of a bombing raid, given data about the points of impact of the bombs dropped. In their simulation, actual points of impact were circularly normally distributed around the intended target. They found that PIP was consistently better than unaided inference, generally assigning a much higher (.2-.3) probability to the true target.

In a more ambitious simulation study, Edwards, Phillips, Hays, and Goodman (1968) had subjects make inferences about the strategies of nations in an imaginary version of a future world. The data with which subjects worked were designed to be good facsimiles of the reports with which real-world intelligence analysts work. As expected, PIP and intuitive inference generally favored the same hypothesis, but PIP assigned much higher posterior odds to this hypothesis. Because there was no "true" data generator in this study, however, it was impossible to determine whether the PIP-assigned posterior odds were in fact more optimal than those produced by intuitive inference.

Wheeler's (1972) thesis data provide by far the strongest support for the PIP concept. Her subjects made inferences about which of two normal distributions had generated a given set of data. As in most simple revision experiments, intuitive inferences over a set of data were very conservative. PIP-generated odds, on the other hand, were nearly optimal over a wide range of data diagnosticity. These results held for individual as well as group data.

Edwards and Seaver (1976) describe an experiment undertaken to determine whether the use of judgmentally averaged log likelihood ratios would contribute significant improvements over the likelihood ratio judgment originally proposed for use in PIP systems. Included as a variable was the diagnosticity of the data used to elicit subjects' responses. It was found that data diagnosticity affected quality of response for both response modes. Estimates became more diagnostic. The primary finding of the study was that quality of estimates did not differ significantly in either verdicality of orderliness between likelihood ratio estimates as originally proposed for the PIP technique and the averaged log likelihood estimates. Both methods were found to produce better estimates than cumulative certainty judgment, as is usual in such comparisons.

The reason for considering an alternative to likelihood ratio judgments is that a problem may arise in applying PIP systems in real-world contexts. The people assessing the likelihood ratios will typically have access to feedback about the posterior odds that are calculated from their likelihood ratios. Goodman (1973), in a reanalysis of data from five studies exploring methods of eliciting judgments about uncertain events, concludes that feedback about the implications of judgments makes them less extreme and is probably the most powerful variable controlling the extremeness of the judgments. Thus, even a PIP system may be susceptible to conservatism in real-world applications. This problem seems less likely to characterize judgments of average certainty due to the very nature of the elicited judgments. Should further research confirm feedback produces conservatism in PIP systems, average certainty judgments may prove to be a useful alternative to PIP.

3.2.2 PIP in complex environments. One natural question to ask before building greater complexity into a system is--does it matter? Lichtenstein (1972), for example, developed actuarial Bayesian models to predict psychosis vs. neurosis using MMPI profiles. One model considered the conditional dependencies between the various MMPI scales whereas the other model ignored them. Lichtenstein found no difference between the models inability to diagnose the correct hypothesis. Her measure of effectiveness, however, was simply the number of times a model assigned a probability greater than .5 to the correct hypothesis. This is a rather insensitive measure for it ignores the magnitudes of the probabilities involved.

Domas and Peterson (1972) studied the effects of data redundancy on inference. In a control condition in which data were conditionally independent, PIP did outperform simple intuition. But in the case of redundant data, intuitive judgments were more optimal than PIP, which assigned excessive probabilities. Simple intuition also outperformed a conditional likelihood ratio form of PIP in which subjects estimated quantities of the form $P(d_1,d_2,\ldots d_n|H_1)/P(d_1,d_2,\ldots d_n|H_2)$. In principle at least, this modified version of PIP should have been able to cope with the nonindependence involved. These results show that ignoring the existence of conditional nonindependence can be considerably more misleading than Lichtenstein's results suggest. They also show that simple intuition may outperform an inappropriate decomposition.[1]

[1] Snapper and Peterson (personal communication) obtained a similar result in a study of real-world weather forecasting, an environment in which data redundancy is very substantial.

Schum, Southard, and Wombolt (1969) took a slightly different approach to the conditional nonindependence problem. In their system, termed semi-PIP, data are first sorted into bundles such that no conditional dependencies exist between data in different bundles. Data within the same bundle, however, are dependent. Any datum which is independent of all other data will define a bundle of size one. In semi-PIP men first group data into bundles, then assess likelihood ratios for each bundle. Because bundles may include many data, men must absorb a considerable portion of the aggregation burden. The success of semi-PIP also depends on the ability of the system operators to make likelihood ratio judgments which appropriately reflect the nonindependencies involved. Schum, Southard, and Wombolt (1969) conducted a series of experiments in which they compared semi-PIP with simple inference. With sets of six data, simple intuition was as veridical as semi-PIP. But as the number of data to be aggregated increased, semi-PIP was substantially better.

Although these results are encouraging, they do not provide a complete test of semi-PIP. First, only two-way dependencies were studied. People might have considerably greater difficulty dealing with higher order dependencies between data. Second, the system operators were alerted as to the possible sources of nonindependence. Thus, these studies bypassed the difficult problem of sorting data into conditionally independent bundles. The success of semi-PIP in the real world will depend heavily on the ability of system operators to identify nonindependence and to sort data appropriately. The identification and sorting problems will be particularly difficult for systems which process information over an extended period of time. An intelligence report received today, for example, might interact with reports received two weeks or even two months ago.

When the amount of information to be processed is great, system operators will be confronted with serious overload problems.

The study of Schum, Southard, and Wombolt (1969) was a part of a long series of studies done at Ohio State University under Air Force sponsorship. Howell (1967) summarized the program and outlined a set of thirteen principles for the design of command and control decision aiding systems based on its results. These principles generally endorse the PIP idea, affirm the interaction between the desirability of PIP and issues of data diagnosticity, and raise such issues as data reliability, feedback concerning system opinion, training, and evaluation of system output. Ten years later, they still look like sound conclusions, though of course much more detailed and sophisticated knowledge is available now than was available then.

PIP systems have also been devised to cope with the problem of multistage inference. A general treatment of multistage inference is provided by Kelly (1972). He provides general mathematical models for multistage inference which can be directly translated into computer algorithms. From a practical standpoint, however, Kelly's models are tractable only when the problem can be structured to eliminate most conditional dependencies between data. When this cannot be done, system operators will be required to make a prohibitively large number of judgments.

Experimental evidence on multistage PIP systems in unfortunately scant. Gettys, Kelly, Peterson, Michel, and Steiger (1973) have conducted two relevant studies. In their first study subjects made inferences about a person's college major based on his height and a hypothetical test score which supposedly discriminated between men and women.

Sex was an intervening variable, linking height and test score with college major. Subjects' intuitive inferences in this task were very good, though slightly conservative. Multistage PIP was better, generating odds which were almost identical with the Bayesian odds. Nevertheless, the difference between multistage PIP and the intuitive inferences was small. The quality of the intuitive inferences left little room for improvement.

In their second experiment, Gettys, et al. used a scenario with a multinomial intervening variable. Here, there were cases where the most likely hypothesis was not favored by the most likely state of the intervening variable. Use of an as-if or best-guess heuristic strategy can lead to very suboptimal inferences with a task of this type. As expected, intuitive inferences were substantially excessive, whereas multistage PIP was close to optimal. Together these studies suggest that multistage PIP systems can be effectively implemented and will, in some cases, yield substantial improvement over simple intuitive inference. More research is required, however, to firmly establish this conclusion.

3.2.3 Pooling the assessments of a group of assessors - One obvious way to overcome the limitations of individual inference is to form a panel of experts. Because different experts will have had access to different information about the question at hand, a consensus probability distribution based on all of this information should be more veridical than the distribution of an expert randomly selected from the panel. In addition, with a panel of experts one can exploit the statistical properties of the averaging process. The benefits of averaging are strikingly illustrated by a simulation experiment conducted by Huber and Delbecq (1972). They considered the problem of point estimation on a theoretically continuous scale. In one example they assumed that each expert's estimate was sampled from a normal distribution with mean equal to the true parameter value and with a standard deviation equal to either 5% or 10% of the scale range. For the 10% standard deviation case, their analysis indicated that the expected absolute error for one randomly selected judge was equal to 7.5% of the scale range. For five randomly selected judges, however, the expected error was only 3.4%, and for ten judges only 2.5%.

Dalkey (1969) has reported results with real subjects which support the Huber and Delbecq analysis. He asked subjects to make point estimates for such quantities as the U.S. gross national product in 1970. As a measure of the error of a group or individual estimate, Dalkey used

$$E = \ln \left| \frac{\text{Estimate}}{\text{True}} \right|.$$

He compared the average error of individual estimates with the average error of groups ranging in size from 2 to 29, the number of subjects in the study. Averaging over groups of five reduced the error score by 42%; averaging over all 29 subjects reduced it 65%. This reflects the decreasing marginal value of additional predictors, a result which we also noted in the Huber and Delbecq analysis. The greatest reduction of error was obtained by going from one to five judges. After ten judges, the reduction in error was fairly negligible.

Winkler (1968) has shown that averaging can be applied to probability distributions as well as to point estimates. Suppose, for example, that we want a probability distribution over the parameter Θ and that a set of experts has assessed the distributions $f_1, f_2, \ldots, f_n$, each of which satisfies all of the normal laws of probability theory. Winkler shows that $f(\Theta) = \sum_{i=1}^{n} w_i f_i(\Theta)$ also satisfies the properties of a probability distribution provided that

$\sum_{i=1}^{n} w_i = 1$, for $o \leq w_i \geq 1$. Here the w_i may be interpreted as weighting factors. In the simple averaging case, of course, $w_i = \frac{1}{n}$.

Alternatively, if all assessors agree on a particular family for the probability distribution, then estimates of the parameter values themselves can be averaged. In a waiting time problem, for example, experts might agree on a Poisson process. Estimates of average time between events could then be averaged to obtain the value of the distribution's parameter. Direct averaging of parameter values raises a number of technical problems, however, which limit the usefulness of this approach. (For a more complete discussion, see Winkler, 1968.)

Virtually all studies of synthesizing the opinions of experts have considered only the benefits to be obtained from averaging together the judgments of individuals. This is at least in part due to the fact that it is much easier to collect judgments from isolated judges and average them than it is to get a group of experts together and have them talk their way to agreement. This practical consideration arises in real-world as well as in experimental environments. In addition, the use of nominal (statistical) as opposed to interacting groups reflects the common belief that direct interactions can result in biases, such as the emergence of dominant figures, which will result in a poorer consensus distribution (Gustafson, Shukla, Delbecq, and Walster, 1973).

Studies of simple averaging generally show that this approach can lead to considerably more veridical inferences. Winkler (1971), for example, conducted a long-term study in which subject assigned probabilities to various point spreads in Big Ten and National Football League games. Winkler evaluated these estimates using quadratic (QSR) and logarithmic scoring rules (LSR). Because these, and all proper scoring rules, are convex, one can show that any average distribution did much better than this, outperforming 95% of the subjects in the study. Using the QSR, the average distribution resulted in a 5% to 10% improvement; with the more sensitive LSR, a 26% to 28% improvement in score was obtained. Winkler also considered various differential weighting schemes in which the weight assigned to a given subject depended either on his own rating of his competence, or upon his score in previous sessions. The scores produced by these differential weighting schemes were but negligibly different from those for the equal weighting model.

Staël v. Holstein (1971) also examined the effects of averaging in his weather forecasting study. Recall that only 7 of the 30 subjects in this study were able to outperform the unconditional climatological probabilities. The average distribution, however, did beat the climatological odds. In addition, the average distribution for the university meteorologists outperformed 28 of the 29 subjects in the study, including both of the official government forecasters.

Staël v. Holstein (1972) obtained similar results in his stock market study. Here, too, the average subject did worse than the simple unconditional historical frequencies. But the average model for the stock market experts outperformed all 72 subjects and was slightly better than the simple historical frequency model. In terms of QST, the average model outperformed the average subject by about 6%.

Next we consider three studies which compared statistical averaging with various modes of direct and indirect interaction. Moskowitz (1971) compared nominal groups with real groups which had to come to a consensus in a Bayesian estimation task. Subjects were asked to evaluate credit risks on the basis of three independent data sources of known diagnosticity. Moskowitz used inferred accuracy ratios as a measure of veridicality. Thus, an accuracy ratio of 1.0 defined optimal performance. In general, nominal groups were substantially more verdical than groups with real interaction; accuracy ratios for the two types of groups were .90 and .63, respectively. The advantage of the nominal groups was most marked for data of high diagnosticity. For low diagnosticity data, in fact, the fact-to-face groups performed better. To summarize these results using the traditional Bayesian terminology, Moskowitz found that direct interaction resulted in substantially greater conservatism than did statistical averaging of the judgments of isolated individuals.

Gustafson, Shukla, Delecq, and Walster (1973) compared the benefits of simple averaging with several types of interaction. Subjects in this experiment were asked to make inferences about a person's sex based on information about the person's height and weight. In the simple Estimate condition, subjects made their judgments as individuals without any interaction with other subjects. In the Talk-Estimate condition, the members of a group talked over each item, then make individual judgments. In the Estimate-Talk-Estimate condition, subjects first made individual judgments, then talked these over, then re-estimated individual judgments. In the Estimate-Feedback-Estimate condition, subjects again began with individual estimates, then re-ceived ananymous feedback about the estimates of the other subjects in their group, then made new estimates. This condition tested the value of the popular Delphi technique (Dalkey & Helmer, 1963). All groups in the last three conditions were comprised of four subjects each.

To provide a control condition for the pure effects of averaging, random groups of four subjects from the Estimate condition were constructed, and the judgments of the members of these "groups" were then averaged together. The error score used to analyze these judgments was percentage deviation from the Bayesian posterior odds. Using this criterion the Estimate-Talk-Estimate groups were clearly the most verdical. The Estimate and Talk-Estimate groups were roughly equal, and were both about 45% worse than the Esimate-Talk-Estimate groups. The Delphi groups came a distant last, with an error score 63% higher than that of the Estimate-Talk-Estimate groups. This study strongly suggests that a combination of averaging processes and face-to-face discussion provides the best mode of aggregating expert opinion. But suprisingly, the discussion process is of little value unless the discussants have previously committed themselves to a prior estimate. The study also suggests that the popular Delphi technique is a very poor way of aggregating opinion.

In an effort to compare various behavioral and mathematical techniques of group probability assessment, Seaver (1977) experimentally compared two aggregation rules, weighted arithmetic means and

and weighted geometric means, and three weighting procedures, equal weights, weights based on self-rating, and DeGroot weights (DeGroot, 1974). Five behavioral interaction techniques were compared, the Delphi method (Dalkey and Helmer, 1963), the Nominal Group technique, developed by Delbecq and Van de Ven (1971), a modified nominal group technique in which group members state their estimates and reasons with no discussion, a concensus technique in which groups were to arrive at concensus in any way they wished, and no interaction or control group in which group members made estimates with no knowledge of other member's estimates.

The quadratic scoring rule was used as the criterion for measuring the quality of group assessments. The well-known insensitivity of that rule may account for the lack of significant differences among behavioral techniques. In general, interaction among group members reduced differences, reduced the calibration of the judgments, and increased the extremeness of judgments. Therefore, deciding whether or not to use group interaction techniques involves a tradeoff between calibration and extremeness of the responses. Although no significant differences were found, slight differences as well as the results of other studies point to slight superiority of the nominal group technique to other group interaction methods.

The data show that little if anything is lost by using mathematical techniques to aggregate individual judgments rather than behavioral interaction. Considering the practical disadvantages of face-to-face meetings of groups, this research suggests that there may be no point in bothering with the sometimes lengthy procedures of behavioral interaction. While results of this experiment dealt totally with point estimates, further studies focused on eliciting continuous distributions are needed.

3.3 Multi-attribute Utility Theory--A Technology for Evaluating Complex Outcomes

In many contexts decision makers may use the output of a PIP system as an input to their decisions, but wish to make the decisions themselves in an intuitive rather than analytic fashion. The full power of the decision-theoretic approach cannot be realized, however, unless alternatives are selected using expected utility maximization as a criterion. This, of course, requires that each possible consequence of the action alternatives being considered be assigned a single numerical value which reflects the utility of that consequence relative to all other possible consequences. Until quite recently it was commonly believed that utility assessment was feasible only when the outcomes in question varied only with respect to a single value-relevant dimension, such as dollar profits. Now, however, we have both the theory and practical methods for assigning utilities to complex outcomes which vary along multiple value-relevant dimensions. This section briefly discusses both the theory and practical scaling techniques. For recent reviews or discussions of the multi-attribute utility approach, see Edwards (1976, 1971), Fischer (1972a), Raiffa (1969), and von Winterfeldt and Fischer (1972). For a detailed book-length treatment, the reader is referred to Keeney and Raiffa's book, Decisions with Multiple Objectives, (1976).

In our discussion of the multi-attribute utility theory (MAU) approach, we will use the following hypothetical example. This example is both oversimplified and uninformed; its purpose is only to illustrate the method of analysis. Suppose that deployment of an anti-ballistic missile system within the continental United States is being considered. The Defense Department has responsibility for deciding upon appropriate ABM sites. The problem is a complicated one, for various reasons. For one thing, ABM systems become available a few at a time, not all at once. For another thing, the Defense Department must balance protection of retaliatory capability, represented by strategic forces; protection of defensive capability, and protection of population, industry, and agriculture.

Utility theory is attractive as an approach to this problem. In particular, it copes with the gradual availability of ABM systems by ranking possible locations; as more systems become available, they can go to progressively lower-ranking sites. (Some technical problems about synergistic effects arise here, but can be ignored for the purpose of this expository example.)

The first task in any utility theory analysis is that of determining whose utilities are to be maximized. In this case there is clearly no single decision maker whose interests are paramount. Ideally, the utility measure should reflect the interests of the American people as a whole. In reality, however, the actual decision will be made by a small number of high-level officials. These officials collectively constitute the decision maker in the analysis. Of course they will not fully agree with one another; that is one of the technicalities we must face.

Next we must consider the question of utility for what. Suppose that in our example the decision makers agree on a vaguely stated overall objective: the protection of the lives and well being of the American people against a strategic nuclear attack. Next, the decision makers must list a set of value-relevant attributes or criteria that bear upon this overall objective. This list of attributes may be obtained in either of two ways. First, the decision makers might simply draw up a list of considerations generally agreed to be important. Or second, the evaluation problem could be attacked hierarchically. The first approach is straightforward. The second merits further discussion.

Suppose that after careful deliberation, the decision-making group decides that two major factors contribute to the defensive posture of the nation: (a) maintenance of a credible deterrent threat, and (b) direct and indirect defense if deterrence fails. Each of these factors can in turn be decomposed into a set of more specific attributes. Assume, for example, that the decision makers decide that in assessing the degree to which a given defensive posture contributes to the "defense if deterrence fails" criterion, three specific criteria are of paramount importance: (a) the percentage of the nation's population that would survive in an all out attack, (b) the percentage of the nation's agricultural capacity that would survive such an attack, and (c) the percentage of the nation's industrial capacity that would survive such an attack. Finally, suppose that the decision makers choose percentage of retaliatory forces surviving an all out first strike as an acceptable measure of the deterrence criterion. These relations can be represented by the hierarchical structure in Figure 3. Equivalently, the decision-making group might simply have drawn up a list of the four specific attributes at the bottom of this value hierarchy. It is obvious that the percentages of retaliatory forces, population, and industrial capacity surviving depend on the sites selected. Agricultural capacity is also influenced, however; for example, the location of ABM sites in remote areas where our own ballistic missiles are located might increase population losses but decrease agricultural losses.

Next, each of the possible action alternatives must be evaluated in terms of the specific attributes at the bottom of the value hierarchy. In the present example, these

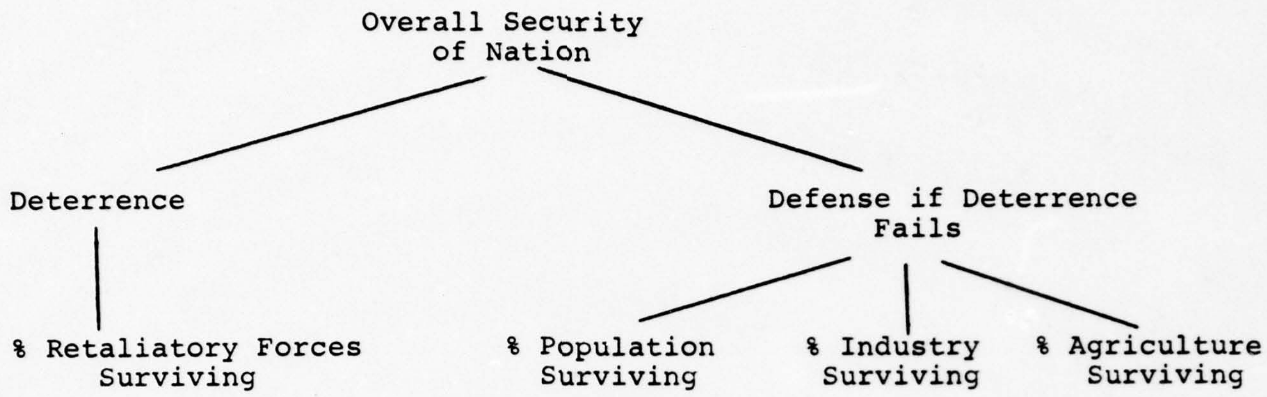

Overall Security of Nation
Deterrence
Defense if Deterrence Fails
% Retaliatory Forces Surviving
% Population Surviving
% Industry Surviving
% Agriculture Surviving

FIGURE 3

VALUE HIERARCHY FOR NATIONAL DEFENSE

alternatives would be the set of ABM sites being evaluated. In assessing the degree to which each site contributed to each of the four attributes, the decision-making group would have to rely upon the judgment of experts. For each set of sites a probability distribution would be assessed over each of the four value attributes. These probability distributions might be generated by simulation models incorporating information concerning the capabilities and intentions of potential aggressors and the characteristics of the ABM's. Or they might be based purely on expert intuition, perhaps formalized as a multistage inference model.

With this information in hand we now turn to the final, and perhaps most difficult, stage in the evaluation process, namely, trading off attainment of one objective against the attainment of others. For only rarely is it the case that one alternative is clearly better than all others with respect to every value-relevant attribute. Multiattribute utility theory (MAU) provides the formal basis for making such tradeoffs. In our discussion of MAU we distinguish between risky and riskless decisions, because the logical underlying utility decomposition is somewhat different in these two cases.

3.3.1 Riskless decision making. A decision is said to be riskless if the decision maker is able to specify with complete certainty the consequences associated with each possible course of action. Under these circumstances decision making requires only that the set of consequences be rank-ordered in terms of their desirability, and then that the alternative associated with the most desirable consequence be selected. The riskless choice assumption is clearly an idealization; all decisions involve uncertainty at some level. Some situations may approximate the certainty assumption, however. Also, there are circumstances in which it may be useful to treat an uncertain decision as if it were riskless. Here one would specify one value for each attribute of an alternative; the expected mean or median outcome with respect to a given dimension are the most likely candidates. This practice of treating an uncertain decision as if it were riskless may be justified in some cases on the grounds that the reduction of time and effort more than offsets the loss in precision. Finally, the theory of riskless choice is useful becasue it deals with the ordering of consequences, and it is a fundamental principle of rational choice that the ordering of consequences should not depend on whether or not the decision involves risk. Thus, there are cases in which we can apply a simple monotonic transformation to a riskless utility function to obtain a utility function which is appropriate for decision making under risk.

A riskless decision strategy is said to be *rational* (Arrow, 1963) if it satisfies two simple and logically compelling principles. The first of these, *comparability*, asserts that for any two outcomes X and Y, either $X \dot{<} Y$, $Y \dot{<} X$, or $X \sim Y$, where $X \dot{<} Y$ denotes "X is not preferred to Y," and $X \sim Y$ denotes "the decision maker is indifferent between X and Y." This condition is trivial, merely asserting that the decision maker can compare any two outcomes in terms of their desirability. The second condition, the transitivity principle, is probably the most important and least controversial principle of rational choice. It asserts that for any three outcomes X, Y and Z, if $X \dot{<} Y$ and $Y \dot{<} Z$, then $X \dot{<} Z$. For a finite set of outcomes, satisfaction of the comparability and transitivity principles is sufficient to guarantee the existence of a *riskless value function* V such that for any outcomes X and Y

$$X \dot{<} Y \text{ if and only if } V(X) \dot{<} V(Y).$$

Simple as the above notion of rationality may seem, the intuitive preferences of individuals sometimes violate the transitivity principle (Luce & Suppes, 1965; Tversky, 1969). To the information-processing-oriented psychologist this immediately suggests that, as in the case of probabilistic inference, people have difficulty aggregating information over the various dimensions of a stimulus. So here, too, decision theorists have focused their attention on methods for decomposing the judgment task into a simpler set of subtasks. They have also been concerned with the assumptions that are required to justify the various types of decompostions.

No decomposition is possible unless preferences satisfy a condition termed simple independence (Krantz and Tversky, 1971). Notationally, let the outcome X be represented by the attribute vector $(x_1, x_2,...,x_n)$, where x_k is the k-th attribute of X. Preferences satisfy the simple independence condition provided that there exists at least one attribute x_k such that preferences for the states of x_k, with all other attributes held fixed, do not depend on the particular states in which the other attributes are held fixed. To illustrate, consider the three attributes in our ABM sites example that relate to the "defense if deterrence fails" goal. Regardless of what proportion of the nation's population and industrial capacity might survive a nuclear attack, more agricultural capacity would probably be preferred to less. Thus, agricultural capacity satisfies simple independence. Similarly, more industrial capacity would probably always be preferred to less, so industrial capacity also satisfies simple independence. In both cases, one could always choose later to operate at less than full capacity.

It is not obvious, however, that one would always wish for a higher rate of survival among the nation's population. For example, suppose that only 10% of the nation's agricultural capacity survived an attack. Would one then prefer a 80% survival rate among the population to a 40% rate? In the former case, there would be a terrible food shortage which might in the long run result in a lower final survival rate, say after five years. In addition, the existence of extreme shortages of food might eliminate any possibility of returning to a semblance of order.

These comments are only speculative. Nevertheless, it should be clear that there are dimensions which will not in general satisfy the simple independence condition. In exploiting the consequences of simple independence it is not necessary, however, that the condition be satisfied for all conceivable outcomes, but only that it be satisfied for the outcomes being evaluated. For example, if no scenario being considered resulted in less than a 30% survival of agricultural capacity, one might well always prefer a higher population survival rate to a lower one, in this restricted set. Thus, tests of simple independence, and all of the other assumptions which we will discuss, should generally be confined to the set of outcomes being evaluated.

When one or more dimensions do satisfy the simple independence condition, it is possible to select one of these as a base dimension and then to <u>trade off</u> all other dimensions against the base dimension. In addition to satisfying simple independence, the base dimension selected should be: a) continuous, and b) capable of assuming a wide range of values. After a base dimension has been chosen, "standard states" for the remaining dimensions must be specified. These standard states should normally be chosen to fall somewhere in the middle of the range of states that the dimension can assume. To illustrate how the <u>sequential trade-off procedure</u> can be used to order outcomes, consider the following two outcomes from the ABM site problem. (Here we consider only the evaluation of outcomes with respect to the defense-if-deterrence-fails criterion.)

	X	Y	Standard
% Ag.	75	10	--
% Pop.	30	70	50
% Ind.	20	70	50

Outcome X represents a case in which urban areas suffer extensive damage resulting in loss of life and industrial capacity, while rural areas are relatively untouched, resulting in a fairly high level of surviving agricultural capacity. Outcome Y reflects an opposite pattern of damage, with urban areas suffering less damage, but with rural agricultural areas being very hard hit. In this example, we have selected agriculture as a base dimension. As noted earlier, the population factor may not satisfy the simple independence assumption. And, more important, people would probable find it extremely difficult to make direct tradeoffs into the population dimension. For both population and industry, fifty percent survival rates were used in this hypothetical example as standard levels.

Considering first outcome X, we begin the process of trading off population and industry survival rates against agriculture by attempting to specify a level of agricultrual survival (a') such that (.75A, .30P, .20I) $\sim$ (a'A, .50P, .20I). That is, we ask the question "How much agricultural capacity would we sacrifice to increase the population survival rate from 30% to its standard level of 50%?" Clearly, there is no objectively correct answer to such a question. Any well informed opinion would have to be based on projections of how a reduction in food per capita would affect long-run survival rates and quality of life for the survivors.

For the sake of illustration, suppose that the decision-making group sets a'=.30. Next, we trade off industrial capacity against agriculture by attempting to specify an a' such that (a'A, .50P, .50I) ~ (.30A, .50P, .20I). In such a crisis, agricultural capacity would presumably be highly valued, so assume the decision makers specify a'=.25. That is, given that food is already scarce, the extra industrial capacity would be of little value. Thus, by transitivity we have (.75A, .30P, .20I) ~ (.25A, .50P, .50I). Suppose that through a similar process for outcome Y, we obtain (.10A, .70P, .50I) ~ (.40A, .50P, .50I). In this case, the trade-offs imply a belief that the high initial survival rates for population and industry will be considerably offset by the absence of food to feed the survivors. Since we have by our trade-offs equilibrated X and Y in the population and industry dimensions, it is now quite straightforward to determine that Y is preferred to X from Y ~ (.40A, .50P, .50I) > (.25A, .50P, .50I) ~ X.

In what sense has this procedure assisted the evaluation process? Clearly the trade-offs involved would be extremely difficult to make. But any decision procedure implicitly implies such trade-offs, and it can be argued that it is better to know what trade-offs you are making so that you can be sure that they reflect your real beliefs and values. In addition, the sequential trade-off procedure simplifies the evaluation process by allowing the decision makers to concentrate primarily on only two dimensions at a time. This could be of considerable benefit in problems involving a large number of dimensions. For a more complete discussion of this approach and its shortcomings, see Raiffa (1969).

A second approach to riskless value assessment utilizes additive rating scale evaluation models. This approach is more restrictive and is appropriate only when stronger assumptions are satisfied. First, within the set of outcomes to be evaluated, every dimension must satisfy the simple independence condition. In addition, all dimensions must satisfy the following joint independence assumption. Preferences for combinations of any subset of dimensions, holding all other dimensions constant, should not depend on the particular levels in which the constant dimensions are held fixed. For example, preferences for various combinations of remaining agricultural and industrial capacity should not depend on the percentages of surviving population.

The joint independences assumptions seems questionable in the context of the example we have been discussing. (The example was in part selected because it would represent a difficult and extremely complex evaluation problem.) In most decision-making centexts, however, dimensions can be defined in such a way that both simple and joint independence will be staisfied. It should be noted, however, that in contrast to the transitivity principle, the simple and joint independence assumptions have no normative appeal. When they are staisfied, however, the evaluation problem is greatly simplified. In particular, when these two assumptions and several technical assimptions relating to the continuity of dimensions and trade-offs are satisfied, then preferences can be represented by a simple additive model of the form

$$V(X) = V_1(x_1) + V_2(x_2) + \dots + V_n(x_n), \qquad (2.1)$$

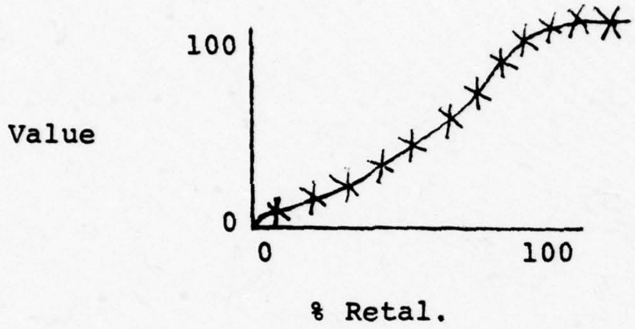
where V_i is a value function defined over the i-th dimension. When the individual value functions are scaled to have a constant range, say 0 to 1, then the model may be written as

$$V(X) = w_1V_1(x_1) + w_2V_2(x_2) + \ldots + s_nV_n(x_n), \quad (2.2)$$

where the w_1 are scaling or weighting factors. The formal theory underlying this additive model is termed the theory of conjoint measurement; for a further discussion see Krantz, Luce, Suppes, and Tversky (1971). It is particularly important to note that the assumptions discussed above guarantee only that an additive model can represent the preference ordering.

Nevertheless, most procedures for developing an additive value index for riskless choice assume that the index has the properties of an interval or ratio scale. Here we will outline a straightforward rating scale procedure that has been fairly widely used. We assume that the initial steps in the evaluation process have been completed; that is, that the utility for whom and what questions have been answered, that the set of value dimensions has been specified, that the set of outcomes to be evaluated has been specified, and that each outcome has been measured or scored with respect to each of the value dimensions. Next, value functions must be assessed over each of the dimensions. Then, the decision makers must specify the most and least desirable states of the dimension that could feasibly occur in the analysis. These states may be arbitrarily assigned within dimension values of 100 and 0, respectively. All intermediately valued states of the dimension are then assigned values using a direct rating scale method. If the dimension is continuous, values may be elicited for three to five intermediate states and a smooth curve interpolated through them.

In the example we have been considering in this discussion, the value function for percentage of retaliatory forces surviving a first strike might look something like the one below.



The s-shape of this function indicates that anything much above 50% survival of our missile forces should provide an adequate deterrent, while anything less than 10% provides very little deterrent whatsoever.

After value functions have been assessed over all dimensions, weighting or scaling factors must be assessed. These factors are necessary because although each of the value scales runs from 0 to 100, they are not scaled in common units. That is, a 10-point change on an important dimension should matter much more than a 10-point change on a trivial dimension. To assess these weights, the decision makers should begin by rank ordering the dimensions in order of their importance. This ordering should be based on the change in overall value induced by moving the dimension from its best feasible state to its worst feasible state. Thus, the importance of a dimension depends upon the range of feasible outcomes with respect to that dimension. After the dimensions have been ordered, quantitative weighting factors can be assigned by having the decision maker make ratio comparisons of the relative importance of pairs of attributes. Typically, either the most or least important dimension is compared with all others. These importance weights should then be normalized to sum to 1.0. Overall values for any given outcome may then be obtained by using the additive form of equation 2.2. For a more complete discussion of this procedure, see Edwards (1971).

Experience suggests that the additive rating scale procedure will generally be easier to use than the sequential trade-off method. This seems to be due to the fact that people are very uncomfortable about making direct trade-offs, particularly when one of the dimensions involves loss of human life. But the additive rating scale model implies a set of direct trade-offs which the decision maker should be willing to live with. In applying the additive rating scale method, care should be taken to assure that the decision makers are aware of the implied trade-offs and that these trade-offs reflect their true values.

The rating scale procedure is also much easier to use in situations in which a large number of outcomes must be evaluated. The sequential trade-off procedure must be directly applied to each outcome being considered. When the number of outcomes is large, the time and effort required could be prohibitive. The additive rating scale method, on the other hand, is unaffected by the number of outcomes to be evaluated. Once the value functions and weighting factors have been assessed, a computer can be used to evaluate any possible outcome for which dimensional measures or scores are available. (Of course, the task of measuring or scoring outcomes with respect to dimensions might also be extremely costly in terms of both time and resources.) Because it is so much easier to use, the additive rating scale method will probably be preferred in many practical situations. It should be noted, however, that the additive model is based on a considerably more restrictive set of assumptions, and when these assumptions are not satisfied, the sequential trade-off method provides an alternative approach. Trade-off methods are also instructive in communicating the true meaning of any multiattribute evaluation model.

3.3.2 Risky decision making. A decision is said to be *risky* when the decision maker is uncertain as to the consequences that will result from each course of action, but is able to specify a probability distribution over these consequences. This probability distribution might be obtained through direct intuitive assessment, as the output of a PIP system, or as the output of a statistical or simulation model. Notationally, let $(p_{1i}x_1, p_{2i}x_2, \ldots, p_{si}x_s)$ be the probability distribution for outcomes associated with the i-th action alternative where p_{ji} is the probability that outcome x_j will occur given that act A_i is selected. Recall from our previous discussion of decision making that the expected utility principle provides a formally optimal rule for choosing between alternatives, each of which gives rise to a probability distribution over a set of outcomes. According to this principle, there exists an interval scale utility function U such that:

a) For any outcomes x_i and x_j x_i < xj if and only if $U(X_i) \leqq U(X_j)$;

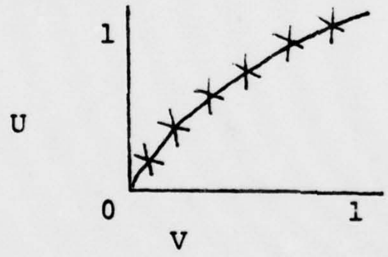
b) For any actions A_m and A_n, $A_m < A_n$ if and only if $\sum_{i=1}^{s} p_{im}U(X_i) < \sum_{i=1}^{s} p_{in}U(X_i)$.

Here we consider the problem of assessing the utility function U. Note that like any riskless value function V, U must preserve the preference ordering for outcomes. In addition, U must possess the interval-scale properties required for the expected utility computation.

When outcomes are along only one value relevant dimension, risky utility assessment is relatively straightforward. One method, discussed by Raiffa (1968), is particularly simple and may be applied in a wide variety of contexts. Given the finite set of outcomes $X_1, X_2, \ldots X_s$ to be evaluated, the decision maker begins by specifying the most and least desired of these outcomes, denoted by X^* and X_*, respectively. Then, for any other outcome X_i, he is asked to specify a probability p_i such that he would be indifferent between receiving the outcome X_i with certainty or accepting the probability distribution of consequences $(p_i X^*, (1-p_i) X_*)$. If the decision maker assigns these probabilities in an expected utility-maximizing fashion, then it can be shown that $U(X_i) = p_i$. Since the experimental studies reviewed earlier indicated that people evaluate simple gambles in a manner which is consistent with the expected utility principle, this seems a reasonable procedure for evaluating unidimensional outcomes. When outcomes are multidimensional, however, information overload problems again become important, so decision theorists have also developed decomposition procedures for risky utility assessment.

Note that since any riskless value function V and any risky utility function U must identically order outcomes, there must exist some monotonic transform R such that $U(X) = R(V(X))$. Thus, given a riskless value function V, we need only to assess the transform R to obtain an interval-scale utility function. Suppose, for example, that all outcomes have been traded off into some base dimension, so that the simple preference ordering over the base dimension itself serves as a riskless value function. Then to obtain the risk transformation R, one need only to assess a utility function over the base dimension itself. This utility decomposition is relatively simple to employ. The only really difficult stage here lies in obtaining the riskless trade off judgments.

When the riskless evaluation function is expressed as an additive rating scale model, it is not quite so simple to obtain the transformation R. Here several multi-attribute outcomes, which span the full range of V, must be selected and overall utilities assigned to them using the utility assessment procedure discussed below. The values of these outcomes are then plotted on the abscissa of a coordinate system, the utilities on the ordinate; a smooth curve is interpolated through them to obtain the transformation relating U to V. This interpolation procedure is illustrated below:



Note that although the value function V is additive in this case, the resulting utility function U need not be. For example, suppose that $U(X) = \log V(X) = \log \Sigma w_i V_i (x_i)$, which is clearly non-additive. In general, U will be additive only if U and V are linearly related. Throughout this paper we refer to these two risk transformation decompositions as <u>R(V) utility decomposition models</u>.

It is possible to obtain direct risky utility decomposition models, thus bypassing the need to construct a riskless value function. But direct utility decomposition requires that an additional assumption, termed utility independence (Keeney and Raiffa, 1976), also be satisfied. The utility independence assumption asserts that preferences for probability distributions over any subset of dimensions, holding all other dimensions constant, should not depend upon the particular states in which the constant dimensions are fixed. This is a slightly stronger version of the simple and joint independence conditions required for riskless additivity. Again, in the example we have discussed, we might expect that the utility independence assumption will be violated because preferences for uncertain outcomes with respect to the population dimension might well depend on the amount of surviving agricultural capacity. When this utility independence assumption and the simple and joint independence assumptions are satisfied, then it can be shown (Keeney and Raiffa, 1976), that a multi-attribute utility function will assume one of the following two forms:

$$U(x_1, x_2, \ldots, x_s) = \sum_{i=1}^{s} w_i U_i(x_i). \quad (2.3)$$

$$U(x_1, x_2, \ldots, x_s) = \frac{1}{k} \prod_{i=1}^{s} [(1 + w_i U_i(x_i)] - \frac{1}{k} . \quad (2.4)$$

Here the U_i are utility functions over each dimension, scaled to run from 0 to 1, the w_i are scaling constants to reduce the dimensions to a common scale, and k is a scaling constant reflecting the type and degree of non-additivity, if any, which is present. Keeney and Raiffa (1976) show that the additive form of 2.3 is a special case of the more general multiplicative model, 2.4. In particular, the additive form holds only if preferences under risk satisfy a final and extremely restrictive condition derived by Fishburn (1965) and frequently referred to as the marginality assumption. Fishburn proved that an additive risky utility assumption is appropriate only if the decision maker is indifferent between all alternatives that have identical marginal probability distributions over outcome dimensions. The following example shows that this assumption is often seriously violated. Consider that probability distributions PD_1 and PD_2 where

$$PD_1 = \left[\begin{array}{l}\text{with probability 1/2 the outcome (.8P, .8A, .6I)} \\ \text{with probability 1/2 the outcome (.1P, .1A, .6I)}\end{array}\right.$$

$$PD_2 = \left[\begin{array}{l}\text{with probability 1/2 the outcome (.8P, .1A, .6I)} \\ \text{with probability 1/2 the outcome (.1P, .8A, .6I).}\end{array}\right.$$

An additive risky utility function is appropriate for the example we have been considering only if the decision makers are indifferent between PD_1 and PD_2. It seems likely that most policy makers would prefer PD_1, which provides a 50% chance of having a large number of survivors with enough food to feed them, to PD_2, which gives either a lot of survivors with no food, or a lot of food with no one to eat it. That is, people care about the joint distribution, not just the marginals.

When the utility independence assumption is satisfied, a decomposed utility model can be constructed using the following procedure devised by Keeney and Raiffa (1976). To obtain within-dimension utility functions, the decision maker need only consider lotteries over the single dimension in question. Consider the i-th dimension and let $X_i{}^*$ and X_{i*} be the most and least desirable states of that dimension which could feasibly occur. Since the overall utility function is defined only on an interval scale, we may arbitrarily set $U_i(x_i{*}) = 1$ and $U_i(x_{i*}) = 0$. To obtain a within-dimension utility for any other state of the i-th dimension, $(x_i o)$ the decision maker must satisfy a probability $p_i 0$ such that he is indifferent between receiving $x_i o$ with certainty or accepting a gamble yielding the probability distribution $(p_i o x_i{*},\ (1 - p_i o)\ x_{i*})$, assuming all other dimensions to be held constant. It can easily be shown that $U_i(x_i o) = p_i o$.

Next, the scaling factors for each dimension must be assessed. Let $(x_i{*},\ x_{\bar{i}*})$ denote the outcome that has the most desirable state on dimension x_i and the least desirable state on all other dimensions. Then to assess the importance or scaling factor for the i-th dimension, the decision maker must specify a probability w_i such that he would be indifferent between receiving the outcome $(x_i{*},\ x_{\bar{i}*})$ with certainty and accepting the gamble $(w_i X{*},\ (1 - w_i)X_*)$. Again, it can be shown that w_i provides an appropriate scaling factor for either equation 2.3 or 2.4.

Finally, a choice must be made between the additive form of model 2.3 and the multiplicative form of model 2.4. Keeney shows that the additive model is appropriate if and only if $\underset{i}{\Sigma}\ w_1 = 1$. Thus, the w_i themselves provide a means to choose between models. In addition, if $\underset{i}{\Sigma}\ w_i \neq 1$, the multiplicative constant k in model 2.4 can be obtained using a simple interactive procedure based only on the w_i. (In practice, the selection of a model form should probably be based not only on the $\Sigma\ w_i$ test, but also on direct tests of the marginality assumption.) The multiplicative model 2.4 can express two types of interaction. First, when $\Sigma\ w_i < 1$, attributes combine in complementary fashion. High values on one dimension mean little unless all other dimensions have similarly high values. The preference for PD_2 in our example of the marginality test implies that $\Sigma w_i < 1$. In addition, the multiplicative model can reflect a substitutability relation. For example, a decision maker might be very loath to risk losing an outcome that was excellent in any attribute. Here, $\Sigma w_i < 1$.

Logically, both of these models are special cases of the more general R(V) formulation. Thus, whenever either of the Keeney models is appropriate, an R(V) model can also be used. In this case, choice of an evaluation procedure should be based on practical considerations such as ease and meaningfulness of assessment.

3.3.3 Validating multi-attribute evaluation models. Two approaches to the validation problem have been proposed. The first, and by far the most widely applied approach, is convergent validation. Studies employing this strategy typically begin by having subjects make overall intuitive judgments about multidimensional outcomes, usually on some type of rating scale. Then decomposed evaluation models are constructed and used to assign values or utilities to the same set of outcomes. Finally, the intuitive and decomposed evaluations are correlated with one another. Most studies which have employed this strategy have considered only riskless rating scale procedures. The correlations obtained in these studies have generally been quite high, ranging from the low .70s to high .90s, with most correlations in the high .80s or better (Pollack, 1964; Hoepfl & Huber, 1970; Huber, Daneshgar, & Ford, 1971; Pai, Gustafson, & Kiner, 1971; Fischer, 1972). In a slight variant of this approach, additive rating scale models have also been validated against hypothetical (Yntema & Klem, 1965) and real choices (Huber, Daneshgar, & Ford, 1971). In both cases the models afforded fairly good predictions.

Trade-off and risky utility assessment procedures have received less experimental attention, perhaps because psychologists have been less aware of the relevant literature dealing with these models. Fischer (1972) compared a special additive case of the sequential trade-off method with intuitive judgments and additive rating scale models and found a high degree of convergence between the three methods. Von Winterfeldt (1976) studied additive risky utility models and found that they did a fair job of predicting intuitive assessments of 14 dimensional outcomes. Fischer (1972) compared the Keeney and R(V) approaches to risky utility decomposition and obtained a high degree of convergence between the two methods. Both types of models also yielded excellent predictions of intuitive utility assessments.

In summary, convergent validation studies have generally indicated a high degree of agreement across different scaling procedures. This agreement suggests that all of the methods do in fact tap the same underlying attribute, namely, subjective value.

The second approach to validating multi-attribute evaluation models rests on the notion of external validity. That is, given that the true value of the outcomes will be ultimately known, it is possible to validate assessment procedures by comparing their outputs with these objectively correct values. In this vein, Yntema and Torgerson (1961) had subjects assign values to visual stimuli varying in size, shape, and color. During the training portion of the experiment, subjects estimated the value of a stimulus, then received feedback on its "true value." These true values were generated by a simple mathematical rule which was arbitrarily specified by the experimenters. On a subsequent series of test trials, the subjects assigned values to the stimuli without feedback. Finally, decomposed additive rating scale models were constructed to evaluate the same set of test stimuli. Both the intuitive and decomposed evaluations correlated highly with the true values of the test stimuli, with the decomposed models doing slightly better. This study provides further support for the belief that decomposition procedures provide a good means for assigning values to outcomes.

A second study involving an external criterion was conducted by Eils (1977). In that study, utility assessments about the credit worthiness of people reflected in credit briefs were elicited from twenty-four groups, each of which consisted of four graduate or undergraduate students who knew each other prior to the experimental session. Group utilities were elicited (via consensus) for ten hypothetical applicants for bank credit cards. The research design completely crossed two factors in assessing group utilities: 1) using a decomposition procedure (MAU) or not, and 2) using a formal group communication strategy (GCS) or not. The quality of each group's utility judgments was defined to be the Pearson product moment correlation between the group's judged utilities and utilities output from a configural (nonlinear) model used by a bank in evaluating applicants' credit cards.

Eils found that the decision technology of MAUA greatly aided groups in reaching decisions that were in some sense consistent with decisions based on a systematic collection and interpretation of a large amount of relevant data (i.e., the bank model). When unit weights were used in place of the elicited differential weights, the MAU groups evidenced even higher correlation with the bank model. The application of a communication strategy did not significantly alter the quality of group evaluations.

Eils' research is perhaps the first to demonstrate a greater degree of fit to an external criterion than wholistic judgments. The formalized bank model used to measure judgmental validity reflects the complex nature of the relationship between applicant characteristics and a subsequent loan, as evidenced in the data used to generate the formal model. Although this criterion is not a totally satisfying one, it is clearly better than none.

3.3.4 The social utility problem. All of our discussion of multi-attribute utility models has implicitly assumed the existence of some decision maker, with a consistent set of preferences, whose utility is to be maximized. Yet our example, and most contexts in which we would like to apply the theory, involve many decision makers or interested parties with conflicting sets of values. Arrow (1963), however, has proved that there is no way to combine a set of individually transitive preference orderings which assures a collectively transitive group preference ordering, at least not if the method of arriving at a group preference ordering satisfies a number of criteria, such as non-dictatorship, which are generally valued in democratic societies. Because the issues here are quite complex, the reader is advised to refer either to Arrow's original work or to the more intuitive discussion presented in Luce and Raiffa (1957).

Arrow's formal arguments have not discouraged those who wish to apply utility theory in complex organizations. Although the resolutions to this problem are heuristic at best, they seem intuitively reasonable, and seem much more desirable than traditional institutional power struggles which also must implicitly cope with the problem of conflicting sets of values. As Edwards (1971) has noted, many participants in a decision-making process are expert on only certain aspects of the problem. In applications of multi-attribute utility theory, such experts only make judgments about the particular problem dimensions falling in their area of expertise. When more than one expert is involved in assessing a given value or utility function, some sort of averaging procedure can be used (for example, O'Connor, 1972). Although no formal rationale can be given for such averaging, it seems reasonable. In other cases, experts may be able to reconcile their differences to arrive at a concensus function for the dimension in question.

The social utility problem is more severe in the matter of specifying trade-offs among dimensions, or equivalently, assigning importance factors to dimensions. If there is one organizational decision maker with overall responsibility for the decision, he should make these trade-offs. (This will often be the case in military organizations.) If several people must share the ultimate responsibility, then they should attempt to resolve their differences to arrive at a consensus set of trade-offs. Sensitivity analyses can be helpful in this regard, in that often the decision will be such that small value conflicts will not affect the final decision. But if all such attempts fail, some sort of averaging process might be used. Nash has shown that, when utilities are defined on an interval scale, a multiplicative averaging process has certain normatively desirable properties possessed by no other strategy for resolving interpersonal utility conflicts. (See Luce & Raiffa, 1957, for a discussion of the group utility problem, including the Nash solution.)

Whatever approach is adopted, it seems clear that the multi-attribute utility approach can assist members of an organization in identifying their implicit value conflicts and force them to communicate directly about them. Despite the heuristic nature of the approaches discussed here, we would argue that they represent a substantial improvement over intuitive goal-setting in organizations which results in different sub-units pursuing conflicting goals or in sequential attention to conflicting goals over time (Cyert & March, 1963; Allison, 1971).

3.4 Bootstrapping--An Alternative Approach to Decision-Aiding Technology

All of the decision-aiding technology discussed above has been based on the assumption that better decision making can be achieved if complex problems are decomposed into a set of relatively simple factors, with people making judgments about these factors, and with a mathematical model being used to aggregate the implications of these judgments for the final decision. A second approach to decision aiding is based on the assumption that intuitive inferences and decisions are unbiased, but subject to a substantial degree of random error (Bowman, 1963; Goldberg, 1970). When this assumption is true, it should be possible to fit a statistical model to a set of intuitive judgments which can capture the systematic aspects of the judgments while filtering out the random error. Yntema and Torgerson (1961) provided the earliest empirical support for this decision-aiding strategy which has come to be called "bootstrapping." They fit additive main-effects models to the test trial responses in their experiment in which subjects learned to assign values to geometrical stumuli. These main-effects models correlated .88 with the true values of the stimuli, whereas the intuitive judgments upon which the main-effects models were based correlated only .84 with the true values.

Other examples of bootstrapping for both decision making and inference will be discussed in the real-world applications section of this paper. For a more comprehensive discussion of the logic of bootstrapping, see Bowman (1963) and Goldberg (1970).

4.0 REAL WORLD APPLICATIONS OF DECISION THEORETIC AIDING CONCEPTS

Over the past ten years, and particularly during the past five years under funding support provided by the Defense Advanced Research Projects Agency, there has been a burgeoning of efforts to apply the decision-aiding concepts discussed in this report. These real-world applications have focused on decision problems in both the civil and military sectors and the aiding concepts employed cover the spectrum of elements of decision-analytic methodology ranging from the explicit numerical expression of uncertainty through probabilistic inference, utility assessment, and comprehensive decision analytic applications.

In this section, we will review applications of this technology with particular emphasis on evidence bearing on the utility of the decision-aiding methods. As will be noted, however, there are severe practical difficulties which, in most applied decision contexts, preclude objective evaluation of the worth of decision aids. Among these difficulties are the frequent absence of an objective criterion against which to assess decision quality, the general lack of parallel decision channels that would permit comparative assessment of alternative decision methods, and the unique, non-repetitive nature of most significant decisions, a factor that further limits objective evaluation. Because of these practical measurement problems, much of the evidence about the utility of decision aids in applied contexts is anecdotal and far short of the level of experimental rigor that would be desired. An excellent discussion of some of the problems attendant to "real-world" evaluations is contained in Fischhoff (1977) and Miller, Kaplan and Edwards (1967).

4.1 Probabilistic Information Processing

Intelligence analysts are almost solely concerned with highly fallible predictions. Traditionally, they have relied on verbal analyses of the type frequently encountered in history or political science. The first significant break with that tradition came when Zlotnick (1968), an analyst, learned about PIP from reading some of Edwards' articles, and decided (with consultative help from Edwards) to try it out on intelligence data. His article reports a re-analysis of the data from the Cuban Missile Crisis--a re-analysis that suggests the possibility that if PIP had been in use at the time, the United States might have had significantly earlier warning that the Russians were putting long-range missiles into Cuba. Other studies not reported in Zlotnick's paper were carried out, with similar results.

Although subsequent to these experiments, interest in PIP wanted in Zlotnick's agency, there as been a recent resurgence of interest. Schweitzer (1976), for example, describes the application of Bayesian inference and Delphi techniques to a number of intelligence estimation problems. Other applications of decision analytic methodology are reported to be underway.

Kelly and Peterson (1972) report on work in a different intelligence agency. This work led to the present situation wherein probabilistic and Bayesian techniques are in production use in elements of the agency, and incoming analysts are routinely trained in the techniques.

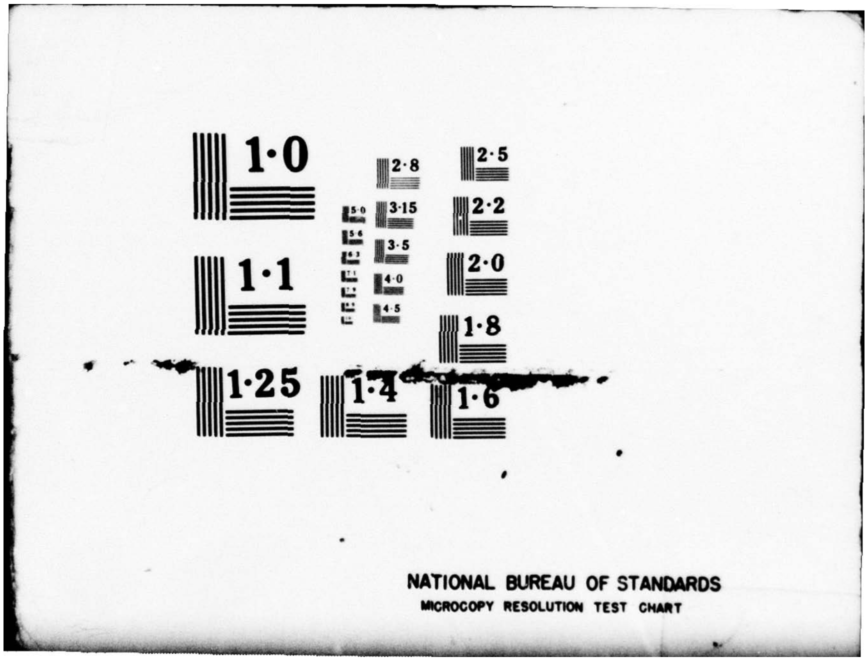
One key to Kelly and Peterson's success, perhaps the salient one, was that they did not set out to do research on PIP or to validate its usefulness in a particular context. Instead, they simply set out to discover how probabilistic techniques could be made a part of the working tools of certain intelligence analysts. To overcome initial resistance to quantitative techniques, Kelly and Peterson sought to demonstrate the ambiguity of verbal expressions of uncertainty. In one demonstration, they asked groups of analysts who co-authored intelligence estimates to express the implied predictions in quantitative terms. Although the analysts felt somewhat uncomfortable about the implied precision of these estimates, the results were so striking that small errors were of little concern. To cite one particularly surprising example, two analysts had co-authored a paper in which they stated that "The ceasefire is holding but it could be broken in the next week." One author interpreted this as implying a .30 probability that the ceasefire would collapse. The other interpreted it as implying a .80 chance of a collapse. Prior to the study, they had not realized that they disagreed about what the quoted sentence meant. In general, the Kelly-Peterson data clearly established that verbal qualifiers (such as "likely," "could," "probably") provide an extremely poor means for conveying subjective beliefs about the likelihood of certain events.

After winning the analysts' support for quantitative probability assessment, Kelly and Peterson focused on the problem of determining which response modes seem to provide the most reliable and meaningful responses. Based on day-to-day experience, Kelly and Peterson concluded that, of the response modes tried, a logarithmically spaced scale calibrated in both odds and probabilities works best. They also discussed problems associated with applying PIP systems in real-world intelligence contexts. For example, analysts found it very difficult to estimate likelihood ratios. Statistical models are generally applied to problems where it is reasonable to assume that some hypothesized state of the environment generates a set of data. In intelligence analysis, however, it frequently appears that causality runs from data to hypothesis. For example,

AD-A053 962 DECISIONS AND DESIGNS INC MCLEAN VA F/G 5/10
DECISION THEORETIC AIDS FOR INFERENCE, EVALUATIONS, AND DECISIO--ETC(U)
FEB 78 G W FISCHER, C W KELLY, W EDWARDS N00014-76-C-0074
UNCLASSIFIED TR-78-1-30 NL

2 OF 2
ADA
053962

END
DATE
FILMED
6-78
DDC


1·0
2·8
2·5
5·0
5·6
3·15
2·2
3·5
2·0
4·0
1·1
4·5
1·8
1·25
1·4
1·6
NATIONAL BUREAU OF STANDARDS
MICROCOPY RESOLUTION TEST CHART

increasing the number of troops stationed at the Sino-Soviet border (a datum) increases the odds of Sino-Soviet military conflict (a hypothesis) at least in part because such troop concentrations might provoke border incidents that could escalate into Sino-Soviet conflict. The analysis techniques used by Kelly and Peterson also typically provide the analysts with feedback on the posterior odds generated by the PIP system. Although laboratory work suggests that inferences are less conservative when no feedback is provided, Kelly and Peterson argue that it would be unreasonable to expect analysts to accept responsibility for the predictions of a PIP model without being aware of the model's predictions. (The same point had previously been raised by Edwards, Phillips, Hays, and Goodman, 1968.)

Peterson and Kelly (1976) have collected data relating intelligence forecasts to actual outcomes, a calibration method made possible only when uncertainties are expressed in the form of explicit probabilities as opposed to verbal qualifiers. They found that analysts' probabilistic estimates were quite realistic in terms of the percentage occurrence of predicted outcomes. There was a slight tendency for analysts to overstate the likelihood of future events, a result which is to be expected in light of the warning function analysts serve and the costs of underestimating in that context.

Decisions and Designs, Incorporated, under Air Force and ARPA sponsorship, has developed and placed in trial use a computer-based Bayesian hierarchical inference model designed to serve an indications-and-warnings function of interest to the intelligence community. The model permits continuous analysis of 120 threat indicators and produces an intelligence conclusion in the form of a probability distribution over several potential courses of hostile action. This model (described in Barclay, 1976; Stewart, 1977; and

Stewart, Chinnis, Kelly and Peterson, 1976) has not been in use long enough to assess its performance against actual outcomes. In terms of user acceptance, preliminary indications are that the users (intelligence analysts at a major Command Headquarters) are uncomfortable with the precisely expressed probabilistic output of the model. Whether this disquiet is over a mistrust of the model dynamics and, hence, of the model output, or whether it stems from the loss of "maneuvering room" afforded by the previously used verbal estimates remains to be seen.

Hierarchical inference models have also been applied to other intelligence assessment problems such as order-of-battle assessments (Kelly and Stewart, 1976), and technology forecasting. Here again, the events against which the models could be assessed have not yet materialized, so there is no basis for formal validation.

Gustafson and his colleagues at the University of Wisconsin have been active in the application of PIP to medical diagnosis and related topics. In his dissertation, for example, Gustafson (1966) showed that Bayesian methods could be used to predict the number of days of hospitalization for patients after hernia operations. In a crowded hospital, such predictions could be useful for scheduling surgery for non-emergency cases and for anticipating acute overcrowding on surgery wards. In a 1969 paper, Gustafson, Edwards, Phillips, and Slack called for more general application of Bayesian techniques in medical diagnosis. Gustafson, Kestly, Greist, and Jensen (1971) have discussed the application of Bayesian inference models to the diagnosis of thyroid disease. In this work, they compared both subjective Bayesian models in which physicians assessed the likelihood ratios for disease states, given various signs and symptoms, and actuarial Bayesian models in which likelihood ratios were estimated from a large sample of clinical records. In general, both approaches worked well, achieving an approximately equal degree of accuracy. This approximate equality is quite noteworthy. For, in terms of cost, the subjective approach is much less expensive to implement. In many real-world contexts, in fact, it would not even in principle be possible to collect a set of data upon which to base a model. In a conference[1] on decision making and subjective probability, Gustafson discussed a similar system used to assess the likelihood that psychiatric outpatients will attempt suicide. The model apparently yields much more accurate predictions than do the intuitive assessments made by clinical psychologists and psychiatrists.

In 1968, Dr. Lee Lusted, a radiologist, wrote a book explaining decision-analytic techniques and calling for their widespread application to medical problems. An application is now in progress, under the leadership of Lusted, Dr. John Loop, Edwards, and others. The problem is assessment of the efficacy of x-rays. In principle, an x-ray might be called efficacious only if the patient is better off as a result of its having been taken than he would have been otherwise. This ideal definition of efficacy turns out to be impractical for large-scale screening in radiological practice because it requires judging what would have happened if the patient had been treated differently from the way he was treated, and physicians find it virtually impossible to make such judgments. Instead, a far more modest definition of efficacy has been adopted: an x-ray is efficacious if it affects the physician's diagnostic thinking. This can be measured by observing his prior distribution over the patient's possible diseases, and then observing his distribution just after receiving the x-ray results. If the two distributions are identical, the procedure was inefficacious.

[1] Ann Arbor Bayesian Conference, May 1973.

To proceed with the study, Lusted and his co-workers trained approximately 50 radiologists in techniques of probability estimation; the radiologists in turn trained attending physicians in the same estimating techniques. In the course of their medical practice, participating physicians, when requesting radiological services, supplied information about their tentative diagnosis, the reason for requesting the x-ray, and their estimates of the probabilities that the most important prospective diagnosis and most likely prospective diagnosis would prove correct. After receipt of the requested x-ray, the physicians re-estimated the probabilities in light of the radiological evidence. As of July 1976, approximately 8,000 such pre- and post-x-ray estimates were in hand. Among the tentative results gleaned to date are the following:

1. The procedure is feasible. The required probabilistic estimates can be made in an orderly way and do provide information about the diagnostic thinking of attending physicians.

2. In the vast majority of cases (over 90%), x-ray examinations do impact diagnostic thinking.

3. About 75% of the examinations produced a lowering of the clinicians' initial probabilities for the tentative most important diagnosis, thus the tests were more a matter of reassurance than of confirming alarm.

This study is of particular significance in supporting the usefulness of the methodology in yielding information about the behavior of individuals performing socially important and policy-relevant functions. Extensions of this work may well lead to recommendations for improving the distribution of health-care services.

The use of probabilistic inference models is also growing rapidly in business applications of Bayesian decision theory. In these applications, the probability model is typically embedded in a larger model which is used to select a utility maximizing option. Spetzler and Stael von Holstein (1972) prepared a report summarizing the insights which have emerged from the work of the Decision Analysis Group of the Stanford Research Institute. In contrast to the medical and military intelligence contexts, most of the important uncertainties in business problems involve continuous variables. Thus, Spetzler and Stael von Holstein's article is oriented primarily toward assessment procedures for continuous variables. They argue that proper specification of the uncertain variable is crucial to the encoding process. In particular, the variable should be so stated that an omniscient observer with perfect foresight could specify the outcome which will actually occur in the form of a simple number--for example, the price of gold on the Paris market at the close of trading on 30 June 1974 will be $X. The quantity should be expressed in terms of a measurement scale which is meaningful to the person making the judgment. Spetzler and Stael von Holstein also strongly emphasize the importance of eliminating implicit assumptions about other uncertain quantities by fully decomposing the inference tree. They have found that "as-iffing" can be a serious problem in real-world contexts. Spetzler and Stael von Holstein go on to discuss specific techniques for eliciting judgments from subjects. Most of these have been discussed in previous sections of this paper. In their work, they typically ask the expert assessor to specify a few points on a distribution, then fit some theoretical function (such as a normal curve) through these points. They argue that the functional form of a distribution should usually be determined by *a priori* modeling considerations. In this approach, the subject's judgments simply specify the parameters of this preselected distribution.

We will encounter other applications of subjective probabilities in our subsequent discussion of decision analysis. These applications typically have used relatively standard procedures and do little to improve the methods already discussed. They do, however, provide evidence for the feasibility of the approach.

It is unfortunate that real-world applications of PIP have not attempted to validate the predictions of the models against actual outcomes. (The medical research of Gustafson and his colleagues is an important exception.) The semi-experimental studies of weather, stock market, and football-score forecasting conducted by Murphy, Winkler, and Stael von Holstein did provide such validation, but these studies involved only unaided intuitive inference. They did not decompose the inference problems as PIP does. Thus, they are not informative about the quality of the predictions which might be generated by a real-world PIP system.

To the totally convinced subjectivist, it is not necessary to validate a subjective probability model. If the model reflects the decision maker's true beliefs, then it is appropriate to use it. We are slightly more skeptical, however, and believe that it should be possible, over the long run, to evaluate real-world forecasting models in terms of the veridicality of their predictions. Such validation is possible, of course, only in decision-making systems which make a relatively large number of predictions over an extended period of time. We believe that such validation studies should receive high priority in future research on Bayesian decision-aiding systems. For if systematic biases are observed, it may be possible to improve the subjective inputs to the system, either through additional training or through mathematical transformation of the assessor's raw estimates.

4.2 Multi-Attribute Utility Assessment

Multi-attribute utility theory (MAU) is being used increasingly in significant applied contexts both as an independent tool for complex assessments and in the context of full-scale decision analyses. In this section, we will review utility theory applications which have been conducted without reference to some larger decision problem, then turn in Section 4.4 to applications in the context of full-scale decision analyses.

Chinnis, Kelly, Minckler, and O'Connor (1975) applied MAU procedures to the assessment of the relative military value of four alternative combat radio net configurations under consideration for the Army inventory in the post-1980 time frame. A hierarchical structure consisting of several levels was developed, starting with military utility, which was partitioned into major dimensions of utility, e.g., technical system utility and operational acceptability. These dimensions were further fractionated into sub-dimensions which, in turn, were further partitioned, each partition becoming more specific until a level was reached at which one or more technical performance characteristics served to describe each of the sub-dimensions. The military utility for different levels of each of the performance characteristics was established by assessing a utility function over the relevant range of that characteristic. The relative importances of the different performance characteristics were assessed by assigning relative importance weights and, using both additive and multiplicative combination rules, an aggregate, weighted utility for each system was derived.

The model was implemented on a computer to permit the user to conduct sensitivity tests and rapid "what if" analyses by varying performance parameters and weights and recalculating utilities. The model was subjected to both internal and external sensitivity tests as a form of validation. As would be expected if the model were valid, it was found to be insensitive to minor changes in assigned weights, yet sensitive to differences between systems. The model also appropriately identified the known lesser utility of a Korean-war-era system which was included in the evaluation as an external criterion. The Army Special Task Force responsible for the evaluation used the computer-based model not only for their own system evaluation purposes, but also as a highly convincing vehicle for presenting and justifying their selection to higher echelons as well. It is interesting to note that this same radio system evaluation problem was approached independently through construction and use of an elaborate simulation model. Both the MAU model and the simulation model produced the same results, but the MAU evaluation was completed in one-third the time and at one-tenth the cost of the companion simulation approach.

In another application (Allen, Buede and O'Connor, 1977), a multi-attribute utility model was developed to handle the extremely complex matter of assessing the combat readiness of military units, in this case, U.S. Marine Corps units. The model structure for evaluating a Marine infantry battalion decomposes the overall battalion mission into thirteen mission-performance standards. Each of these are further decomposed into successively finer-grained tasks and sub-tasks yielding at the lowest level of the hierarchy up to 800 specific characteristics for which ratings are required. These input values are then aggregated over weighted value dimensions to provide an overall measure of combat readiness. The disaggregation to 800 rating elements is probably far more than necessary for readiness-assessment purposes (an excess also recognized by the analysts). In this case, the level of detail was driven principally by the sponsor's interest in a comprehensive model and one that could be used to pinpoint specific things needing correction.

In applications of the model in the context of Marine Corps exercises, user acceptance was found to be high. Marine Corps officials responsible for combat-readiness evaluations reported that use of the model reduced evaluation efforts by an estimated 75% over previously used methods and provided the best method for readiness evaluation yet devised. The model, or refinements of it, has been adopted for Marine Corps-wide combat-readiness assessment applications.

Similar multi-attribute utility assessment models have been developed and applied in evaluation contexts, such as determining the optimal mix of aircraft for naval aviation (O'Connor, Rhees, and Allen, 1976), selecting optimal design proposals for a Navy Electronic warfare suite (Hays, O'Connor, and Peterson, 1975), system design evaluation for a variety of Army weapon system procurements (unpublished), and the assessment of alternative designs of a hostile weapons location system (Barclay, Chinnis and Minckler, 1975). A particularly interesting and valuable use of multi-attribute utility theory is presented by Barclay and Peterson (1976) in a report on a quantitative method for optimizing outcomes in complex multi-issue negotiations. The computer-based negotiation model developed by Barclay and Peterson has been used to explore alternative bargaining postures for the U.S. in two significant international negotiations.

In a medical context, Gustafson, Feller, Crane, and Holloway (1971), have developed an index of the severity of burns using additive utility models. In this model, they used five measures of burn severity: size of full thickness burn, size of partial thickness burn, age of the patient, number of past serious medical problems, and location of burns. The first four measures represented continuous, or near continuous dimensions, whereas the last measure, location of burn, was divided into nine dichotomous scores for face and head, front of body, and so on. Four physicians assessed functions relating each of these five measures to the overall severity of a burn. They also assigned weights to each of the measures using a ratio comparison procedure. The experimenters then averaged together the models of the individual physicians to obtain an overall model.

This model was validated in two ways. First, the same four doctors rank-ordered 28 hypothetical descriptions of burn patients in terms of the severity of their cases. The experimenters then averaged these ranks and correlated them with the rank-ordering predicted by the average overall model. The correlation obtained was .92, indicating that the overall model did a reasonable job of reflecting the subjective beliefs of the physicians studied. (The use of averaged ranks as a criterion variable seems undesirable, but it is probable that similar results would have been obtained if an interval-scale measure of severity had been used.) In addition, the model was validated against the survival rates of a real sample of burn patients. In general, severity of burn (as indicated by the model) showed a strong negative relationship with ultimate survival rates. Thus, the model shows considerable promise as a tool for classifying patients who will require varying degrees of sophistication in their treatment. The authors suggested that the model might also be useful for evaluating the effectiveness of burn treatment operations. The index could be used to standardize for the severity of the cases encountered by a particular facility.

Patrick, Bush, and Chen (1972) provide another example of the application of multi-attribute utility theory to medical problems. Their primary goal has been the development of an overall index of health status. Their index takes account of three classes of variables: age, disease state, and the ability of the person to carry out various types of mental, physical, and social functions. In the study cited above, they asked a group of medical professionals, including both students and administrators, to evaluate a set of profiles which described persons in terms of the variables listed above. Each evaluator made three sets of judgments, one based on ratio comparisons of cases, one on a simple categorical rating scheme, and the third using a complex indifference-judgment response. Averaging these intuitive ratings over judges within particular response modes, they obtained high (.90+) correlations between methods, especially the ratio and categorical rating methods. Within-judge reliability was also quite good, with test-retest correlations ranging from .74 to .83. Most encouraging was the considerable degree of consensus found across judges. Correlations between the ratings of individual judges and an overall model obtained by averaging over judges ranged from .7 to .8. Based on these results, the authors argued that it might well be feasible to develop an overall index of health status which would reflect the subjective preferences of the American public. Such an index would have to be based on survey research data, with respondents probably making categorical judgments about age-disease-functional state profiles. Such a model would not only provide a tool for evaluating the quality of medical care in different locations, but might also be useful for allocating research money. Whether or not the particular approach adopted by Patrick, Bush, and Chen ultimately proves useful, it seems likely that some kind of multi-dimensional health status index will be developed as a tool for guiding and evaluating the health care service provided in this country.

O'Connor (1972) has constructed similar indices for the evaluation of water quality. In this study, water-quality experts from across the country first specified a list of factors relevant to water quality, then constructed scales relating levels of each factor to overall water quality, then assigned weighting factors to each value dimension. Through repeated visits with the experts during which they were shown the weights and functions assessed by other experts, O'Connor was able to obtain consensus about the functions relating each dimension to overall quality. But conflicts between importance weights seemed to reflect real differences among the regional interests of the experts. Thus, O'Connor averaged their weighting assessments to obtain the overall water-quality index. (In fact, two indexes were constructed, the first for the evaluation of water for public use, and the second for water sustaining fish and wildlife.)

To validate his evaluation models, O'Connor used the overall index as well as the indexes for each of the individual experts to assign values to a real sample of water specimens. He found extremely high correlations between the models of the individual judges and the overall model. Although the experts disagreed about the weighting factors, the degree of disagreement proved quite minor, and had very little effect on the overall values assigned to water samples. O'Connor's work clearly established the feasibility of developing a set of water-quality indexes which could be used nationwide to evaluate water quality and to allocate resources for cleaning up our rivers, lakes, and streams. Natural resource and public health professionals have continued this work.[1]

In another project, Snapper, Guttentag, and Edwards applied multi-attribute utility measurement techniques to the problem of measuring the benefits of research on child development sponsored by the Office of Child Development of the Department of Health, Education, and Welfare. Value

[1] O'Connor, personal communication.

dimensions were elicited from a large group of experts on child development, both inside and outside the Office of Child Development (OCD). Importance weights were estimated. It turned out that the group could easily agree that some dimensions were clearly unimportant, but that they could not agree at all about the importance of the more important ones. The problem was resolved by using the weights assigned by the Director. Location measures on the dimensions, fortunately, presented far fewer problems of disagreement. In order to convert the interval-scale utilities obtained from multi-attribute utility measurement to ratio-scale numbers suitable for use in cost-benefit ratios, it was necessary to determine the true zero point. Fortunately, this proved easy: a research program has positive utility if and only if it would be undertaken if it had zero cost. Development of an explicit value system for OCD's research activities has served a catalytic function in stimulating OCD thinking about its program and its values.

Multi-attribute utility models also have considerable potential as a tool for measuring public preference for goods and services. In one application, Lehmann (1971) used utility models to predict stated preferences for television shows. Respondents in a survey-based marketing research study evaluated a set of television programs on each of six dimensions. They also weighted these dimensions in terms of their importance. Then they rank-ordered a set of 20 programs in terms of how much they liked them. Here, the mean rank-order correlation between additive utility models and the respondents' stated preferences was .72. Clearly, actual viewing habits would be much more difficult to predict, for one would have to consider the alternatives available at the time of the program. Nevertheless, the study does demonstrate the feasibility of using multi-attribute utility models in survey studies of public preference.

In another context, Klahr (1969) compared decomposed additive utility models with multi-dimensional scaling models in predicting the decisions of a graduate admissions committee. He found that both did a good job, with the simple utility model doing slightly better. His results suggest that, at least for preliminary screening, additive utility models might substitute for highly paid academic professionals who serve on admissions committees. Dawes (1971) has obtained similar results using bootstrapping techniques.

Although each of the applications described above concerned decision making under risk, the scaling procedures used did not explicitly reflect this fact. Keeney (1972a) has described a case study in which he had hospital managers develop a risky multi-attribute utility model for use in bloodbank decision making. The principal trade-off to be made in this context involves shortages versus the deterioration (and subsequent nonuse) of blood. The more blood held in stock, the less likely one is to be unable to fill a request. Unfortunately, large stocks also increase the probability that blood will go unused too long and will have to be disposed of as unfit for use. Keeney's article provides an extremely detailed discussion of the procedures involved in a realistic application of risky multi-attibute utility models. It is also noteworthy that the function developed was substantially non-additive. Ellis and Keeney (1972) and de Neufville and Keeney (1972) describe similar applications of risky multi-attribute utility models in the areas of air pollution regulation and the evaluation of alternative locations for a major international airport.

Most of the applications described thus far have resolved the social utility problem by averaging across evaluators. Spetzler (1968) presents an alternate approach in his discussion of the development of a risk policy for a large corporation. In all, thirty-six corporate officials, including the top officers, participated in the development of a utility function for the evaluation of risky capital investments. Initially, each official evaluated a large number of simple (and hypothetical) investment decisions. These decisions were used to estimate, for each official, the parameters of a fairly complex logarithmic utility function. Later, the whole group of managers met to discuss the implications of their earlier assessments. This discussion revealed a striking degree of risk aversion, even for decisions which involved investments which were very small relative to the total corporate assets. The final utility function agreed upon was much less risk averse than the utility function of the average manager. The main effect of the discussion, apparently, was to bring home the realization that the corporation need not display risk aversion for the large number of relatively small investment decisions which they made. Only for very large decisions was risk aversion desirable. In this case, discussion led to a consensus utility function which was much different from the one that would have been obtained had the original utility functions been averaged together.

Together, these studies illustrate that utility models can be implemented in realistic settings. Decision makers can be persuaded that the models can be of value to them, and can be induced to devote the time required to develop such models.

4.3 Computer-Based Decision Aids

Decision analytic applications have long suffered from the laboriousness of the method and the usual requirement for shepherding by a skilled decision analyst. As a result, a great deal of attention has been given to the development of computer-based routines to reduce these burdens. Most of the effort in this area has been focused on the development of computer tools suitable for use by decision analysts. More recently, there has been increased emphasis on the development of computer-based decision aids which embody decision analytic concepts and which are intended for on-line use by decision makers and their staffs, rather than by decision analysts. Ulvila (1975) presents a review of computer routines developed to aid in decision analysis applications.

In the remainder of this section, we will focus attention on recently developed computer-based, decision-analytic aids intended for direct on-line use by decision makers. Such aids are apparently few in number, but constitute important harbingers of future directions in the application of decision analytic concepts.

The first of these aids, known as OPINT, is designed for the rapid screening of decision options. The model is capable of handling decision problems having only one main decision--a single choice among several options--and is further restricted to decision situations having but one major uncertain event. To use the model, the decision maker specifies the structure of the problem by listing the decision options, the possible outcomes of the uncertain event, and the dimensions along which the value of the consequences of options and outcomes will be measured. The decision maker then addresses probabilities of occurrence for different outcomes, the desirability of each consequence along each value dimension, and the relative importance of each of these value dimensions. All this information serves as input to the computer program. The output of the computer program is a summary measure of expected value for each option. By adjusting the various inputs to the computer program, the decision maker can then test the sensivity of these expected values for the different options to various assumptions, for example, assumptions such as the probabilities of the outcomes and the weights of different dimensions of values for the consequences.

This evaluation procedure provides a way for decision makers to study a set of options in an approximate fashion when a rapid analysis is necessary either to identify areas for further detailed study or to support a decision that must be made immediately. The procedure can also be used before a problem arises for contingency planning. The capabilities and limitations of this model are fully detailed in Selvidge (1976).

The second computer-based decision aid is an additive multi-attribute utility model for option evaluation under certainty. Designated EVAL, this model was programmed for operation on the IBM 5100 portable computer and, like the OPINT model, is self-tutorial in that the program guides the user through the process of problem structuring and value elicitation. The EVAL model is capable of handling comparative assessment of eight alternatives simultaneously, each having up to eight hierarchical levels.

Under the technical cognizance of the third author, these decision aids were placed in trial use in a major U.S. military command headquarters in October 1976. Following introductory training in the concepts of decision analysis and the specific dynamics of the decision aids, the aids were used extensively by various staff elements of both the parent and subordinate headquarters in addressing and resolving real decision problems of immediate interest to the Commands. During the same time period, less extensive applied tests of the decision aids were conducted through participation in political-military simulation exercises, through applications in intelligence contexts and through use in a number of problem-solving demonstrations in a classroom environment. In all, during twelve months of use, approximately 45 different decision problems were addressed using the two generic decision aids. Most of the decision models were initiated and constructed by staff officers, to address decision problems confronted in their areas of responsibility. Some of the decision models were executed by decision analysts working in close conjunction with users.

Because of the factors discussed earlier that limit formal evaluation of decision aids in applied contexts, the kind of quantitative validation that would constitute firm evidence as to the effectiveness of the decision aids is again lacking. Kelly, however, offers the following anecdotal assessment of the trial application of the decision aids based upon his personal observations and user comments.

1. Introduction of the decision aids clearly stimulated the informal use of decision analytic concepts within the Command. The use of principles of problem disaggregation, probability and expected value became evident in staff decision recommendations completed without reference to the decision aids.

2. The decision models served a highly valuable communication function. The visible problem structure, implicit rationale, and sensitivity testing features provided by the computer-based models proved to be an efficient and compelling means for conveying staff recommendations for courses of action. This was particularly evident in simulation exercises conducted at the National War College where, for comparative evaluation purposes, groups of students, aided and unaided, decided on courses

of action in political-military simulations. While both groups reached the same general conclusion, it was the reported judgment of the faculty receiving the teams' briefings at the conclusion of the exercises, that the team using the decision aid had a far more compelling case for their decisions than did the unaided team. As another example, a staff officer at the major command test site reported that using the decision aid as an analytical and communications vehicle enabled him to obtain Command approval of his staff recommendation in three days in contrast to the norm of two and one-half weeks in a process requiring fifteen concurrences.

3. Use of the decision aids appeared to force users to distinguish between option value and likelihoods and to treat both explicitly. Early in the trial application experience, it was noted that unaided decision makers tended to allow consideration of option value to dominate their thinking. The decision aids, of course, force explicit treatment of both value and probability to derive expected values for options.

4. The benefit of numerical expression of uncertainty (required by the decision models) as opposed to verbal qualifiers, was again evident. In reviewing the details of a staff recommendation developed via the decision aid, a senior commanding officer, noting probabilistic estimates by his intelligence staff of the likelihood of significant events, reported that it was the first time he realized that his intelligence staff regarded one of the events as much more likely than the other and exactly opposite to his prior belief.

5. The structure and explicit information requirements of the decision models appeared to facilitate coordinated, efficient action on the part of staff elements involved in a given decision. The model made the information required of each staff element quite specific and, hence, more directly and efficiently addressed and communicated.

On the other hand, a number of deficiencies and needs became apparent during the trial application experience:

1. Even though the models were designed to assist users in structuring decision problems, it was found that this aspect of decision modeling remained troublesome. About one-third of the user population would not or could not correctly structure decision problems. The reasons for this are not immediately apparent, and there is a serious need for further research on problem structuring and option generation.

2. Many users voiced a requirement for a capability to reflect in the decision model verbal rationale for the values, probabilities and structure. Over time, users forget why and on what basis values were arrived at. This problem is particularly acute in those instances wherein the model reflects multiple staff inputs. A rationale capturing capability is needed.

3. The capability to conduct sensitivity tests is an important feature of the decision aids under study. On both the OPINT and EVAL models, this feature was restricted to manipulation of values for a single variable at a time and results were

displayed in a matrix format. Users expressed a need for a simultaneous multivariate sensivity testing capability and for a graphical display of results.

Freedy, et. al. (1976) presents a report detailing an evaluation of a highly promising computer-based decision aid which employs adaptive "learning" techniques to capture the decision makers utilities and to suggest courses of action consistent with those values. The decision aid termed ADDAM (Adaptive Dynamic Decision Aiding Methodology) continuously observes both the decision maker's choice behavior and the decision environment, "learns" his decision policy and offers decision suggestions based on the apparent value of the alternatives to the decision makers.

As currently configured, ADDAM is used to assist operator performance in two related decision tasks: deciding on a means of information acquisition and, on the basis of acquired information, deciding on a course of action. In information acquisition, ADDAM infers the operator's utility structure, combines the utilities with estimates of information availability and recommends the information source with the highest expected utility. In the action-selection task, aiding is provided by a Bayesian probability updating program.

The adaptive decision aid was tested in a simulated anti-submarine warfare exercise involving deployment of sensors of various types and different levels of reliability to track a submarine. In the simulation, points were awarded for correct submarine location reports, and penalties were deducted for incorrect reports. A cost factor was also introduced depending on the sensor resources allocated. The quantitative elements served as the basis for an objective game score, where the score was defined as: points-penalties-cost. Results of the tests indicated that the decision aid improved mean performance scores by about 88% over unaided trials, and decision consistency was significantly enhanced for those using the decision aid.

4.4 Decision Analysis in Real and Simulated Environments

Decision analysis brings all of the elements of statistical decision theory to bear on complex problems. Bayesian inference models are used to predict the consequences of decisions, utility models are used to evaluate these consequences, and selection of an alternative is based on the principle of expected utility maximization. In addition, the probability models used in decision analysis frequently take the form of complex Monte Carlo simulations. The initial impetus for this approach to decision making came from men interested in the analysis of corporate decisions under risk. As Howard Raiffa, Robert Schlaifer, and their colleagues at the Harvard Business School became actively involved in practical business decision-making problems, they became increasingly convinced of the merits of the Bayesian interpretation of probability, and of the need to solve complex problems through analysis rather than intuition (Raiffa, 1968). Students in the leading business schools are now routinely trained in the discipline of decision analysis, and excellent business oriented texts are available (for example, Raiffa, 1968; and Schlaifer, 1969). A number of articles also provide a good introduction to the decision analysis approach (Howard, 1966; v. Holstein, 1972). Moreover, numerous real-world applications of decision analysis have been carried out in corporate settings. Unfortunately, few of these analyses have been published in journals. Proprietary problems arise, for many of the decisions influence corporate outcomes in a competitive market. (But for an exceptionally interesting analysis of the decision to seed hurricanes, see Howard, Matheson, and North, 1972).

Brown (1971), however, has discussed some of the results which emerged from a project in which he attempted to assess the degree to which business managers have found decision analysis useful. His analysis is based on a survey of a large number of companies which have applied decision analysis, and on an in-depth study of four companies which have made extensive use of decision analysis. In some cases, the decision analysis approach has not lived up to expectations. Brown attributes most of the disappointing results to a failure to involve top-level decision makers in the analysis.

Companies in which middle-level managers or staff men carried out the analyses were generally less enthusiastic about the approach than those in which top-level managers were actively involved. When top decision makers were not involved, analyses often failed to solve the right problem. Important options were neglected and important value considerations left out. On the other hand, some companies routinely apply decision analysis to all important decisions; then top executives are thoroughly familiar with the approach. Not surprisingly, decision analysis has the greatest support in these companies. Thus, Brown argues that successful application of the decision analysis approach in business settings requires that key decision makers be familiar with and actively involved in such analyses.

Analysts on the staff of Decisions and Designs have conducted a number of decision analyses directed at various defense and related issues. These are reported in Brown, Kelly, Stewart, and Ulvila (1975); Brown, Peterson, and Ulvila (1975); Decisions and Designs, Incorporated (1973); and Peterson, Chinnis, and Hoblitzell (1975). Focused on topics ranging from policy regarding the export of computer technology through strategic decisions and resource-allocation considerations, these studies are logically compelling, but their recommendations are objectively unverified and unverifiable for reasons discussed previously.

A collection of articles edited by Drake, Keeney, and Morse (1972) describes the application of operations research and decision analysis to public policy decision making in the non-defense area. The articles by Ellis and Keeney (1972) and de Neufville and Keeney (1972) are particularly interesting because they adopt a fairly sophisticated approach to the evaluation of social alternatives under risk. (See the previous section of this paper for a further discussion of the utility measurement techniques utilized in these studies.) Keeney and Raiffa (1976) devote two chapters to these and other applications.

Most of the applications cited above have involved the analysis of "one-shot" decisions which had to be made at a given point in time. They did not deal with repeated or relatively routine decisions. One area of application in which decisions must be made repeatedly is that of medical treatment. Ginsberg (1969) has described the application of the decision analysis approach to patient management. His model, which deals with one class of illness--the pleural effusion syndrome, assists the physician in selecting diagnostic procedures, evaluating their outcomes, and determining a course of therapeutic action. In his work, Ginsberg was forced to confront the very difficult problem of assigning utilities to outcomes characterized by probability of death or severe disability, number of days in bed, number of days in severe pain, and so on. He adopted the tack of developing different utility models for different patients. The three persons who participated in his project assigned utilities using both the probabilistic utility models we described earlier and direct monetary bids. The degree of within-patient convergence between assessment methods was strikingly high and very encouraging. The utility assessment problem is quite difficult in medical contexts. For it does not seem practical to ask each patient to develop a utility model before one treats him. In many cases, the patient will be too sick for that. One possibility is to develop an "every person's" utility function for medical outcomes which would reflect not only the interests of the average patient but also the needs of society.

Betaque and Gorry (1971) describe a slightly less ambitious medical patient management decision analysis which, if modified to include more possibilities, could also be applied to repeated decisions. In their work, two renal experts separately developed decision-tree models for the management of renal patients. After developing the models, the doctors were presented with 28 hypothetical patient descriptions and asked to decide what would be the first stage in the treatment of the patient. These decisions were then compared with the decisions which would have been made by the models. One physician and his model agreed 26 out of 28 times. The other agreed only 22 times with his model, but stated that in two of the six disagreements, the model's decision was at least as good as his own. It is also interesting to note that the models agreed on 23 cases while the physicians agreed on only 21.

The final application of decision analysis in a repetitive decision making environment is the previously discussed tactical fighter strike allocation model--JUDGE. Although the JUDGE simulation experiments (Miller, Kaplan, and Edwards, 1967, 1969) did not involve the use of subjective probabilities--except as they were specified as parameters of the programming model--they provide what is to date the most clear-cut evidence of the superiority of decision analysis over intuitive decision making. They also clearly establish the feasibility of implementing decision analysis in on-line decision-making contexts where repeated decisions must be made. Similar research applications in realistic simulated environments are clearly called for.

5.0 CONCLUDING REMARKS

We feel that a strong research basis exists for a number of conclusions.

1. People are seriously suboptimal processors of information, particularly in probabilistic inference and complex decision situations. The use of formal decision-aiding algorithms improves performance in these contexts.

   While these conclusions are based mostly on laboratory experiments using student subjects, we have little doubt that they transfer to real situations and to expert, highly trained information processors and decision makers. Such evidence as is available on the behavior of experts shows them to be no better than students at probabilistic inference. Expertise, we believe, is profoundly important; neither PIP nor any other facet of decision analysis could function without the judgments of experts. But the functions of expertise are primarily those of knowing what kinds of information bear on the problem, and how. These functions are necessarily performed by people in all decision contexts.

2. A considerable body of experimental and experiential evidence supports the idea that the divide-and-conquer approach to information processing and to decision making can lead to substantially better inferences and decisions than would be obtained otherwise. Another, less pompous way of putting this conclusion would be: analysis often helps. The main reason why it helps is that

analysis permits partitioning of a complex intellectual task into components for at least some of which suitable formal tools are available. It is scarcely surprising that people, if asked to solve dynamic programming problems in their heads, do so suboptimally. The effect of analysis is to separate, for example, the task of assigning values to targets from the dynamic programming task of dispatching aircraft so as to maximize values of targets destroyed. Once the tasks have been thus separated, it is obvious and easy to let the expert judge the target value, and to let the computer do the dynamic programming.

3. Experience with applications of decision-analytic tools, though still somewhat scanty and even more scantily reported, seems to bear out the expectation, based on laboratory studies and plausibility arguments, that these tools are practical and valuable decision aids. But the experience also shows some clear technical and conceptual traps for the unwary--and it seems unlikely that all significant traps have been found and marked.

4. The technology based on multi-attribute utility theory is as exciting and promising now, in 1978, as the Bayesian technology was over a decade ago--and as much in need of further development. Multi-attribute utility measurement techniques seem to do a good job of representing preferences, especially in riskless situations. Among the unsolved problems are time preferences, risk preferences, and best elicitation techniques.

5. The approaches discussed in this paper seem most impressive in dealing with rather discrete, slow,

important problems. Although continuous models for inference and decision exist, most of the interesting experimental and practical results deal with highly discrete cases, and the technology seems far simpler for such cases. One obvious consequence of this conclusion is that decision-theoretical tools are at their most promising when applied to inference and decision problems, not to continuous control tasks.

6. Most of the most persuasive studies bearing on the merits of decision-analytic approaches are simulation studies set in complex but controlled environments. However, these studies have seldom been simulations of specific existing or contemplated systems. Instead, they have been simulations that captured the essential features of a real inference or decision task without attempting detailed, high-fidelity simulation of an actual task environment. They differ from traditional laboratory experiments primarily in making a serious attempt to capture the complex, untidy, redundant messiness of the real world and of real tasks in it.

Research on flat maxima in decision analysis has strong implications for elicitation technology. In both probabilistic inference and multi-attribute utility, there is a clear conflict between elicitation techniques that emerge from axiomatic formulations of the underlying models and thus guarantee the appropriateness of those models, and much simpler elicitation techniques (typically direct-estimation or rating-scale techniques) that have far less intimate connections with axiom systems but that seem far more practical for harried decision makers. No one has done (yet) the obvious and necessary comparison studies, especially in the utility field. But the general flatness of decision-theoretical maxima invite the hypothesis that subtleties of elicitation technology may often be beside the point; ball-park accuracy may be all that is required. If so, the simpler, easier-to-use techniques clearly win. But the research basis for this conclusion is by no means nailed down as yet; it should be.

In the area of multi-attribute utility, the most urgent research need obviously is some external standard of value with which to compare values subjectively elicited by various techniques and under various circumstances. A conclusion that different techniques produce different results is far less useful than one which adds that the result of technique A is consistently much closer to objective correctness than the result of technique B. The preceding sentence will raise the hackles of those who take seriously the idea that tastes and preferences are ultimately subjective, so that no external prescription of them can ever be appropriate. Our own view is that tastes and preferences are neither more nor less subjective than probabilities. Both are subjective quantities with external referents, and sometimes those external referents should be definitive, sometimes not. The situations in which definitive external referents exist are obviously convenient for experiments, and equally obviously unlikely to occur in application. Discovery of such situations for utility is a deeply felt research need--and, in our view, by no means hopeless.

A final comment concerns the relation between decision technology and experiment, model, or statistic. It should be clearly recognized that decision technology based on subjective estimates is no substitute for activities aimed at making those subjective estimates more objective. Use of estimates should never be allowed to inhibit attempts to replace those estimates with data obtained from experiments or from field investigations, or with suitable and valid models. In this sense, the use of estimates can be seen as a sort of first-aid measure--what to do till the statistician comes. But often it takes so long for the statistician to come and obtain the evidence he needs that subjective estimates, even though flawed, offer the only feasible approach. This is especially likely to be the case when the stakes are astronomical and the decisions unique.

And even after an objectively based model is available, it is an empirical question whether it performs better or worse than procedures based on subjective estimates. Instances of both kinds have been reviewed in this report. Until much more experience at comparative evaluation has accumulated, the availability of competing objective and subjective approaches should be taken as a signal for a comparative evaluation of them, rather than for an automatic assumption that either is preferable to the other.

In the next few years, we believe that the following will be the most stimulating and successful research directions for decision technology.

1. Further work on elicitation and validation of multi-attribute utility measurement.

2. Research on the simplification of MAU models; the trade-off between modeling error and assessment error.

3. Development of real-time, on-line decision aids.

4. The packaging of decision analysis--training techniques, standardization of elicitation techniques, and so on.

5. Further work on problems of interpersonal disagreement, especially about values.

6. Accumulation of further evidence bearing on the applied utility of decision-analytic technology.

7. Research on decision problem structuring and option generation.

6.0 REFERENCES

Allen, J., Buede, D., and O'Connor, M. The use of multiattribute value assessment techniques in the development of a marine corps combat readiness evaluation system (MCCRES). Decisions and Designs, Incorporated Technical Report, 1977 (in press).

Allison, G. Essence of Decision. Little, Brown, 1971.

Arrow, K. Social Choice and Individual Values. Yale, 1963.

Barclay, S. Interactive graphic aid for bayesian hierarchical inference. Decisions and Designs, Incorporated Technical Report, December 1976.

Barclay, S., Chinnis, J. O., and Minckler, R. D. Prototype projectile tracking radar evaluation. Decisions and Designs, Incorporated Technical Report 75-14, November 1975.

Barclay, S., and Peterson, C. R. Multi-attribute utility models for negotiations. Decisions and Designs, Incorporated Technical Report 76-1, March 1976.

Beach, L. R., Wise, J. A., and Barclay, S. Sample proportion and subjective probability revision. Org. Beh. Hum. Perf., 1970, 5, 183-190.

Betaque, N. E. and Gorry, G. A. Automating judgmental decision making for a serious medical problem. Management Science, 1971, 17, B421-434.

Bjorkman, M. Learning of linear functions: Comparison between a positive and negative slope. Report No. 183, Psychological Laboratories of the University of Stockholm, 1965.

Bjorkman, M. The effect of training and number of stimuli on the response variance in correlation learning. Report No. 2, Department of Psychology, University of Umea, 1968.

Bowman, E. H. Consistency and optimality in managerial decision making. Management Science, 1963, 9, 310-321.

Brehmer, B. Cognitive dependence on additive and configural cue-criterion relations. Amer. Jr. Psychol., 1969, 82, 490-503.

Brown, R. V. Marketing applications of personalist decision analysis. Marketing Science Institute, Cambridge, Massachusetts, 1971.

Brown, R. V., Kelly, C. W., Stewart, R. R., and Ulvila, J. W. The Timeliness of NATO response to an impending Warsaw Pact attack. Decisions and Designs, Incorporated Technical Report 75-7, December 1975.

Brown, R. V., Peterson, C. R., and Ulvila, J. W. An analysis of alternative mideastern oil agreements. Decisions and Designs, Incorporated Technical Report 75-6, December 1975.

Chinnis, J. O., Kelly, C. W., Minckler, R. D., and O'Connor, M. F. Single channel ground and airborne radio system (SINCGARS) evaluation model. Decisions and Designs, Incorporated Technical Report, September 1975.

Chinnis, J. O. and Peterson, C. R. Nonstationary processes and conservative inference. J. exp. Psychol., 1970, 84, 248-251.

Cyert, R. and March, J. A Behavioral Theory of the Firm. Prentice-Hall, 1963.

Dalkey, N. C. The Delphi Method: An experimental study of group opinion. RAND Memorandum RM-5888-PR, June 1969.

Dalkey, N. C. and Helmer, O. An experimental application of the Delphi Method to the use of experts. Management Science, 1963, 9.

Dawes, R. M. A case study of graduate admissions. Amer. Psychologist, 1971, 26, 180-188.

Decisions and Designs, Incorporated. Computer sales to the soviet bloc. Decisions and Designs, Incorporated Technical Report 73-4, October, 1973.

DeGroot, M. H. Optimal Statistical Decisions. McGraw-Hill, 1971.

DeNeufville, R. and Keeney, R. L. Use of decision analysis in airport development for Mexico City. In Drake, et. al. (Eds.) Analysis of Public Systems, 1972.

Domas, P. A. and Peterson, C. R. Probabilistic information processing systems: Evaluation with conditionally dependent data. Org. Beh. Hum. Perf., 1973.

Drake, A. W., Keeney, R. L., and Morse, P. M. (Eds.) Analysis of Public Systems. MIT Press, 1972.

DuCharme, W. M. Response bias explanation of conservative human inference. J. exp. Psychology, 1970, 85, 66-74.

DuCharme, W. M. and Peterson, C. R. Intuitive inference about normally distributed populations. J. exp. Psychology, 1968, 78, 269-275.

Edwards, W. Optimal strategies for seeking information: Models for statistics, choice reaction times, and human information processing. J. Math. Psychol., 1965, 2, 312-329.

Edwards, W. Conservatism in human information processing. In Kleinmuntz, B. (Ed.) Formal Representation of Human Judgment. Wiley, 1968.

Edwards, W. How to Use Multi-Attribute Utility Measurement for Social Decision Making. In IEEE Transactions on Systems, Man, and Cybernetics. March 1977. 326-340.

Edwards, W. Social utilities. In Decision and Risk Analysis: Powerful New Tools for Management, Proceedings of Sixth Triannual Symposium, June 1971, Hoboken: The Engineering Economist, 1972, 119-129.

Edwards, W., Lindman, H., and Phillips, L. D. Emerging technologies for making decisions. In New Directions in Psychology II. Holt, Rinehart, and Winston, 1965, 261-325.

Edwards, W., Phillips, L. D., Hays, W. L., and Goodman, B. C. Probabilistic information processing systems: Design and evaluation. IEEE Trans. Syst. Sci., Cybernetics, 1968, 248-265.

Edwards, W. and Seaver, D. A. Research on the technology of inference and decision. SSRI Research Report 76-7. Social Sciences Research Institute, University of Southern California, October 1976.

Eils, E. C. Effects of communication strategy and decision technology on the process and product of decision making groups. Doctoral dissertation, University of Southern California, June 1977.

Ellis, H. M. and Keeney, R. L. A rational approach for government decisions concerning air pollution. In Drake, et. al. (Eds.) Analysis of Public Systems, 1972.

Fischer, G. W. Multi-dimensional value assessment for decision making. The University of Michigan, Engineering Psychology Laboratory Technical Report 037230-2-T, June 1972a.

Fischer, G. W. Four methods for assessing multi-attribute utilities: an experimental validation. The University of Michigan, Engineering Psychology Laboratory Technical Report 037230-6-T, September 1972b.

Fischhoff, B. Decision analysis: clinical art or clinical science? Proceedings of the Sixth Conference on Subjective Probability, Utility and Decision Making. Warsaw, September 1977.

Freedy, A. Davis, K. B., Steeb, R., Samet, M. G., and Gardiner, P. C. Adaptive computer aiding in dynamic decision processes: methodology, evaluation, and applications. Perceptronics, Incorporated Technical Report PFTR 1016-76, August 1976.

Fried, L. S. and Peterson, C. R. Information seeking: Optional versus fixed stopping. J. exp. Psychology, 1969, 80, 525-529.

Gettys, C. F., Kelly, C. W., and Peterson, C. R. The best guess hypothesis in multi-stage inference. Org. Beh. Hum. Perf., 1973.

Gettys, C. F., Kelly, C. W., Peterson, C. R., Michel, C., and Steiger, J. H. Multiple-stage probabilistic information processing. Org. Beh. Hum. Perf., 1973.

Ginsberg, A. S. Decision analysis in clinical patient management with an application to the pleural-effusion syndrome. The RAND Corporation, r-751-RC/NLM, July 1971.

Goldberg, L. R. Simple models or simple processes? Some research on clinical judgments. Amer. Psychologist, 1968, 23, 483-496.

Goldberg, L. R. Man versus model of man: A rationale, plus some evidence for a method of improving on clinical references. Psychol. Bull., 1970, 73, 422-432.

Gustafson, D. H. Comparison of methodologies for predicting and explaining hospital length of stay. Doctoral Dissertation, University of Michigan, 1965.

Gustafson, D. H. Feller, I., Crane, K., and Holloway, D. C. A decision theory approach to measuring severity in illness. The University of Wisconsin, Department of Industrial Engineering, 1971.

Gustafson, D. H., Kestly, J. J., Greist, J. H., and Jenson, N. M. Initial evaluation of a subjective Bayesian diagnostic system. Health Services Research, Fall 1971, 204-213.

Gustafson, D. H., Shukla, R. K., Delbecq, A., and Wallster, G. W. A comparative study of difference in subjective likelihood estimates made by individuals, interest groups, Delphi groups, and nominal groups. Org. Beh. Hum. Perf., 1973, 9, 280-291.

Hammond, K. R. and Summers, D. A. Cognitive dependence on linear and non-linear cues. Psychol. Rev., 1965, 72, 215-234.

Hays, M. L., O'Connor, M. F., and Peterson, C. R. An application of multi-attribute utility theory: design-to-cost evaluation of the U.S. Navy's electronic warfare system. Decisions and Designs, Incorporated Technical Report 75-3, October 1975.

Hoepfl, R. T. and Huber, G. P. A study of self-explicated utility models. Behavioral Science, 1970, 15, 408-414.

Hoffman, P. J., Slovic, P., and Rorer, L. G. An analysis of variance model for the assessment of configural cue utilization in clinical judgment. Psychol. Bull., 1968, 69, 338-349.

Howard, R. A. The foundations of decision analysis. IEEE Trans. Syst. Sci. and Cybernetics, 1968, 4, 211-219a.

Howard. R. A. Decision analysis: applied decision theory. In: D. B. Hertz and J. Melese (Eds.) Proceedings of the Fourth International Conference on Operational Research, Wiley, 1968b.

Howard, R. A., Matheson, J. E., and North, D. W. The decision to seed hurricanes. Science, 1972, 176, 1191-1202.

Howell, W. C. Some principles for the design of decision systems: A review of six years of research on a command-control simulation. AMRL-TR-67-136. Aerospace Medical Research Laboratory, Wright-Patterson AFB, Ohio, 1967.

Huber, G. P., Daneshgar, R., and Ford, D. L. An empirical comparison of five utility models for predicting job preferences. Org. Beh. Hum. Perf., 1971, 6, 267-282.

Huber, G. P., and Delbecq, A. Guidelines for combining the judgments of individual members in decision conferences. Academy of Management Journal, 1972, 161-174.

Kahneman, D. K. and Tversky, A. Subjective probability: A judgment of representativeness. Cognitive Psychology, 1972, 3, 450-454.

Kaplan, R. J., and Newman, J. R. Studies in probabilistic information processing. IEEE Trans. Hum. Fact. Electronics, 1966, 7, 49-63.

Keeney, R. L. An illustrated procedure for assessing multiattributed utility functions. Sloan Management Review, 1972, 14, 37-49. (a)

Keeney, R. L. Multiplicative utility functions. Technical Report No. 70, MIT Operations Research Center, 1972 (b).

Keeney, R. L., and Raiffa, H. Decisions with Multiple Objectives, John Wiley and Sons, 1976.

Kelly, C. W. Application of Bayesian procedures to hierarchical inference. Doctoral Dissertation, The University of Michigan, 1972.

Kelly, C. W. Innovations in intelligence production. Report to the Commission on the Organization of the Government for the Conduct of Foreign Policy. Decisions and Designs, Incorporated, 1976.

Kelly, C. W. and Peterson, C. R. Decision theory research. Decisions and Designs, Incorporated Technical Report 75-5, September 1975.

Kelly, C. W. and Peterson, C. R. Probability estimates and probabilistic procedures in current intelligence analysis. Report on Phase I. Federal Systems Division, IBM, FSC 71-5047, 1971.

Klahr, D. Decision making in a complex environment: the use of similarity judgments to predict preferences. Management Science, 1969, 15, 595-618.

Krantz, D. H., Luce, R. D., Suppes, R., and Tversky, A. Foundations of Measurement: Additive and Polynomial Representation, I, Acadamic Press, 1971.

Krantz, D. H. and Tversky, A. Conjoint measurement analysis of composition rules in psychology. Psychol. Rev., 1971, 78, 151-169.

Kunreuther, H. Limited knowledge and insurance protection. Public Policy, 1976, 24, 227-261.

Lehmann, D. R. Television show preferences: Application of a choice model. J. Marketing Research, 1971, 8, 47-55.

Lichtenstein, S. Conditional non-independence of data in a practical Bayesian decision task. Org. Beh. Hum. Perf., 1972, 8, 21-25.

Lichtenstein, S. and Feeney, G. J. The importance of the data generating model in probability estimation. Org. Beh. Hum. Perf., 1968, 3, 62-67.

Lichtenstein, S. and Slovic, P. Reversals of preferences between bids and choices in gambling decisions. J. exp. Psychol., 1971, 89, 46-55.

Luce, R. D. and Raiffa, H. Games and Decisions. Wiley, 1957.

Luce, R. D. and Suppes, P. Preference, utility, and subjective probability. In Luce, et. al. (Eds.), Handbook of Mathematical Psychology, III. Wiley, 1965.

Lusted, L. Introduction to Medical Decision Making. C. C. Thomas, Springield, 1968.

March, J. and Simon, H. Organizations. Wiley, 1958.

Matheson, J. E. Decision analysis practice: Examples and insights. Stanford Research Institute, 1969.

Meehl, P. Clinical Versus Statistical Prediction. University of Minnesota Press, 1954.

Miller, L., Kaplan, R. J. and Edwards, W. JUDGE: A value-judgment-based tactical command system. Org. Beh. Hum. Perf., 1967, 2, 329-374.

Miller, L., Kaplan, R. J. and Edwards, W. JUDGE: A laboratory evaluation. Org. Beh. Hum. Perf., 1969, 4 97-111.

Moskowitz, H. Conservatism in group information processing behavior under varying management information systems. Paper No. 333, Krannert School of Industrial Administration, Purdue, 1971.

Moskowitz, H. The value of information in aggregate production planning. Paper No. 347, Krannert School of Industrial Administration, Purdue, April 1972.

Moskowitz, H. R&D managers choices of development policies in simulated R&D environments. IEEE Trans. Eng. Management, 1972, 19, 22-30.

Murphy, A. H. and Winker, R. L. Subjective probability forecasting of temperature: some experimental results. Preprint volume, Third Conference on Probability and Statistics in Atmospheric Science, American Meterological Society, Boulder, Colorado, 1973.

Naylor, J. C. and Clark, R. D. Intuitive inference strategies in interval learning tasks as a function of validity magnitude and sign. Org. Beh. Hum. Perf., 1968, 3, 47-61.

O'Connor, M. F. The application of multi-attribute scaling techniques to the development of indices of water quality. Doctoral Dissertation, The University of Michigan, 1972.

O'Connor, M. F., Rhees, T. R., and Allen, J. J. A multi-attribute utility approach for evaluating alternative naval aviation plans. Decisions and Designs, Incorporated Technical Report 76-16, September 1976.

Oskamp, S. Overconfidence in case-study judgments. J. Consulting Psychol., 1965, 29, 261-265.

Pai, G. K., Gustafson, D. H., and Kiner, G. W. Comparison of three non-risk methods for determining a preference function. University of Wisconsin, January 1971.

Patrick, D. L., Bush, J. W., and Chen, M. M. Measuring levels of well-being for a health status index. Department of Community Medicine, University of California, San Diego, 1972.

Peterson, C. R. and Beach, L. R. Man as an intuitive statistician. Psychol. Bull., 1967, 68, 29-46.

Peterson, C. R., Chinnis, J. O., and Hoblitzell, C. M. A decision analytic assessment of the value of information in strategic policy decisions. Decisions and Designs, Incorporated Technical Report 75-10, August 1975.

Peterson, C. R., Chinnis, J. O., and Hoblitzell, C. M. A Decision Analytic Assessment of the Value of Information in Middle East Policy Decisions. Decisions and Designs, Incorporated Technical Report 75-9, August 1975.

Peterson, C. R., DuCharme, W., and Edwards, W. Sampling distributions and probability revisions. J. exp. Psychol., 1968, 76, 236-243.

Peterson, C. R., Hammond, K. and Summers, D. A. Multiple probability learning with shifting cue weights. Amer. J. Psychol., 1965, 78, 660-663.

Peterson, C. R., Snapper, K. J. and Murphy, A. H. Credible intervals for temperature forecasting. Bull. Amer. Meteorological Society, 1972, 53, 966-972.

Phillips, L. D. Some components of prcbabilistic inference. University of Michigan, Human Performance Center Technical Report No. 1, 1966.

Phillips, L. D. and Edwards, W. Conservatism in a simple probability inference task. *J. exp. Psychol.*, 1966, *72*, 346-357.

Phillips, L. D., Hays, W. L., and Edwards, W. Conservatism in complex probabilistic inference. *IEEE Trans. Hum. Fact. Electronics*, 1966, 7, 7-18.

Pitz, G. F., Downing, L. and Reinhold, H. Sequential effects in the revision of subjective probabilities. *Canadian J. Psychol.*, 1967, *21*, 381-393.

Pitz, G. F. and Reinhold, H. Payoff effects in sequential decision making. *J. exp. Psychol.*, 1968, *77*, 249-257.

Pitz, G. F., Reinhold, H. and Geller, E. S. Strategies of information seeking in deferred decision making. *Org. Beh. Hum. Perf.*, 1968, *4*, 1-19.

Pollack, I. Action selection and the Yntema-Torgerson worth function. In E. Bennet (Ed.), *Information Systems Science and Engineering: Proceedings of the First Congress on the Information Systems Sciences*, McGraw-Hill, 1964.

Raiffa, H. *Decision Analysis*. Addison-Wesley, 1968.

Raiffa, H. Preferences for multi-attributed alternatives. RAND Memorandum RM-5868-DOT/RC, April 1968.

Rapoport, A. Sequential decision making in a computer-controlled task. *J. Math. Psychol.*, 1964, *1*, 351-374.

Rapoport, A. Variables affecting decisions in a multi-stage inventory task. L. L. Thurstone Laboratory, University of North Carolina Report No. 49, 1966.

Rapoport, A. Dynamic programming models for multi-stage decision making tasks. *J. Math. Psychol.*, 1967, *4*, 48-71.

Rapoport, A. Effects of information cost on sequential search behavior. Report No. 75, L. L. Thurstone Laboratory, University of North Carolina, 1969.

Rapoport, A. and Calder, B. J. Are inventory decisions optimal? Report No. 101, L. L. Thurstone Laboratory, University of North Carolina, 1971.

Rapoport, A. Lissitz, R. W. and McAllister, H. A. Search behavior with and without optional stopping. Report No. 89, L. L. Thurstone Laboratory, University of North Carolina, 1970.

Rapoport, A. and Wallsten, T. S. Individual decision behavior. Annual Review of Psychology, 1972, 23, 131-176.

Robinson, G. H. Continuous estimation of a time-varying probability. Ergonomics, 1964, 7, 7-21.

Schlaiffer, R. S. Introduction to Decision Making Under Uncertainty, McGraw-Hill, 1969.

Schum, D. A. Inferences on the basis of conditionally non-independent data. J. exp. Psychol., 1966, 72, 401-409.

Schum, D. A., DuCharme, W. M., DePitts, K. E. Research on human multi-stage inference processes. Rice University, Applied Mathematics and Systems Theory, No. 46-11, 1971.

Schum, D. A., Goldstein, I. L., Howell, W. C., and Southard, J. F. Subjective probability revisions under cost-payoff arrangements. Org. Beh. Hum. Per., 1967, 2, 84-104.

Schum, D. A., Southard, J., and Wombolt, L. Aided human processing of inconclusive evidence in diagnostic systems: a summary of experimental evaluations. AMRL-TR-69-11, Aerospace Medical Research Laboratory, Wright-Patterson AFB, 1969.

Schweitzer, N. Bayesian analysis for intelligence: some focus on the Middle East. Paper presented at the International Studies Association, Toronto, Canada, 1976.

Seaver, D. A. How groups can assess uncertainty: human interaction versus mathematical models. Social Sciences Research Institute, University of Southern California, 1977.

Selvidge, J. Rapid Screening of decision options. Decisions and Designs, Incorporated Technical Report 76-12, October 1976.

Shanteau, J. Descriptive versus normative models of sequential inference judgment. Kansas State University, 1971.

Shanteau, J. and Anderson, N. H. Integration theory applied to judgments of the value of information. Kansas State University, 1971.

Shepard, R. N. On subjectively optimum selection among multi-attribute alternatives. In M. W. Shelly and G. L. Bryan (Eds.) *Human Judgments and Optimality*. Wiley, 1964.

Simon, H. A. and Newell, A. Human problem solving: The state of the theory in 1970. *Amer. Psychologist*, 1971, *26*, 145-159.

Slovic, P. Analyzing the expert judge: a descriptive study of a stockbroker's decision processes. *J. Applied Psychol.*, 1969, *53*, 255-263.

Slovic, P. Psychological study of human judgment: Implications for investment decision making. *Oregon Research Institute Monograph*, *11*, September 1971.

Slovic, P, Fischhoff, B., and Lichtenstein, S. Behavioral Decision Theory. *Annual Review of Psychology*, 1977, *28*, 1-39.

Slovic, P., Fischhoff, B., Lichtenstein, S., Corrigan, B., and Combs, B. Preferences for insuring against probable small losses. *The Journal of Risk and Insurance*, 1977, 44, 2, 237-258.

Slovic, P. and Lichtenstein. Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Org. Beh. Hum. Perf.*, 1971, *6*, 649-744.

Smedlund, J. *Multiple-probability learning*. Oslo: Akaderisk Forlap, 1955.

Spetzler, C. S. The development of a corporate risk policy for capital investment decisions. *IEEE Trans. Sci. Cybernetics*, 1968, *4*, 279-300.

Spetzler, C. and Stael von Holstein, C. A. Probability encoding in decision analysis. Paper presented at ORSA-TIMS-AIEEE 1972, Joint National Meeting, Atlantic City, New Jersey, 8-10 November 1972.

Steiger, J. H. and Gettys, C. F. The best guess error in multi-stage inference. *J. exp. Psychol.*, 1972, *92* 1-7.

Stael von Holstein, C. A. Measurement of subjective probability, *Acta. Psychologica*, 1970, *34*, 146-159.

Stael von Holstein, C. A. An experiment in probabilistic weather forecasting. *J. Applied Meteorology*, 1971, *10*, 635-645.

Stael von Holstein, C. A. Probabilistic forecasting: an experiment related to the stock market. *Org. Beh. Hum. Perf.*, 1972, *8*, 139-158.

Stewart, R. R. Computer-based analytical aids program for intelligence analysts. Decisions and Designs, Incorporated interim progress report, January 1977.

Stewart, R. R., Chinnis, J. O., Kelly, C. W., and Peterson, C. R. Development of hierarchical inference software modules for indications and warning analysis. Decisions and Designs, Incorporated Technical Report. December 1976.

Tversky, A. Additivity, utility, and subjective probability. *J. Math. Psychol.*, 1967, *4*, 175-202.

Tversky, A. Intranstivity of preferences. *Psychol. Rev.* 1969, *76*, 31-48.

Tversky, A. and Kahneman, D. The judgment of probability by retrieval and construction of instances. *Oregon Research Institute Bulletin*, 1971.

Tversky, A. and Kahneman, D. The belief in the "law of small numbers." *Psychol. Bull.*, 1971, *76*, 105-110.

Ulvila, J. A pilot survey of computer programs for decision analysis. Decisions and Designs, Incorporated Technical Report 75-2, January 1975.

von Winterfeldt, D. and Edwards, W. Costs and payoffs in perceptual research. University of Michigan, Engineering Psychology Laboratory Technical Report 011313-1-T, 1973.

von Winterfeldt, D. and Edwards, W. Evaluation of complex stimuli using multi-attribute utility procedures. University of Michigan, Engineering Psychology Laboratory Technical Report 011313-2-T, 1973.

von Winterfeldt, D. and Fischer, G. Multi-attribute utility theory: models and assessment procedures. University of Michigan, Engineering Psychology Laboratory Technical Report 011313-7-T, 1973.

Wallsten, T. S. Probabilistic information processing, Bayes' rule, and conjoint measurement. Report 98, Thurstone Laboratory, University of North Carolina, 1971.

Wendt, D. Value of information for decisions. *J. Math. Psychol.*, 1969, *6*, 430-334.

Wheeler, G. Misaggregation versus response bias as explanations for conservative inference. Doctoral Dissertation, University of Michigan, 1972.

Wheeler, G. and Beach, L. R. Subjective sampling distributions and conservatism. Org. Beh. Hum. Perf., 1968, 3, 36-46.

Wiggins, N. and Kohen, E. S. Man versus model of man revisited: The forecasting of graduate school success. J. Pers. Soc. Psychol., 1971, 19, 100-106.

Winkler, R. L. The consensus of subjective probability distributions. Management Science, 1968, 15, B61-75.

Winkler, R. L. Probabilistic prediction: Some experimental results. J. Amer. Stat. Assoc., 1971, 66, 675-685.

Yntema, D. B. and Klem, L. Telling a computer how to evaluate multi-dimensional situations. IEEE Trans. Hum. Fact. Electronics, 1965, 6, 3-13.

Yntema, D. B. and Torgerson, W. S. Man-computer cooperation in decisions requiring common sense. IRE Trans. Hum. Fact. Electronics, 1961, 2, 20-26.

Youssef, Z. I. The effects of cascaded inference on the subjective value of information. Org. Beh. Hum. Perf., 1973.

Zlotnick, J. A theory for prediction. Foreign Service Journal, 1968, 45, 20.

CONTRACT DISTRIBUTION LIST
(Unclassified Technical Reports)

Director 2 copies
Advanced Research Projects Agency
Attention: Program Management Office
1400 Wilson Boulevard
Arlington, Virginia 22209

Office of Naval Research 3 copies
Attention: Code 455
800 North Quincy Street
Arlington, Virginia 22217

Defense Documentation Center 12 copies
Attention: DDC-TC
Cameron Station
Alexandria, Virginia 22314

DCASMA Baltimore Office 1 copy
Attention: Mr. K. Gerasim
300 East Joppa Road
Towson, Maryland 21204

Director 6 copies
Naval Research Laboratory
Attention: Code 2627
Washington, D.C. 20375

Office of Naval Research 6 copies
Attention: Code 102IP
800 North Quincy Street
Arlington, Virginia 22217

SUPPLEMENTAL DISTRIBUTION LIST
(Unclassified Technical Reports)

Department of Defense

Director of Net Assessment
Office of the Secretary of Defense
Attention: MAJ Robert G. Gough, USAF
The Pentagon, Room 3A930
Washington, DC 20301

Assistant Director (Net Technical Assessment)
Office of the Deputy Director of Defense
Research and Engineering (Test and Evaluation)
The Pentagon, Room 3C125
Washington, DC 20301

Assistant Director (Environmental and Life Sciences)
Office of the Deputy Director of Defense
Research and Engineering (Research and Advanced Technology)
Attention: COL Henry L. Taylor
The Pentagon, Room 3D129
Washington, DC 20301

Director, Defense Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, VA 22209

Director, Cybernetics Technology Office
Defense Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, VA 22209

Director, ARPA Regional Office (Europe)
Headquarters, U.S. European Command
APO New York 09128

Director, ARPA Regional Office (Pacific)
Staff CINCPAC, Box 13
APO San Francisco 96610

Dr. Don Hirta
Defense Systems Management School
Building 202
Ft. Belvoir, VA 22060

Chairman, Department of Curriculum Development
National War College
Ft. McNair, 4th and P Streets, SW
Washington, DC 20319

Defense Intelligence School
Attention: Professor Douglas E. Hunter
Washington, DC 20374

Vice Director for Production
Management Office (Special Actions)
Defense Intelligence Agency
Room 1E863, The Pentagon
Washington, DC 20301

Command and Control Technical Center
Defense Communications Agency
Attention: Mr. John D. Hwang
Washington, DC 20301

Department of the Navy

Office of the Chief of Naval Operations (OP-951)
Washington, DC 20450

Office of Naval Research
Assistant Chief for Technology (Code 200)
800 N. Quincy Street
Arlington, VA 22217

Office of Naval Research (Code 230)
800 North Quincy Street
Arlington, VA 22217

Office of Naval Research
Naval Analysis Programs (Code 431)
800 North Quincy Street
Arlington, VA 22217

Office of Naval Research
Operations Research Programs (Code 434)
800 North Quincy Street
Arlington, VA 22217

Office of Naval Research
Information Systems Program (Code 437)
800 North Quincy Street
Arlington, VA 22217

Director, ONR Branch Office
Attention: Dr. Charles Davis
536 South Clark Street
Chicago, IL 60605

Director, ONR Branch Office
Attention: Dr. J. Lester
495 Summer Street
Boston, MA 02210

Director, ONR Branch Office
Attention: Dr. E. Gloye
1030 East Green Street
Pasadena, CA 91106

Director, ONR Branch Office
Attention: Mr. R. Lawson
1030 East Green Street
Pasadena, CA 91106

Office of Naval Research
Scientific Liaison Group
Attention: Dr. M. Bertin
American Embassy - Room A-407
APO San Francisco 96503

Dr. A. L. Slafkosky
Scientific Advisor
Commandant of the Marine Corps (Code RD-1)
Washington, DC 20380

Headquarters, Naval Material Command
(Code 0331)
Attention: Dr. Heber G. Moore
Washington, DC 20360

Head, Human Factors Division
Naval Electronics Laboratory Center
Attention: Mr. Richard Coburn
San Diego, CA 92152

Dean of Research Administration
Naval Postgraduate School
Attention: Patrick C. Parker
Monterey, CA 93940

Superintendent
Naval Postgraduate School
Attention: R. J. Roland, (Code 52R1)
C^3 Curriculum
Monterey, CA 93940

Naval Personnel Research and Development
Center (Code 305)
Attention: LCDR O'Bar
San Diego, CA 92152

Navy Personnel Research and Development
Center
Manned Systems Design (Code 311)
Attention: Dr. Fred Muckler
San Diego, CA 92152

Naval Training Equipment Center
Human Factors Department (Code N215)
Orlando, FL 32813

Naval Training Equipment Center
Training Analysis and Evaluation Group
(Code N-00T)
Attention: Dr. Alfred F. Smode
Orlando, FL 32813

Director, Center for Advanced Research
Naval War College
Attention: Professor C. Lewis
Newport, RI 02840

Naval Research Laboratory
Communications Sciences Division (Code 5403)
Attention: Dr. John Shore
Washington, DC 20375

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

Chief, Intelligence Division
Marine Corps Development Center
Quantico, VA 22134

Department of the Army

Alan H. Curry
Operations and Management Science Division
U.S. Army Institute for Research in Management Information and Computer Science
730 Peachtree St., N.E. (Suite 900)
Atlanta, Georgia 30308

Deputy Under Secretary of the Army
(Operations Research)
The Pentagon, Room 2E621
Washington, DC 20310

Director, Army Library
Army Studies (ASDIRS)
The Pentagon, Room 1A534
Washington, DC 20310

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

Director, Organizations and Systems
Research Laboratory
U.S. Army Institute for the Behavioral
and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

Technical Director, U.S. Army Concepts
Analysis Agency
8120 Woodmont Avenue
Bethesda, MD 20014

Director, Strategic Studies Institute
U.S. Army Combat Developments Command
Carlisle Barracks, PA 17013

Commandant, Army Logistics Management Center
Attention: DRXMC-LS-SCAD (ORSA)
Ft. Lee, VA 23801

Department of Engineering
United States Military Academy
Attention: COL A. F. Grum
West Point, NY 10996

Commanding General
Headquarters, DARCOM
Attention: DRCED - Richard Murray
5001 Eisenhower Avenue
Alexandria, VA 22333

Marine Corps Representative
U.S. Army War College
Carlisle Barracks, PA 17013

Chief, Studies and Analysis Office
Headquarters, Army Training and Doctrine
Command
Ft. Monroe, VA 23351

Commander, U.S. Army Research Office
(Durham)
Box CM, Duke Station
Durham, NC 27706

Department of the Air Force

Assistant for Requirements Development
and Acquisition Programs
Office of the Deputy Chief of Staff for
Research and Development
The Pentagon, Room 4C331
Washington, DC 20330

Air Force Office of Scientific Research
Life Sciences Directorate
Building 410, Bolling AFB
Washington, DC 20332

Commandant, Air University
Maxwell AFB, AL 36112

Chief, Systems Effectiveness Branch
Human Engineering Division
Attention: Dr. Donald A. Topmiller
Wright-Patterson AFB, OH 45433

Deputy Chief of Staff, Plans, and
Operations
Directorate of Concepts (AR/XOCCC)
Attention: Major R. Linhard
The Pentagon, Room 4D 1047
Washington, DC 20330

Director, Advanced Systems Division
(AFHRL/AS)
Attention: Dr. Gordon Eckstrand
Wright-Patterson AFB, OH 45433

Commander, Rome Air Development Center
Attention: Mr. John Atkinson
Griffis AFB
Rome, NY 13440

IRD, Rome Air Development Center
Attention: Mr. Frederic A. Dion
Griffis AFB
Rome, NY 13440

HQS Tactical Air Command
Attention: LTCOL David Dianich
Langley AFB, VA 23665

Other Government Agencies

Chief, Strategic Evaluation Center
Central Intelligence Agency
Headquarters, Room 2G24
Washington, DC 20505

Director, Center for the Study of Intelligence
Central Intelligence Agency
Attention: Mr. Dean Moor
Washington, DC 20505

Mr. Richard Heuer
Methods & Forecasting Division
Office of Regional and Political Analysis
Central Intelligence Agency
Washington, DC 20505

Office of Life Sciences
Headquarters, National Aeronautics and Space Administration
Attention: Dr. Stanley Deutsch
600 Independence Avenue
Washington, DC 20546

Other Institutions

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

Institute for Defense Analyses
Attention: Dr. Jesse Orlansky
400 Army Navy Drive
Arlington, VA 22202

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

Perceptronics, Incorporated
Attention: Dr. Amos Freedy
6271 Variel Avenue
Woodland Hills, CA 91364

Stanford University
Attention: Dr. R. A. Howard
Stanford, CA 94305

Director, Applied Psychology Unit
Medical Research Council
Attention: Dr. A. D. Baddeley
15 Chaucer Road
Cambridge, CB 2EF
England

Department of Psychology
Brunel University
Attention: Dr. Lawrence D. Phillips
Uxbridge, Middlesex UB8 3PH
England

Decision Analysis Group
Stanford Research Institute
Attention: Dr. Miley W. Merkhofer
Menlo Park, CA 94025

Decision Research
1201 Oak Street
Eugene, OR 97401

Department of Psychology
University of Washington
Attention: Dr. Lee Roy Beach
Seattle, WA 98195

Department of Electrical and Computer Engineering
University of Michigan
Attention: Professor Kan Chen
Ann Arbor, MI 94135

Department of Government and Politics
University of Maryland
Attention: Dr. Davis B. Bobrow
College Park, MD 20747

Department of Psychology
Hebrew University
Attention: Dr. Amos Tversky
Jerusalem, Israel

Dr. Andrew P. Sage
School of Engineering and Applied Science
University of Virginia
Charlottesville, VA 22901

Professor Raymond Tanter
Political Science Department
The University of Michigan
Ann Arbor, MI 48109

Professor Howard Raiffa
Morgan 302
Harvard Business School
Harvard University
Cambridge, MA 02163

Department of Psychology
University of Oklahoma
Attention: Dr. Charles Gettys
455 West Lindsey
Dale Hall Tower
Norma, OK 73069

Institute of Behavioral Science #3
University of Colorado
Attention: Dr. Kenneth Hammond
Room 201
Boulder, Colorado 80309

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

14 **REPORT DOCUMENTATION PAGE**

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
TR-78-1-30

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
6 DECISION THEORETIC AIDS FOR INFERENCE, EVALUATION, AND DECISION MAKING: A REVIEW OF RESEARCH AND EXPERIENCE.

5. TYPE OF REPORT & PERIOD COVERED
9 Technical Report,

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
C.W. Kelly
G.W. Fischer - Institute for Policy Sciences, Duke University
W. Edwards - Social Science Research Inst., USC

8. CONTRACT OR GRANT NUMBER(s)
15 N~~000~~14-76-C-~~00~~74,
F33615-73-C-4056

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Decisions and Designs, Inc.
Suite 600, 8400 Westpark Dr., P.O. Box 907
McLean, Virginia 22101

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
Defense Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, Virginia 22209

12. REPORT DATE
11 Feb [redacted] 78

13. NUMBER OF PAGES
146 12 142 p.

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
Office of Naval Research
800 North Quincy Street
Arlington, Virginia 22217

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)

10 Gregory W./Fischer, Clinton W./Kelly III
Ward/Edwards

18. SUPPLEMENTARY NOTES

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

Behavioral decision theory
Decision aiding
Decision analysis
Decision making
Decision theory
Multi-attributed utility theory

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

Over the past twenty years, there has been increasing emphasis on research concerned with human decision making abilities and with the development of formal methods to aid decision makers in reaching logically consistent choices. This broad area of research is of particular importance in national security contexts where key decision makers must resolve extremely complex decision problems characterized by uncertainty, conflicting information, and enormously high stakes.

This technical report presents a summary of major portions of the → next page

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

137 390 664

alt

literature bearing on people's ability to process information and to reach decisions. It also contains a review of laboratory and field assessments of judgmentally-based decision aiding systems embodying decision analytic concepts.

The evidence reviewed provides a strong research basis for the conclusion that unaided human judgment in complex inference and decision tasks is highly fallible. Formal algorithms (decision models) applied in these contexts typically yield better results than global human judgment. The data supporting these conclusions suggest that people are better at making simple judgments than they are at aggregating large amounts of information to form overall decisions.

Consistent with these findings, the decision aiding technologies reviewed in this report are based on principles of task disaggregation. A decision problem is divided into its relevant attributes, each of which is well within the judgmental capacities of the decision maker. People make judgments about attribute probabilities and values, and formal models are used to aggregate these judgments to arrive at a decision. A large body of experimental and experiential evidence supports the notion that this divide-and-conquer approach leads to substantially better inferences and decisions than otherwise would be obtained. This research divides naturally into two parts, one dealing with probability judgments, the other with value (utility) judgments.

Probabilistic Information Processing (PIP) systems decompose the task of probabilistic inference. People identify relevant states of the environment and information sources; they also estimate likelihood ratios linking the data sources to the environmental states. Aggregating information across data is assigned to Bayes' theorem. The literature leading to the formulation of PIP systems and evaluating their application is extensively reviewed.

The second major input to decision making is judgments of value. This requires that each possible consequence of the action alternatives being considered be assigned a single numerical value reflecting the utility of that consequence relative to all other possible consequences. Both the theory and methods for assigning utilities to complex outcomes have recently become available. The technology based on multi-attribute utility theory is exciting and promising, but still relatively in its infancy. The growing body of evidence, both published and unpublished, on development and application of this technology, as well as some of its as yet unsolved problems, is reviewed in depth.