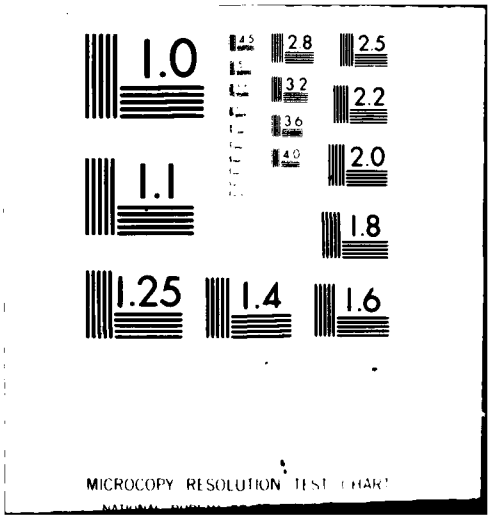
AD-A098 617 DECISIONS AND DESIGNS INC MCLEAN VA F/G 9/2
STRUCTURING AND JUDGMENT IN DECISION TECHNOLOGY.(U)
JAN 81 W EDWARDS, R S JOHN, D V WINTERFELDT MDA903-80-C-0194

UNCLASSIFIED NL

probabilistic information which must be combined. We studied a simple cascaded inference task in which one individual has information about the diagnosticity of an event and the other has information about the probability that the event has occurred (reliability information). Our main purpose was to compare two people working together with an individual combining both types of information, and to assess the relative impact of reliability and diagnosticity. Both nomothetic and idiographic analyses of the responses indicated that diagnosticity has a greater impact than reliability on judged responses. This effect was not mediated by subject sex, nor by whether the information was integrated by an individual working alone or by two people, each given either diagnosticity or reliability. Naming either the "reliability" person or the "diagnosticity" person as responsible for the group response did not alter the findings appreciably. Our results are consistent with the bulk of the subjective inference/prediction literature that suggests that subjects over-emphasize the impact of diagnostic information by not taking full account of other relevant information, such as imperfect correlation in prediction problems, and base-rate information in Bayesian inference.


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
MICROCOPY RESOLUTION TEST CHART

(12)

FINAL RESEARCH REPORT SSRI 81-4

STRUCTURING AND JUDGMENT IN DECISION TECHNOLOGY

by

Ward Edwards, Richard S. John, and Detlof v. Winterfeldt

Sponsored by
(15)
Defense Advanced Research Projects Agency
Prime Contract MDA903-80-C-0194
Under Subcontract from
Decisions and Designs, Incorporated

(11) January, 1981

Social Science Research Institute
University of Southern California
Los Angeles, California 90007
(213) 743-6955

results; these can be found in the self-contained technical reports. Instead, we briefly review the studies and results, focusing on interpretation. Our current research has led to many new and exciting ideas, and we will use this report as a means of explicating them.

II. Computer vs. Analyst

Increasingly, applyers of decision analysis are turning to computers to supplement the role of the decision analyst. For the most part, the state of the art in decision software is at a level of data storage, display, and computation as an aid to a sophisticated user. Undoubtably, these developments are useful, and we see no serious intellectual issue to be raised concerning this supplementary role of computers in decision analysis. But the next generation of decision software will surely be designed to perform a larger range of analyst functions.

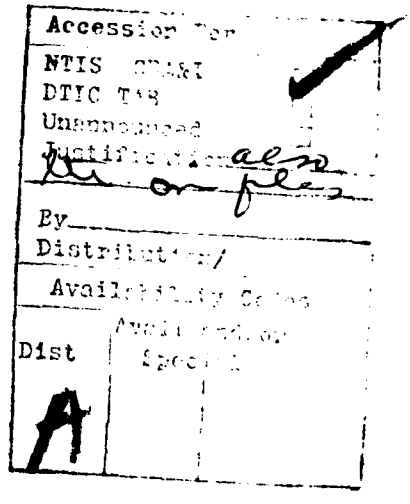
Our focus is on identifying potential problems challenging the computerization of decision analysis, and assessing the extent to which these problems can be overcome. Two problems seem particularly salient to us. First, much of what goes on in decision analysis, especially during structuring, is more accurately described as "art" than as "science". To what extent can this often ill-defined art be transformed into software? Secondly, past consumers of decision analysis have expressed satisfaction with both the process and the conclusions of analyses. To what extent is this satisfaction a function of the formal methods and procedures embodied in decision theory and the technology of decision analysis, and to what extent do other factors such as personal interaction and the establishment of a rapport account for client approval?

To answer these questions, John, v. Winterfeldt, and Edwards (see Technical Report No. 81-1, for a more complete description) examined the quality and user acceptance of simple MAU analysis performed by an analyst vs. a stand alone computer package, called MAUD (for Multi Attribute Utility Decomposition; see Humphreys and Wisudha, 1980; Hunphreys and McFadden, 1980). Unlike other packages from DARPA that we could obtain and look at, MAUD "was designed to work in direct interaction with the decision maker, without a decision analyst, counselor, or other 'expert' as intermediary (Humphreys and McFadden, 1980)."

Given a predetermined set of alternatives (at least 4 and not more than 8), MAUD guides the decision maker through a highly structured series of interactions resulting in aggregate alternative values and an implied ordering

TABLE OF CONTENTS

Acknowledgement ii
Disclaimer iii
Summary iv
I. Introduction 1
II. Computer vs. Analyst 2
III. Hierarchical vs. nonhierarchical MAU structures 6
IV. Group structure, datum diagnosticity and source reliability in hierarchical inference 8
References 12



ACKNOWLEDGEMENT

This research was supported by the Advanced Research Projects Agency of the Department of Defense, prime contract MDA 903-80-C-0194 (ARPA), and subcontracted from Decisions and Designs, Inc., #79-312-0731. The authors would like to thank Peggy Giffin, Greg Griffin, J. Robert Newman, and William Stillwell, who made many valuable contributions to the research summarized in this report.

DISCLAIMER

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 Advanced Research Projects Agency or the United States Government.

scale (rather than 9 points), and weights were obtained by direct ratio estimation of "importance". Most sessions, both analyst and MAUD, lasted between 1 and 2 hours.

Although subjects overwhelmingly yielded more favorable reports for the analyst session than for the MAUD session, subjects' agreement with and acceptance of the analyst and MAUD results (implied ordering and most preferred alternative) did not differ. Specifically, subjects indicated a desire to use the analyst rather than MAUD in future decisions and confidence that the analyst rather than MAUD "found" the best alternative by a ratio of 8 to 1. Most subjects thought that the analyst session was more helpful (5 to 1), more comfortable (4 to 1), and more effective in discovering new aspects of the problem (3 to 1) than the MAUD session. Yet, only a slim majority (54%) chose the analyst ratings over the MAUD ratings. Furthermore, only small differences were found between analyst and MAUD ratings with respect to the number of order reversals (Kendall Tau) with direct holistic ratings by the subject. Likewise, MAUD and analyst ratings did not differ substantially in the number of times they matched the most preferred alternatives. In short, although subjects reported liking analyst sessions better than MAUD sessions, no differences emerged in subjects' acceptance of resulting evaluations. These results were not substantially mediated by session order, problem type, analyst, or subject sex and race.

Two analysts tended to differ systematically in subjects' approval of their sessions and subjects' acceptance of the final ratings, as well as in the number of attributes generated. The median number of attributes elicited was greater for analyst sessions (7.5) than for MAUD sessions (5.9); however, one analyst averaged 10 attributes per session, while another averaged only a little over 5. The 10-attribute analyst was rated higher than the other four analysts in terms of subjects' impressions of the sessions, but received the lowest amount of acceptance of the resulting alternative orderings. The five-attribute analyst, however, received the lowest subjective ratings of all, but achieved the greatest degree of acceptance of final alternative orderings. Our findings seem to indicate that subjects feel better taken care of when more attributes are included in the analysis, but that subjects' holistic ratings are better accounted for by analyses with smaller rather than larger numbers of attributes. Of course, it is easy to overinterpret these results, and replication is certainly required.

SUMMARY

This report summarizes 15 months of research on the technology of inference and decision, including topics such as: the quality and user acceptance of decision analysis performed by computer vs. analyst; the effect of hierarchical vs. nonhierarchical structures on MAU importance weights and ratings; and the relative impacts of group structure, source reliability, and datum diagnosticity on hierarchical inference judgments. The purpose of this report is to summarize findings and explain how they integrate into an overall program of research on decision technology.

For the most part, the state of the art in decision software is at a level of data storage, display, and computation as an aid to a sophisticated user. Almost certainly, the next generation of decision software will be designed to perform a larger range of analyst functions. We have focused on identifying potential problems challenging the computerization of decision analysis, and on assessing the extent to which these problems can be overcome. Two questions are particularly salient: First, to what extent can the often ill-defined art of structuring be transformed into software; and secondly, to what extent is past consumers' satisfaction with decision analysis a function of the formal methods and procedures of the theory and rationale of decision theory, and to what degree do other factors such as personal interaction and the establishment of a rapport account for client approval? We compared multiattribute utility analyses of personal decision problems of undergraduates performed by a human analyst vs. those performed by a "stand-alone" software package, Multi Attribute Utility Decomposition (MAUD). Although subjects overwhelmingly yielded more favorable reports for the analyst session than for the MAUD session, subjects' agreement with and acceptance of the analyst and MAUD results (implied ordering and most perferred alternative) did not differ. We did find that subjects feel better taken care of when more attributes are included in the analysis, but that subjects' holistic ratings are better accounted for by analyses with smaller rather than larger numbers of attributes. We found that MAUD is not as "stand-alone" as its developers have advertised. In particular, our subjects needed at least some instruction in the attribute elicitation phase of the program.

Overall, our subjects became quite involved in both MAUD and analyst sessions. Subjective ratings of both sessions were greatly skewed toward the high end. Subjects were highly motivated, and their responses seemed more thoughtful and considered than is often the case for thought experiments with hypothetical scenarios, typical of laboratory experiments with college subjects.

Unfortunately, it is somewhat difficult to interpret subjects' acceptance of the resulting ordering of an analysis within this paradigm. Low acceptance could mean that the analysis has totally gone awry, or it could be indicative of a deeper, more valid evaluation than the subject is capable of in his/her holistic ratings. Regardless of the interpretation, it is important that analyst and MAUD orderings did not differ in terms of subject acceptance.

Of course, our findings cannot be interpreted in a vacuum. Proper consideration should be given to the subject population, problem types, analyst experience and method (SMART), and the particular MAUA software we employed (MAUD). In particular, we should comment on the peculiarities of the MAUD program. We found that MAUD is not as "stand alone" as its developers have advertised. In particular, our subjects needed at least some instruction in the attribute elicitation phase of the program. Typical mistakes included: repetition of attributes (up to 15 times); including more than one attribute in a given attribute definition; and thinking about other attributes when specifying the "ideal point" and/or scale values on an attribute. MAUD should give the subject more information concerning attribute elicitation, as the "difference questions" are simply too abstract and nondirective.

We also found that very few subjects are able to answer the brlts weighting question properly. In particular, most exhibit a sort of risk aversion, in which they only equate the sure thing to the gamble when the gamble odds favor the more favorable outcome. Of course, this response strategy will render the weights almost totally meaningless, since they will be solely dependent upon the order (essentially random in MAUD) in which the two varying attributes are presented. In short, the risk aversion problem with brlts may result in random weights in MAUD. To circumvent this, we intervened at the point when the subject begins the brlts portion of the program, and attempted to explain the brlts question in terms of "importance" of the two varying attributes. In particular, indifference between the sure thing and the gamble for odds of 1:1 was equated to attributes of equal importance. Odds of greater and less than 1:1 were also explained in terms of the relative importance of the attributes.

We also found that most subjects are unable to answer the brlts weighting question properly; uninstructed responses exhibit a sort of risk aversion that renders the weights virtually meaningless. Overall, subjects were highly motivated, and their responses seemed more thoughtful and considered than is often the case for thought experiments with hypothetical scenarios, typical of laboratory experiments with college subjects.

Hierarchical MAU structuring (and weighting) is a particularly attractive approach to building value trees when the number of attributes is large; partly because it reduces the number of necessary judgments, and partly because it avoids weighting questions in which only remotely related attributes need to be compared. But there are several potential problems: for example, respondents may add to an upper level value meaning not captured by its lower branches; ranges of alternatives, which can often be made explicit in specific attributes, become more vague at higher levels, perhaps distorting range dependent importance weights; furthermore, it is not clear whether numbers elicited hierarchically and non-hierarchically are consistent--if not, which should we trust more? We studied subjects' weighting and rating judgments for both hierarchical and non-hierarchical value trees relevant to evaluation of alternatives for electricity production. Hierarchical weights were found to be more variable than non-hierarchical weights, essentially replicating a result reported in 1973. We also found that subjects are often inconsistent in their attribute ratings across different levels of the value hierarchy. Random error, value lability, and misunderstanding of higher level attributes are all possible explanations for this result. Finally, we found that subjects' weight sets did not differ to any great extent as a function of their preferred alternatives; rather, location measures were the primary determinant of preference orderings. Policy decision making, at least in some highly charged arenas, may be more a matter of one's perception of how well each strategy will accomplish the stated goals, and not one sensitive to the tradeoffs among different goals.

Many important and interesting probabilistic information processing tasks are essentially hierarchical or cascaded in formal structure, and involve situations in which different people have different types of

It is important to note that all interventions into the MAUD sessions (both for attribute elicitation and brlts questions) were kept short and detached from the flow of the interaction between MAUD and the subject. Information and instructions were given only for clarifying those points necessary for the MAUD assessments. In short, all extra-MAUD interaction within the MAUD session was kept as unobtrusive as possible.

Results concerning quality of attribute sets (in terms of completeness, value independence, etc.) are included in the full technical report.

III. Hierarchical vs. nonhierarchical MAU structures

Complex evaluation problems can usually be aided by the construction of a value tree which organizes general values, intermediate objectives, and final value relevant attributes in a hierarchy. MAU models can then be built in two ways:

1) By ignoring the hierarchical structure and performing the weighting and rating tasks on lowest attributes (twigs) only;
2) By weighting branches at each level of the tree under a given node and computing final attribute weights by multiplying down the tree.

The hierarchical weighting model can, furthermore, be coupled with ratings of options on different levels of the tree to examine the internal consistency of MAU models and judgments at various levels of aggregation.

Hierarchical weighting is especially attractive when the number of attributes is very large; partly because it reduces the number of necessary judgments, and partly because it avoids weighting questions in which only remotely related attributes need to be compared. But there are also problems; for example, respondents may add to an upper level value meaning not captured by its lower branches; ranges of alternatives, which can often be made explicit in specific attributes, become more vague at higher levels, perhaps distorting range dependent importance weights; furthermore, it is not clear whether numbers elicited hierarchically and nonhierarchically are consistant -- if not, which numbers should we trust more?

To answer some of these questions Stillwell, v. Winterfeldt, and Edwards (Technical Report No. 81-2) performed an experiment in which 37 undergraduate students evaluated three scenarios for electricity production (coal vs. nuclear vs. geothermal coupled with strict conservation measures). MAU ratings and weights were elicited for 13 attributes, which a previous study had found relevant for this evaluation.

probabilistic information which must be combined. We studied a simple cascaded inference task in which one individual has information about the diagnosticity of an event and the other has information about the probability that the event has occurred (reliability information). Our main purpose was to compare two people working together with an individual combining both types of information, and to assess the relative impact of reliability and diagnosticity. Both nomothetic and idiographic analyses of the responses indicated that diagnosticity has a greater impact than reliability on judged responses. This effect was not mediated by subject sex, nor by whether the information was integrated by an individual working alone or by two people, each given either diagnosticity or reliability. Naming either the "reliability" person or the "diagnosticity" person as responsible for the group response did not alter the findings appreciably. Our results are consistent with the bulk of the subjective inference/prediction literature that suggests that subjects over-emphasize the impact of diagnostic information by not taking full account of other relevant information, such as imperfect correlation in prediction problems, and base-rate information in Bayesian inference.

These same attributes were then arranged in a hierarchical fashion, and subjects judged importance on each of three levels of the hierarchy. Upper levels are simply combinations of lower level attribute sets. Final weight for a lower level attribute (twig) is computed by multiplying weights that include that twig at each of the three levels of the hierarchy. Hierarchical weights were found to be more variable than nonhierarchical weights; this finding essentially replicates a result reported by Sayeki and Vesper in 1973.

From the standpoint of structuring, a more interesting result followed from subjects' attribute ratings of alternatives (location measures) at each level of the hierarchy. Surprisingly, subjects are often inconsistent in their ratings. For example, coal may be rated higher than nuclear on an upper level attribute, while nuclear is rated higher than coal on all of the sub-attributes making up that attribute. Three explanations seem plausible. First, subjects may simply err in expressing their values. A second possibility is that our subjects' values are extremely labile; between the time subjects were asked to rate options at the lower and higher levels, their values changed. The third explanation is that subjects imbue higher level attributes with a richer meaning than the structure below that attribute warrants, creating a sort of "super dimension" of undelineated aspects. Further research is warranted on this topic, as a case can be made for all three interpretations.

Finally, we found that subjects' weight sets did not differ to any great extent as a function of their preferred alternatives, e.g., subjects favoring the nuclear option assigned weights which were similar to those preferring coal. All groups gave the highest weight to the health/safety/environment factor. Instead of weights, location measures were the primary determinant of preference orderings among the three alternatives. This suggests that preferences are <u>not</u> determined by the extent to which one is willing to trade off one attribute for another, as has ofter been asserted. Contrarily, it appears from these data that one's perception of the alternatives' standing on the various attributes determines preference (or vice versa). Weights seem to play little role. Brickman, Shaver, and Archibald (1969) reported much the same finding in an attitude study of various foreign policies of the United States towards the Vietnam conflict. In a study on attitudes towards nuclear power, Otway, Maurer, and Thomas

I. Introduction

This final report summarizes work by the Social Science Research Institute, University of Southern California supported by the Advanced Research Projects Agency of the Department of Defense under prime contract MDA903-80-C-0194, under subcontract from Decisions and Designs, Inc. The research conducted during this contract period from November 1, 1979 to January 31, 1981, under the direction of Professor Ward Edwards, the Principle Investigator, grew out of a program of research supported by ARPA for the study of the technology of inference and decision. Edwards (1973, 1975), Edwards and Seaver (1976), Edwards, John and Stillwell (1977, 1979), and Edwards and Stillwell (1980) summarize previous research.

Our past research was concerned with the simplification of decision analysis, and, specifically, with developing and validating simple techniques for multiattribute utility analysis (MAUA). Both our laboratory and real world validation studies demonstrated that very simple rating and ranking methods perform just as well as more complicated techniques. These validation studies, combined with earlier sensitivity analysis, led us to conclude that within a given structure the precise methods of eliciting numbers matter little. Varying problem settings and structure could, of course, have strong effects on elicited numbers and results of the analysis.

The research summarized in this proposal examined such variations in problem setting and structure. Specifically, we performed three experiments on the following topics:

(1) the quality and user acceptance of decision analysis performed by computer vs. analyst;
(2) the effect of hierarchical vs. nonhierarchical structures on MAU importance weights and ratings;
(3) the relative impacts of group structure, source reliability, and datum diagnosticity on hierarchical probabilisitic inference judgments.

Our research on these issues and problems is reported in three technical reports which are now being prepared.

The purpose of this report is to summarize our findings and to explain how they integrate into an overall program of research on decision technology. Thus, we do not report detailed descriptions of procedures and

result is not encouraging for those who view attribute ratings of alternatives as a task for "technicians" who should know the "facts", and importance weighting as the primary function of the decision maker charged with the value laden task of trading off one deserving goal for another. Policy decision making, at least in arenas as highly charged as nuclear power and Vietnam, may be more a matter of one's perception of how well each strategy will accomplish the stated goals, and not one sensitive to the tradeoffs among different goals.

This result, coupled with the finding that attribute ratings of alternatives are highly inconsistent, is disconcerting. In effect, this study suggests that the structure and labels of attributes may, to a great extent, determine the final preference ordering of the analysis.

IV. Group structure, datum diagnosticity, and source reliability in hierarchical inference

Many important and interesting probabilistic information processing tasks are essentially hierarchical or cascaded in formal structure, and involve situations in which different people have different types of probabilistic information which must be combined. Griffin and Edwards (for a more complete description see Technical Report No. 81-3) studied a simple cascaded inference task in which one individual has information about the diagnosticity of an event and the other has information about the probability that the event has occurred (reliability information). The main purpose of this experiment was to compare two people working together when combining such information with an individual with both types of information, and to assess the relative impact of reliability and diagnosticity in both situations.

In the symmetric, two hypothesis case, reliability and diagnosticity should be equally important in determining the aggregate odds:

$$L = (L_r L_d + 1)/(L_r + L_d) , \qquad (1)$$

where L is the aggregate odds, L_r is the reliability likelihood ratio, and L_d is the diagnosticity likelihood ratio. Thus, diagnostic information at odds of 10 to 1, that has only a 2 to 1 chance in favor of being true, should be equally as convincing as diagnostic information at odds of 2 to 1, that has a 10 to 1 chance in favor of being true.

Undergraduates (31 two-person groups and 10 individuals working alone) were presented with a scenario in which a judgment had to be made about the

results; these can be found in the self-contained technical reports. Instead, we briefly review the studies and results, focusing on interpretation. Our current research has led to many new and exciting ideas, and we will use this report as a means of explicating them.

II. Computer vs. Analyst

Increasingly, applyers of decision analysis are turning to computers to supplement the role of the decision analyst. For the most part, the state of the art in decision software is at a level of data storage, display, and computation as an aid to a sophisticated user. Undoubtably, these developments are useful, and we see no serious intellectual issue to be raised concerning this supplementary role of computers in decision analysis. But the next generation of decision software will surely be designed to perform a larger range of analyst functions.

Our focus is on identifying potential problems challenging the computerization of decision analysis, and assessing the extent to which these problems can be overcome. Two problems seem particularly salient to us. First, much of what goes on in decision analysis, especially during structuring, is more accurately described as "art" than as "science". To what extent can this often ill-defined art be transformed into software? Secondly, past consumers of decision analysis have expressed satisfaction with both the process and the conclusions of analyses. To what extent is this satisfaction a function of the formal methods and procedures embodied in decision theory and the technology of decision analysis, and to what extent do other factors such as personal interaction and the establishment of a rapport account for client approval?

To answer these questions, John, v. Winterfeldt, and Edwards (see Technical Report No. 81-1, for a more complete description) examined the quality and user acceptance of simple MAU analysis performed by an analyst vs. a stand alone computer package, called MAUD (for Multi Attribute Utility Decomposition; see Humphreys and Wisudha, 1980; Hunphreys and McFadden, 1980). Unlike other packages from DARPA that we could obtain and look at, MAUD "was designed to work in direct interaction with the decision maker, without a decision analyst, counselor, or other 'expert' as intermediary (Humphreys and McFadden, 1980)."

Given a predetermined set of alternatives (at least 4 and not more than 8), MAUD guides the decision maker through a highly structured series of interactions resulting in aggregate alternative values and an implied ordering

likelihood that a job applicant will be successful, assuming that he or she is hired. The diagnosticity information was a test result with known validity; the reliability information was the odds that an unreliable tester actually reported the true versus a random test result. Subjects were asked to give odds based on 12 different L_d, L_r pairs; they were told that a monetary payoff of up to $5.00 would be given contingent on their performance. Both nomothetic and idiographic analyses of the responses indicated that diagnosticity has a greater impact than reliability on judged responses. This effect was not mediated by subject sex, nor by whether the information was integrated by an individual working alone or by two people, each given either diagnosticity or reliablity. Furthermore, naming either the "reliability" person or the "diagnosticity" person as "responsible" for the group response did not alter the findings appreciably.

Substantively, this experiment has replicated a robust finding in the subjective inference/prediction literature. Kahneman and Tversky (1973, 1979) and others have demonstrated that subjects predicting a criterion score (e.g., GPA) from a predictor score (e.g., SAT score) tend to ignore the fact that the two are not perfectly correlated. They treat all predictors as though they were equally valid, avoiding optimal modification of the predictor score by regressing it toward the mean. In effect, subjects ignore certain information about the worth or credibility of extreme diagnostic information, just as our subjects devalued the <u>explicit</u> information about diagnosticity conveyed in the reliability probability.

This finding is also consistent with the so-called "base-rate fallacy" (Kahneman and Tversky, 1973; Lyon and Slovic, 1976). According to Bayes' Theorem, a prior of 2 to 1 coupled with diagnostic information of 100 to 1 should be as compelling as a prior of 100 to 1 and diagnostic information of 2 to 1. Many studies have shown that the base-rate information will not be weighed into the posterior odds judgments to the appropriate extent (see Nisbett and Ross, 1980, for a review). In all three of these examples, subjects over-emphasized the impact of diagnostic information by not taking full account of other information: imperfect correlation in the prediction problem, base-rate information in the Bayesian inference problem, and reliability of the diagnositc information in the hierarchical inference problem.

It is important to note the critical role of stimulus presentation on these types of experiments. The present study conveyed the reliability and diag-

of the alternative set. MAUD developes a set of attributes by asking for descriptions of how the various alternatives differ. Single attribute value functions are elicited by placing each alternative on a nine point rating scale, determining an ideal point, and normalizing under an assumption of piece-wise linearity. Finally, importance weights are assessed under an additivity assumption via the basic reference lottery tickets (brlts) procedure (Keeney and Raiffa, 1976). The final aggregate values resulting from a MAUD analysis are a hybrid of riskless rating scale single attribute values and risky brlts importance weights.

Thirty-five undergraduates of sixty-seven interviewed (52% selection ratio) volunteered to undergo MAUA with both an analyst and MAUD for a choice dilemma that (1) was personally important and relevant, (2) involved four or more viable alternatives, and (3) required information that was readily accessible. The experiment is unique in that the multiattribute evaluation problems were generated by the subjects, and not by the experimenter as a thought experiment on a hypothetical scenario. Problems included choosing among majors (11), colleges to which to transfer (9), places to live (6), careers (4), travel plans (2), automobiles (1), sports activities (1), and strategies for handling a roommate difficulty (1).

After problems and alternative sets were discussed and agreed on with the experimenter (not an analyst), subjects were assigned to either the MAUD-first or analyst-first condition, and underwent the first analysis. The second analysis, either MAUD or analyst, came approximately one week after the first. Five different analysts were utilized, including two research faculty, one seventh year graduate student, and two first year graduate students. All subject-analyst assignments were made at the convenience of both parties. None of the analysts had more than cursory experience with applying MAUA for personal decision problems, and the two first-year students learned of MAU ideas only a few weeks before their involvement in the study.

Although details of procedure may have varied across analysts, all used a version of MAUA much like the simple multiattribute rating technique (SMART; Edwards, 1977). This differed from the MAUD analyses in that no procedure, however vague, was specified for obtaining attributes. Analysts used one or more of the following methods: (1) suggestion, (2) MAUD-like difference question, (3) asking "How is this particular alternative attractive?", (4) requiring the subject to find one aspect on which each alternative is attractive, (5) asking "What dimensions/attributes do you want to consider?", and (6) asking "What factors are relevant to the decision?" Also, single dimension values were assigned via a 100 point rating

nosticity information in the form of contingency tables with appropriate summary statistics. (The cover story involved an unreliable report of pass/fail on a test that was somewhat diagnostic of success/failure in a job situation.) Past research on the base-rate fallacy has shown that priors will be utilized more when they are either causal to the event hypotheses (Bar-Hillel, 1980; Tversky and Kahneman, 1980) or imparted to the subject in a concrete, trial by trial manner (Manis, Dovalina, Avis, and Cordoze, 1980) (See also Kassin, 1979). We can only speculate that such manipulations would have had similar affects on the utilization of reliability information in our experiment. Although contingency tables are more concrete than a single summary number, they are still relatively abstract (c.f., Manis, et al., 1980).

Methodologically, this experiment points out two important contrasts in the way judgment and decision researchers view their data: (1) nomothetic vs. idiographic analysis and (2) no effect vs. optimal model null hypothesis testing. (See Einhorn and Hogarth, 1981). Nomothetic analyses (exploring response patterns of group mean responses) are required when a "between subjects" design is used. However, when a "within subjects" design is employed, idiographic analysis (exploring typical response patterns of individual subject responses) is usually preferred, although either is appropriate. Both ways of looking at our data suggest that responses are based on only slightly modified diagnostic information; however, only the nomothetic gives an idea as to the specific heuristic strategy subjects actually employ. Patterns of mean responses indicate that subjects' log responses are quite close to the product of log L_d and the reliability probability. That is, subjects tend to use the reliability probability to adjust the log diagnostic odds downward. This interpretation can be distinguished from the alternative heuristic wherein subjects adjust the diagnostic probability or the diagnostic likelihood ratio with the reliability probability directly. Because these heuristic models are so highly correlated with each other and also with the optimal (modified Bayes' theorem), an idiographic analysis cannot distinguish among the four possibilities.

One of the greatest sources of confusion in the judgment literature (especially with regard to base-rates) is the difference between testing the null hypothesis that an information manipulation (e.g. reliability) had no effect, versus testing the optimal model hypothesis. The finding in our study is the usual one, namely, that the manipulation (reliability)

scale (rather than 9 points), and weights were obtained by direct ratio estimation of "importance". Most sessions, both analyst and MAUD, lasted between 1 and 2 hours.

Although subjects overwhelmingly yielded more favorable reports for the analyst session than for the MAUD session, subjects' agreement with and acceptance of the analyst and MAUD results (implied ordering and most preferred alternative) did not differ. Specifically, subjects indicated a desire to use the analyst rather than MAUD in future decisions and confidence that the analyst rather than MAUD "found" the best alternative by a ratio of 8 to 1. Most subjects thought that the analyst session was more helpful (5 to 1), more comfortable (4 to 1), and more effective in discovering new aspects of the problem (3 to 1) than the MAUD session. Yet, only a slim majority (54%) chose the analyst ratings over the MAUD ratings. Furthermore, only small differences were found between analyst and MAUD ratings with respect to the number of order reversals (Kendall Tau) with direct holistic ratings by the subject. Likewise, MAUD and analyst ratings did not differ substantially in the number of times they matched the most preferred alternatives. In short, although subjects reported liking analyst sessions better than MAUD sessions, no differences emerged in subjects' acceptance of resulting evaluations. These results were not substantially mediated by session order, problem type, analyst, or subject sex and race.

Two analysts tended to differ systematically in subjects' approval of their sessions and subjects' acceptance of the final ratings, as well as in the number of attributes generated. The median number of attributes elicited was greater for analyst sessions (7.5) than for MAUD sessions (5.9); however, one analyst averaged 10 attributes per session, while another averaged only a little over 5. The 10-attribute analyst was rated higher than the other four analysts in terms of subjects' impressions of the sessions, but received the lowest amount of acceptance of the resulting alternative orderings. The five-attribute analyst, however, received the lowest subjective ratings of all, but achieved the greatest degree of acceptance of final alternative orderings. Our findings seem to indicate that subjects feel better taken care of when more attributes are included in the analysis, but that subjects' holistic ratings are better accounted for by analyses with smaller rather than larger numbers of attributes. Of course, it is easy to overinterpret these results, and replication is certainly required.

is used in some heuristic, non-optimal manner. Invariably, the very same data (ours included) that could be used to reject the no effect hypothesis can also be used to prove that the subject is not following the optimal model. The reason is simple: subjects made some use of the manipulated information (reliability), but not optimal use. This often results in seemingly contradictory findings: reliability information had an effect on responses (no effect rejected), yet subjects' responses were not even ordinally consistent with the model, due to a neglect of reliability information. Both of these statements are true of our study; unfortunately, researchers sometimes choose to emphasize no effect rejections and ignore optimal model rejections (and vice versa), in order to support a particular theoretical position. It makes more sense to look at the data both ways, and attempt to discover the exact heuristic employed, if possible.

Overall, our subjects became quite involved in both MAUD and analyst sessions. Subjective ratings of both sessions were greatly skewed toward the high end. Subjects were highly motivated, and their responses seemed more thoughtful and considered than is often the case for thought experiments with hypothetical scenarios, typical of laboratory experiments with college subjects.

Unfortunately, it is somewhat difficult to interpret subjects' acceptance of the resulting ordering of an analysis within this paradigm. Low acceptance could mean that the analysis has totally gone awry, or it could be indicative of a deeper, more valid evaluation than the subject is capable of in his/her holistic ratings. Regardless of the interpretation, it is important that analyst and MAUD orderings did not differ in terms of subject acceptance.

Of course, our findings cannot be interpreted in a vacuum. Proper consideration should be given to the subject population, problem types, analyst experience and method (SMART), and the particular MAUA software we employed (MAUD). In particular, we should comment on the peculiarities of the MAUD program. We found that MAUD is not as "stand alone" as its developers have advertised. In particular, our subjects needed at least some instruction in the attribute elicitation phase of the program. Typical mistakes included: repetition of attributes (up to 15 times); including more than one attribute in a given attribute definition; and thinking about other attributes when specifying the "ideal point" and/or scale values on an attribute. MAUD should give the subject more information concerning attribute elicitation, as the "difference questions" are simply too abstract and nondirective.

We also found that very few subjects are able to answer the brlts weighting question properly. In particular, most exhibit a sort of risk aversion, in which they only equate the sure thing to the gamble when the gamble odds favor the more favorable outcome. Of course, this response strategy will render the weights almost totally meaningless, since they will be solely dependent upon the order (essentially random in MAUD) in which the two varying attributes are presented. In short, the risk aversion problem with brlts may result in random weights in MAUD. To circumvent this, we intervened at the point when the subject begins the brlts portion of the program, and attempted to explain the brlts question in terms of "importance" of the two varying attributes. In particular, indifference between the sure thing and the gamble for odds of 1:1 was equated to attributes of equal importance. Odds of greater and less than 1:1 were also explained in terms of the relative importance of the attributes.

References

Bar-Hillel, M. The base rate fallacy in probability judgments. Acta Psychologica, 1980, 44, 211-233.

Brickman, P., Shaver, P., & Archibald, P. American tactics and goals as perceived by social scientists. In W. Isard (Ed.), Vietnam: issues and alternatives. Cambridge, Ma.: Schenkman Publishing Company, 1969.

Edwards, W., Research on the technology of inference and decision (Tech. Rep. 011313-F). Ann Arbor, Mich.: University of Michigan, Engineering Psychology Laboratory, November, 1973.

Edwards, W., Research on the technology of inference and decision (SSRI Tech. Rep. 75-10). Los Angeles: University of Southern California, Social Science Research Institute, August, 1975.

Edwards, W., How to use multiattribute utility measurement for social decision-making. IEEE Transactions on Systems, Man, and Cybernetics, 1977, SMC-7, 326-340.

Edwards, W., John, R.S., & Stillwell, W., Research on the technology of inference and decision (SSRI Tech. Rep. 77-6). Los Angeles: University of Southern California, Social Science Research Institute, November, 1977.

Edwards, W., John, R.S., & Stillwell, W., Research on the technology of inference and decision (SSRI Tech. Rep. 79-1). Los Angeles: University of Southern California, Social Science Research Institute, January, 1979.

Edwards, W., & Seaver, D.A., Research on the technology of inference and decision (SSRI Tech. Rep. 76-7). Los Angeles: University of Southern California, Social Science Research Institute, October, 1976.

Edwards, W. & Stillwell, W., Validation, error, and simplification of decision technology (SSRI Tech. Rep. 80-5). Los Angeles: University of Southern California, Social Science Research Institute, June, 1980.

Einhorn, H. J. & Hogarth, R. M. Behavioral decision theory: processes of judgment and choice. In M. R. Rosenweig and L. W. Porter (Eds.), Annual review of psychology, vol. 32. Palo Alto, Ca.: Annual Reviews Inc., 1981.

It is important to note that all interventions into the MAUD sessions (both for attribute elicitation and brlts questions) were kept short and detached from the flow of the interaction between MAUD and the subject. Information and instructions were given only for clarifying those points necessary for the MAUD assessments. In short, all extra-MAUD interaction within the MAUD session was kept as unobtrusive as possible.

Results concerning quality of attribute sets (in terms of completeness, value independence, etc.) are included in the full technical report.

III. Hierarchical vs. nonhierarchical MAU structures

Complex evaluation problems can usually be aided by the construction of a value tree which organizes general values, intermediate objectives, and final value relevant attributes in a hierarchy. MAU models can then be built in two ways:

1) By ignoring the hierarchical structure and performing the weighting and rating tasks on lowest attributes (twigs) only;
2) By weighting branches at each level of the tree under a given node and computing final attribute weights by multiplying down the tree.

The hierarchical weighting model can, furthermore, be coupled with ratings of options on different levels of the tree to examine the internal consistency of MAU models and judgments at various levels of aggregation.

Hierarchical weighting is especially attractive when the number of attributes is very large; partly because it reduces the number of necessary judgments, and partly because it avoids weighting questions in which only remotely related attributes need to be compared. But there are also problems; for example, respondents may add to an upper level value meaning not captured by its lower branches; ranges of alternatives, which can often be made explicit in specific attributes, become more vague at higher levels, perhaps distorting range dependent importance weights; furthermore, it is not clear whether numbers elicited hierarchically and nonhierarchically are consistant -- if not, which numbers should we trust more?

To answer some of these questions Stillwell, v. Winterfeldt, and Edwards (Technical Report No. 81-2) performed an experiment in which 37 undergraduate students evaluated three scenarios for electricity production (coal vs. nuclear vs. geothermal coupled with strict conservation measures). MAU ratings and weights were elicited for 13 attributes, which a previous study had found relevant for this evaluation.

Humphreys, P., & McFadden, W. Experiences with MAUD: aiding decision structuring versus bootstrapping the decision maker. Unpublished manuscript, 1980. (Available from Decision Analysis Unit, Brunel Institute of Organization and Social Studies, Brunel University, Uxbridge, Middlesex, England.)

Humphreys, P., & Wisudha, A. Multi Attribute Utility Decomposition, (Technical report 79-2/2). Uxbridge, Middlesex, England: Brunel University, Brunel Institute of Organization and Social Studies, Decision Analysis Unit, 1980.

Kahneman, D., & Tversky, A. On the psychology of prediction. Psychological Review, 1973, 80, 237-251.

Kahneman, D., & Tversky, A. Intuitive prediction: biases and corrective procedures. TIMS Studies in Management Science, 1979, 12, 313-327.

Kassin, S.M. Consensus information, prediction, and causal attribution: a review of the literature and issues. Journal of Personality and Social Psychology, 1979, 37, 1966-1981.

Keeney, R.L., & Raiffa, H. Decisions with multiple objectives. New York: Wiley, 1976.

Lyon, D.,& Slovic, P. Dominance of accuracy information and neglect of base rates in probability estimation. Acta Psychologica, 1976, 40, 287-298.

Manis, M., Dovalina, I., Avis, N.E., & Cardoze, S. Base rates can affect individual predictions. Journal of Personality and Social Psychology, 1980, 38, 231-248.

Nisbett, R., & Ross, L. Human inference: strategies and shortcomings of social judgment. Englewood Cliffs, N.J.: Prentice-Hall, 1980.

Otway, H., Maurer, D., & Thomas, K. Nuclear power: the question of public acceptance. Futures, 1978, 10, 109-118.

Sayeki, Y., & Vesper, K.H. Allocation of importance and goal-dependent utility. Management Science, 1973, 19(6), 667-675.

Tversky, A., & Kahneman, D. Causal schemas in judgments under certainty. In M. Fishbein (Ed.), Progress in social psychology. Hillsdale, N.J.: Erlbaum, 1980.

These same attributes were then arranged in a hierarchical fashion, and subjects judged importance on each of three levels of the hierarchy. Upper levels are simply combinations of lower level attribute sets. Final weight for a lower level attribute (twig) is computed by multiplying weights that include that twig at each of the three levels of the hierarchy. Hierarchical weights were found to be more variable than nonhierarchical weights; this finding essentially replicates a result reported by Sayeki and Vesper in 1973.

From the standpoint of structuring, a more interesting result followed from subjects' attribute ratings of alternatives (location measures) at each level of the hierarchy. Surprisingly, subjects are often inconsistent in their ratings. For example, coal may be rated higher than nuclear on an upper level attribute, while nuclear is rated higher than coal on all of the sub-attributes making up that attribute. Three explanations seem plausible. First, subjects may simply err in expressing their values. A second possibility is that our subjects' values are extremely labile; between the time subjects were asked to rate options at the lower and higher levels, their values changed. The third explanation is that subjects imbue higher level attributes with a richer meaning than the structure below that attribute warrants, creating a sort of "super dimension" of undelineated aspects. Further research is warranted on this topic, as a case can be made for all three interpretations.

Finally, we found that subjects' weight sets did not differ to any great extent as a function of their preferred alternatives, e.g., subjects favoring the nuclear option assigned weights which were similar to those preferring coal. All groups gave the highest weight to the health/safety/environment factor. Instead of weights, location measures were the primary determinant of preference orderings among the three alternatives. This suggests that preferences are <u>not</u> determined by the extent to which one is willing to trade off one attribute for another, as has ofter been asserted. Contrarily, it appears from these data that one's perception of the alternatives' standing on the various attributes determines preference (or vice versa). Weights seem to play little role. Brickman, Shaver, and Archibald (1969) reported much the same finding in an attitude study of various foreign policies of the United States towards the Vietnam conflict. In a study on attitudes towards nuclear power, Otway, Maurer, and Thomas (1978) also found that attitude differences between pro-nuclear groups and anti-nuclear groups are largely due to beliefs, not to values. This

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: 81-4	2. GOVT ACCESSION NO.: AD-A098617	3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): Structuring and Judgment in Decision Technology		5. TYPE OF REPORT & PERIOD COVERED: Tech. 11/1/79 - 1/31/81
		6. PERFORMING ORG. REPORT NUMBER: 81 - 4
7. AUTHOR(s): W. Edwards, R.S. John, and D. v. Winterfeldt		8. CONTRACT OR GRANT NUMBER(s): Prime Contract MDA 903-80-C-0194 (ARPA); Subcontract 79-312-0731 (DDI)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Social Science Research Institute, University of Southern California, Los Angeles, California 90007		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS:
11. CONTROLLING OFFICE NAME AND ADDRESS: Defense Advanced Research Projects Agency, 1400 Wilson Blvd, Arlington, VA 22209		12. REPORT DATE: January, 1981
		13. NUMBER OF PAGES: 13
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report): unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE
DISTRIBUTION UNLIMITED

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

18. SUPPLEMENTARY NOTES

Decisions & Designs, Inc.
8400 Westpark Drive - Suite 600
McLean, VA 22101

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

multiattribute utility; validation; biases; elicitation; weighting; hierarchical weighting; cascaded inference; reliability; diagnosticity; group probability assessment; artificial intelligence; hierarchical inference; structuring; decision analysis; value trees; modified Bayes' Theorem; utility; probability; outcome study

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

This report summarizes 15 months of research on the technology of inference & decision, including topics such as: the quality and user accentance of decision analysis performed by computer vs. analyst; the effect of hierarchical vs. non-hierarchical structures on MAU importance weights and ratings; and the relative impacts of group structure, source reliability, and datum diagnosticity on hierarchical inference judgments. The purpose of this report is to summarize findings and explain how they integrate into an overall program of research on decision technology.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

result is not encouraging for those who view attribute ratings of alternatives as a task for "technicians" who should know the "facts", and importance weighting as the primary function of the decision maker charged with the value laden task of trading off one deserving goal for another. Policy decision making, at least in arenas as highly charged as nuclear power and Vietnam, may be more a matter of one's perception of how well each strategy will accomplish the stated goals, and not one sensitive to the tradeoffs among different goals.

This result, coupled with the finding that attribute ratings of alternatives are highly inconsistent, is disconcerting. In effect, this study suggests that the structure and labels of attributes may, to a great extent, determine the final preference ordering of the analysis.

IV. Group structure, datum diagnosticity, and source reliability in hierarchical inference

Many important and interesting probabilistic information processing tasks are essentially hierarchical or cascaded in formal structure, and involve situations in which different people have different types of probabilistic information which must be combined. Griffin and Edwards (for a more complete description see Technical Report No. 81-3) studied a simple cascaded inference task in which one individual has information about the diagnosticity of an event and the other has information about the probability that the event has occurred (reliability information). The main purpose of this experiment was to compare two people working together when combining such information with an individual with both types of information, and to assess the relative impact of reliability and diagnosticity in both situations.

In the symmetric, two hypothesis case, reliability and diagnosticity should be equally important in determining the aggregate odds:

$$L = (L_r L_d + 1)/(L_r + L_d) , \qquad (1)$$

where L is the aggregate odds, L_r is the reliability likelihood ratio, and L_d is the diagnosticity likelihood ratio. Thus, diagnostic information at odds of 10 to 1, that has only a 2 to 1 chance in favor of being true, should be equally as convincing as diagnostic information at odds of 2 to 1, that has a 10 to 1 chance in favor of being true.

Undergraduates (31 two-person groups and 10 individuals working alone) were presented with a scenario in which a judgment had to be made about the

For the most part, the state of the art in decision software is at a level of data storage, display, and computation as an aid to a sophisticated user. Almost certainly, the next generation of decision software will be designed to perform a larger range of analyst functions. We have focused on identifying potential problems challenging the computerization of decision analysis, and on assessing the extent to which these problems can be overcome. Two questions are particularly salient: First, to what extent can the often ill-defined art of structuring be transformed into software; and secondly, to what extent is past consumers' satisfaction with decision analysis a function of the formal methods and procedures of the theory and rationale of decision theory, and to what degree do other factors such as personal interaction and the establishment of a rapport account for client approval? We compared multiattribute utility analyses of personal decision problems of undergraduates performed by a human analyst vs. those performed by a "stand-alone" software package, Multi Attribute Utility Decomposition (MAUD). Although subjects overwhelmingly yielded more favorable reports for the analyst session than for the MAUD session, subjects' agreement with & acceptance of the analyst and MAUD results (implied ordering and most preferred alternative) did not differ. We did find that subjects feel better taken care of when more attributes are included in the analysis, but that subjects' holistic ratings are better accounted for by analyses with smaller rather than larger numbers of attributes. We found that MAUD is not as "stand-alone" as its developers have advertised. In particular, our subjects needed at least some instruction in the attribute elicitation phase of the program. We also found that most subjects are unable to answer the brits weighting question properly; uninstructed responses exhibit a sort of risk aversion that renders the weights virtually meaningless. Overall, subjects were highly motivated, and their responses seemed more thoughtful and considered than is often the case for thought experiments with hypothetical scenarios, typical of laboratory experiments with college subjects.

Hierarchical MAU structuring (and weighting) is a particularly attractive approach to building value trees when the number of attributes is large; partly because it reduces the number of necessary judgments, and partly because it avoids weighting questions in which only remotely related attributes need to be compared. But there are several potential problems: for example, respondents may add to an upper level value meaning not captured by its lower branches; ranges of alternatives, which can often be made explicit in specific attributes, become more vague at higher levels, perhaps distorting range dependent importance weights; furthermore, it is not clear whether numbers elciited hierarchically and non-hierarchically are consistent--if not, which should we trust more? We studied subjects' weighting and rating judgments for both hierarchical and non-hierarchical value trees relevant to evaluation of alternatives for electricity production. Hierarchical weights were found to be more variable than non-hierarchical weights, essentially replicating a result reported in 1973. We also found that subjects are often inconsistent in their attribute ratings across different levels of the value hierarchy. Random error, value lability, and misunderstanding of higher level attributes are all possible explanations for this result. Finally, we found that subjects' weight sets did not differ to any great extent as a function of their preferred alternatives; rather, location measures were the primary determinant of preference orderings. Policy decision making, at least in some highly charged arenas, may be more a matter of one's perception of how well each strategy will accomplish the stated goals, and not one sensitive to the tradeoffs among different goals.

Many important and interesting probabilistic information processing tasks are essentially hierarchical or cascaded in formal structure, and involve situations in which different people have different types of probabilistic

likelihood that a job applicant will be successful, assuming that he or she is hired. The diagnosticity information was a test result with known validity; the reliability information was the odds that an unreliable tester actually reported the true versus a random test result. Subjects were asked to give odds based on 12 different L_d, L_r pairs; they were told that a monetary payoff of up to $5.00 would be given contingent on their performance. Both nomothetic and idiographic analyses of the responses indicated that diagnosticity has a greater impact than reliability on judged responses. This effect was not mediated by subject sex, nor by whether the information was integrated by an individual working alone or by two people, each given either diagnosticity or reliablity. Furthermore, naming either the "reliability" person or the "diagnosticity" person as "responsible" for the group response did not alter the findings appreciably.

Substantively, this experiment has replicated a robust finding in the subjective inference/prediction literature. Kahneman and Tversky (1973, 1979) and others have demonstrated that subjects predicting a criterion

nosticity information in the form of contingency tables with appropriate summary statistics. (The cover story involved an unreliable report of pass/fail on a test that was somewhat diagnostic of success/failure in a job situation.) Past research on the base-rate fallacy has shown that priors will be utilized more when they are either causal to the event hypotheses (Bar-Hillel, 1980; Tversky and Kahneman, 1980) or imparted to the subject in a concrete, trial by trial manner (Manis, Dovalina, Avis, and Cordoze, 1980) (See also Kassin, 1979). We can only speculate that such manipulations would have had similar affects on the utilization of reliability information in our experiment. Although contingency tables are more concrete than a single summary number, they are still relatively abstract (c.f., Manis, et al., 1980).

Methodologically, this experiment points out two important contrasts in the way judgment and decision researchers view their data: (1) nomothetic vs. idiographic analysis and (2) no effect vs. optimal model null hypothesis testing. (See Einhorn and Hogarth, 1981). Nomothetic analyses (exploring response patterns of group mean responses) are required when a "between subjects" design is used. However, when a "within subjects" design is employed, idiographic analysis (exploring typical response patterns of individual subject responses) is usually preferred, although either is appropriate. Both ways of looking at our data suggest that responses are based on only slightly modified diagnostic information; however, only the nomothetic gives an idea as to the specific heuristic strategy subjects actually employ. Patterns of mean responses indicate that subjects' log responses are quite close to the product of log L_d and the reliability probability. That is, subjects tend to use the reliability probability to adjust the log diagnostic odds downward. This interpretation can be distinguished from the alternative heuristic wherein subjects adjust the diagnostic probability or the diagnostic likelihood ratio with the reliability probability directly. Because these heuristic models are so highly correlated with each other and also with the optimal (modified Bayes' theorem), an idiographic analysis cannot distinguish among the four possibilities.

One of the greatest sources of confusion in the judgment literature (especially with regard to base-rates) is the difference between testing the null hypothesis that an information manipulation (e.g. reliability) had no effect, versus testing the optimal model hypothesis. The finding in our study is the usual one, namely, that the manipulation (reliability)

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, D.C. 20301

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

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

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

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

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

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

Department of the Navy

Office of the Chief of Naval
Operations
(OP-951)
Washington, D.C. 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 No. Quincy Street
Arlington, VA 22217

Office of Naval Research
Naval Analysis Programs (Code 431)
800 No. Quincy Street
Arlington, VA 22217

Office of Naval Research
Operations Research Programs (Code
800 No. Quincy Street
Arlington, VA 22217

Office of Naval Research
Information Systems Program (Code
800 No. Quincy Street
Arlington, VA 22217

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

is used in some heuristic, non-optimal manner. Invariably, the very same data (ours included) that could be used to reject the no effect hypothesis can also be used to prove that the subject is not following the optimal model. The reason is simple: subjects made some use of the manipulated information (reliability), but not optimal use. This often results in seemingly contradictory findings: reliability information had an effect on responses (no effect rejected), yet subjects' responses were not even ordinally consistent with the model, due to a neglect of reliability information. Both of these statements are true of our study; unfortunately, researchers sometimes choose to emphasize no effect rejections and ignore optimal model rejections (and vice versa), in order to support a particular theoretical position. It makes more sense to look at the data both ways, and attempt to discover the exact heuristic employed, if possible.

References

Bar-Hillel, M. The base rate fallacy in probability judgments. Acta Psychologica, 1980, 44, 211-233.

Brickman, P., Shaver, P., & Archibald, P. American tactics and goals as perceived by social scientists. In W. Isard (Ed.), Vietnam: issues and alternatives. Cambridge, Ma.: Schenkman Publishing Company, 1969.

Edwards, W., Research on the technology of inference and decision (Tech. Rep. 011313-F). Ann Arbor, Mich.: University of Michigan, Engineering Psychology Laboratory, November, 1973.

Edwards, W., Research on the technology of inference and decision (SSRI Tech. Rep. 75-10). Los Angeles: University of Southern California, Social Science Research Institute, August, 1975.

Edwards, W., How to use multiattribute utility measurement for social decision-making. IEEE Transactions on Systems, Man, and Cybernetics, 1977, SMC-7, 326-340.

Edwards, W., John, R.S., & Stillwell, W., Research on the technology of inference and decision (SSRI Tech. Rep. 77-6). Los Angeles: University of Southern California, Social Science Research Institute, November, 1977.

Edwards, W., John, R.S., & Stillwell, W., Research on the technology of inference and decision (SSRI Tech. Rep. 79-1). Los Angeles: University of Southern California, Social Science Research Institute, January, 1979.

Edwards, W., & Seaver, D.A., Research on the technology of inference and decision (SSRI Tech. Rep. 76-7). Los Angeles: University of Southern California, Social Science Research Institute, October, 1976.

Edwards, W. & Stillwell, W., Validation, error, and simplification of decision technology (SSRI Tech. Rep. 80-5). Los Angeles: University of Southern California, Social Science Research Institute, June, 1980.

Einhorn, H. J. & Hogarth, R. M. Behavioral decision theory: processes of judgment and choice. In M. R. Rosenweig and L. W. Porter (Eds.), Annual review of psychology, vol. 32. Palo Alto, Ca.: Annual Reviews Inc., 1981.

Humphreys, P., & McFadden, W. Experiences with MAUD: aiding decision structuring versus bootstrapping the decision maker. Unpublished manuscript, 1980. (Available from Decision Analysis Unit, Brunel Institute of Organization and Social Studies, Brunel University, Uxbridge, Middlesex, England.)

Humphreys, P., & Wisudha, A. Multi Attribute Utility Decomposition, (Technical report 79-2/2). Uxbridge, Middlesex, England: Brunel University, Brunel Institute of Organization and Social Studies, Decision Analysis Unit, 1980.

Kahneman, D., & Tversky, A. On the psychology of prediction. Psychological Review, 1973, 80, 237-251.

Kahneman, D., & Tversky, A. Intuitive prediction: biases and corrective procedures. TIMS Studies in Management Science, 1979, 12, 313-327.

Kassin, S.M. Consensus information, prediction, and causal attribution: a review of the literature and issues. Journal of Personality and Social Psychology, 1979, 37, 1966-1981.

Keeney, R.L., & Raiffa, H. Decisions with multiple objectives. New York: Wiley, 1976.

Lyon, D.,& Slovic, P. Dominance of accuracy information and neglect of base rates in probability estimation. Acta Psychologica, 1976, 40, 287-298.

Manis, M., Dovalina, I., Avis, N.E., & Cardoze, S. Base rates can affect individual predictions. Journal of Personality and Social Psychology, 1980, 38, 231-248.

Nisbett, R., & Ross, L. Human inference: strategies and shortcomings of social judgment. Englewood Cliffs, N.J.: Prentice-Hall, 1980.

Otway, H., Maurer, D., & Thomas, K. Nuclear power: the question of public acceptance. Futures, 1978, 10, 109-118.

Sayeki, Y., & Vesper, K.H. Allocation of importance and goal-dependent utility. Management Science, 1973, 19(6), 667-675.

Tversky, A., & Kahneman, D. Causal schemas in judgments under certainty. In M. Fishbein (Ed.), Progress in social psychology. Hillsdale, N.J.: Erlbaum, 1980.

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

Other Government Agencies

Chief, Strategic Evaluation Center
Central Intelligence Agency
Headquarters, Room 2G24
Washington, D.C. 20505

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

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

Other Institutions

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

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

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

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

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

Dr. Andrew P. Sage
School of Engineering and Applied Science
University of Virginia
Charlottesville, VA 22903

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
Norman, OK 73069

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

Decisions and Designs, Incorporated
Suite 600, 8400 Westpark Drive
P.O. Box 907
McLean, VA 22101

Decision Science Consortium, Inc.
Suite 421
7700 Leesburg Pike
Falls Church, VA 22043

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: 81-4	2. GOVT ACCESSION NO.: AD-A098617	3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): Structuring and Judgment in Decision Technology		5. TYPE OF REPORT & PERIOD COVERED: Tech. 11/1/79 - 1/31/81
		6. PERFORMING ORG. REPORT NUMBER: 81 - 4
7. AUTHOR(s): W. Edwards, R.S. John, and D. v. Winterfeldt		8. CONTRACT OR GRANT NUMBER(s): Prime Contract MDA 903-80-C-0194 (ARPA); Subcontract 79-312-0731 (DDI)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Social Science Research Institute, University of Southern California, Los Angeles, California 90007		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS:
11. CONTROLLING OFFICE NAME AND ADDRESS: Defense Advanced Research Projects Agency, 1400 Wilson Blvd, Arlington, VA 22209		12. REPORT DATE: January, 1981
		13. NUMBER OF PAGES: 13
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report): unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE
DISTRIBUTION UNLIMITED

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

18. SUPPLEMENTARY NOTES

Decisions & Designs, Inc.
8400 Westpark Drive - Suite 600
McLean, VA 22101

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

multiattribute utility; validation; biases; elicitation; weighting; hierarchical weighting; cascaded inference; reliability; diagnosticity; group probability assessment; artificial intelligence; hierarchical inference; structuring; decision analysis; value trees; modified Bayes' Theorem; utility; probability; outcome study

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

This report summarizes 15 months of research on the technology of inference & decision, including topics such as: the quality and user accentance of decision analysis performed by computer vs. analyst; the effect of hierarchical vs. non-hierarchical structures on MAU importance weights and ratings; and the relative impacts of group structure, source reliability, and datum diagnosticity on hierarchical inference judgments. The purpose of this report is to summarize findings and explain how they integrate into an overall program of research on decision technology.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

For the most part, the state of the art in decision software is at a level of data storage, display, and computation as an aid to a sophisticated user. Almost certainly, the next generation of decision software will be designed to perform a larger range of analyst functions. We have focused on identifying potential problems challenging the computerization of decision analysis, and on assessing the extent to which these problems can be overcome. Two questions are particularly salient: First, to what extent can the often ill-defined art of structuring be transformed into software; and secondly, to what extent is past consumers' satisfaction with decision analysis a function of the formal methods and procedures of the theory and rationale of decision theory, and to what degree do other factors such as personal interaction and the establishment of a rapport account for client approval? We compared multiattribute utility analyses of personal decision problems of undergraduates performed by a human analyst vs. those performed by a "stand-alone" software package, Multi Attribute Utility Decomposition (MAUD). Although subjects overwhelmingly yielded more favorable reports for the analyst session than for the MAUD session, subjects' agreement with & acceptance of the analyst and MAUD results (implied ordering and most preferred alternative) did not differ. We did find that subjects feel better taken care of when more attributes are included in the analysis, but that subjects' holistic ratings are better accounted for by analyses with smaller rather than larger numbers of attributes. We found that MAUD is not as "stand-alone" as its developers have advertised. In particular, our subjects needed at least some instruction in the attribute elicitation phase of the program. We also found that most subjects are unable to answer the brits weighting question properly; uninstructed responses exhibit a sort of risk aversion that renders the weights virtually meaningless. Overall, subjects were highly motivated, and their responses seemed more thoughtful and considered than is often the case for thought experiments with hypothetical scenarios, typical of laboratory experiments with college subjects.

Hierarchical MAU structuring (and weighting) is a particularly attractive approach to building value trees when the number of attributes is large; partly because it reduces the number of necessary judgments, and partly because it avoids weighting questions in which only remotely related attributes need to be compared. But there are several potential problems: for example, respondents may add to an upper level value meaning not captured by its lower branches; ranges of alternatives, which can often be made explicit in specific attributes, become more vague at higher levels, perhaps distorting range dependent importance weights; furthermore, it is not clear whether numbers elciited hierarchically and non-hierarchically are consistent--if not, which should we trust more? We studied subjects' weighting and rating judgments for both hierarchical and non-hierarchical value trees relevant to evaluation of alternatives for electricity production. Hierarchical weights were found to be more variable than non-hierarchical weights, essentially replicating a result reported in 1973. We also found that subjects are often inconsistent in their attribute ratings across different levels of the value hierarchy. Random error, value lability, and misunderstanding of higher level attributes are all possible explanations for this result. Finally, we found that subjects' weight sets did not differ to any great extent as a function of their preferred alternatives; rather, location measures were the primary determinant of preference orderings. Policy decision making, at least in some highly charged arenas, may be more a matter of one's perception of how well each strategy will accomplish the stated goals, and not one sensitive to the tradeoffs among different goals.

Many important and interesting probabilistic information processing tasks are essentially hierarchical or cascaded in formal structure, and involve situations in which different people have different types of probabilistic

information which must be combined. We studied a simple cascaded inference task in which one individual has information about the diagnosticity of an event and the other has information about the probability that the event has occurred (reliability information). Our main purpose was to compare two people working together with an individual combining both types of information, and to assess the relative impact of reliability and diagnosticity. Both nomothetic and idiographic analyses of the responses indicated that diagnosticity has a greater impact than reliability on judged responses. This effect was not mediated by subject sex, nor by whether the information was integrated by an individual working alone or by two people, each given either diagnosticity or reliability. Naming either the "reliability" person or the "diagnosticity" person as responsible for the group response did not alter the findings appreciably. Our results are consistent with the bulk of the subjective inference/prediction literature that suggests that subjects over-emphasize the impact of diagnostic information by not taking full account of other relevant information, such as imperfect correlation in prediction problems, and base-rate information in Bayesian inference.

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 Technical Information Center — 12 copies
Attention: DDC-TC
Cameron Station
Alexandria, Virgina 22314

DCASMA Baltimore Office — 1 copy
Attention: Mrs. Betty L. Driskill
300 East Joppa Road
Towson, Maryland 21204

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

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, D.C. 20301

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

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

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

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

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

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

Department of the Navy

Office of the Chief of Naval
Operations
(OP-951)
Washington, D.C. 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 No. Quincy Street
Arlington, VA 22217

Office of Naval Research
Naval Analysis Programs (Code 431)
800 No. Quincy Street
Arlington, VA 22217

Office of Naval Research
Operations Research Programs (Code
800 No. Quincy Street
Arlington, VA 22217

Office of Naval Research
Information Systems Program (Code
800 No. Quincy Street
Arlington, VA 22217

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

Dean of Research Administration
Naval Postgraduate School
Attention: Patrick C. Parker
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

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

Naval Research Laboratory
Communications Sciences Division (Code 54)
Attention: Dr. John Shore
Washington, D.C. 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

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

Director, Army Library
Army Studies (ASDIRS)
The Pentagon, Room 1A534
Washington, D.C. 20310

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

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

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

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

Department of the Air Force

Assistant for Requirements Development and Acquistion Programs
Office of the Deputy Chief of Staff for Research and Development
The Pentagon, Room 4C331
Washington, D.C. 20330

Air Force Office of Scientific Research
Life Sciences Directorate
Building 410, Bolling AFB
Washington, D.C. 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, D.C. 20330

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

Other Government Agencies

Chief, Strategic Evaluation Center
Central Intelligence Agency
Headquarters, Room 2G24
Washington, D.C. 20505

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

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

Other Institutions

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

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

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

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

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

Dr. Andrew P. Sage
School of Engineering and Applied Science
University of Virginia
Charlottesville, VA 22903

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
Norman, OK 73069

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

Decisions and Designs, Incorporated
Suite 600, 8400 Westpark Drive
P.O. Box 907
McLean, VA 22101

Decision Science Consortium, Inc.
Suite 421
7700 Leesburg Pike
Falls Church, VA 22043