AD-A073 369   DECISIONS AND DESIGNS INC MCLEAN VA   F/G 5/10

RESEARCH ON THE TECHNOLOGY OF INFERENCE AND DECISION.(U)

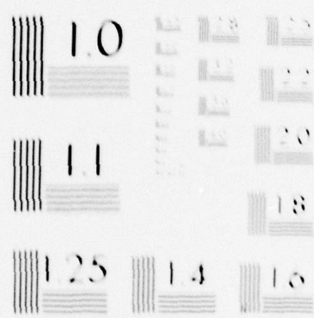
JAN 79 W EDWARDS, R S JOHN, W STILLWELL   N00014-76-C-0074

UNCLASSIFIED   NL

| OF |

AD
A073369

END
DATE
FILMED
9-79
DDC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

social science research institute

USC

UNIVERSITY OF SOUTHERN CALIFORNIA

FINAL RESEARCH REPORT

RESEARCH ON THE TECHNOLOGY OF INFERENCE AND DECISION

WARD EDWARDS
RICHARD S. JOHN
WILLIAM STILLWELL

SPONSORED BY:
ADVANCED RESEARCH PROJECTS AGENCY
DEPARTMENT OF DEFENSE

MONITORED BY:
ENGINEERING PSYCHOLOGY PROGRAMS
OFFICE OF NAVAL RESEARCH
CONTRACT No. N00014-76-C-0074, ARPA

JANUARY, 1979

SSRI RESEARCH REPORT 79-1

B

Social Science Research Institute
University of Southern California
Los Angeles, California 90007
(213) 741-6955

INSTITUTE GOALS:

The goals of the Social Science Research Institute are threefold:

- To provide an environment in which scientists may pursue their own interests in some blend of basic and methodological research in the investigation of major social problems.
- To provide an environment in which graduate students may receive training in research theory, design and methodology through active participation with senior researchers in ongoing research projects.
- To disseminate information to relevant public and social agencies in order to provide decision makers with the tools and ideas necessary to the formulation of public social policy.

HISTORY:

The Social Science Research Institute, University of Southern California, was established in 1972, with a staff of six. In fiscal year 1978-79, it had a staff of over 90 full- and part-time researchers and support personnel. SSRI draws upon most University academic Departments and Schools to make up its research staff, e.g. Industrial and Systems Engineering, the Law School, Psychology, Public Administration, Safety and Systems Management, and others. Senior researches have point appointments and most actively combine research with teaching.

FUNDING:

SSRI Reports directly to the Executive Vice President of USC. It is provided with modest annual basic support for administration, operations, and program development. The major sources of funding support are federal, state, and local funding agencies and private, foundations and organizations. The list of sponsors has recently expanded to include governments outside the United States. Total funding has increased from approximately $150,000 in 1972 to almost $3,000,000 in the fiscal year 1978-1979.

RESEARCH INTERESTS:

Each senior SSRI scientist is encouraged to pursue his or her own reaearch interests, subject to availability of funding. These interests are diverse: a recent count identified 27. Four major interests persist among groups of SSRI researchers: crime control and criminal justice, methods of dispute resolution and alternatives to the courts, use of administration records for demographic and other research purposes, and exploitation of applications of decision analysis to public decision making and program evaluation. But many SSRI projects do not fall into these categories. Most project combine the skills of several scientists, often from different disciplines. As SSRI research personnel change, its interests will change also.

(1)

RESEARCH ON THE TECHNOLOGY OF INFERENCE AND DECISION

Final
Research Report 79-1

January, 1979

Ward Edwards

Richard S. John

William Stillwell

Social Science Research Institute

University of Southern California

Approved for Public Release

Distribution Unlimited

CONTENTS

Acknowledgments

This research was supported by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Office of Naval Research under Contract N00014-76-C-0074 under subcontract from Decisions and Designs, Inc.

The authors would like to thank Hut Barron, Lee Eils, Tsuneko Fujii-Eustace, William Gabrielli, J.Robert Newman, David Seaver and Detlof von Winterfeldt, 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.

I. INTRODUCTION

This report summarizes the work by the Social Science Research Institute, University of Southern California on subcontract P.O. 78-072-0720 from Decisions and Designs, Inc., prime contract N00014-76-C-0074 from the Advanced Research Projects Agency, monitored by the engineering Psychology Programs, Office of Naval Research. The research conducted during this contract period from October 1, 1977 to December 31, 1978 under the direction of Professor Ward Edwards, the Principal Investigator, was an ongoing program of Research on the Technology of Inference and Decision. Edwards (1973, 1975), Edwards and Seaver (1976) and Edwards, John and Stillwell (1977) summarized previous research.

The proposal leading to this subcontract called for research on six specific topics: group procedures for probability assessment, the role of expertise in group probability assessment, measurement and validation of group utilities, assessing small probabilities, the tradeoff between modeling error and judgmental error in multiattribute utility measurement, and development of a computer program to assess probability distributions. Our research on these and other topics is reported in seven technical reports which have been produced or are now being prepared. Summaries of these technical reports appear at the end of this report.

The purpose of this report is to explain how this research integrates into an overall program of research on decision technology. Thus we do not report in detail findings that are set forth in the self-contained technical reports. Instead, we do discuss in detail work that will not result in technical reports as well as new ideas to which current work has led.

II. TECHNICAL OVERVIEW

The research conducted under this program has sought to determine the factors that affect the quality of decisions. With the motivation often coming from practical problems in decision making, both theoretical and applied issues have been investigated. On the practical side, we have compared group behavioral techniques with mathematical aggregation models for group probability assessment (this is a continuation of work started on the previous contract), we have continued our investigation of aids for the assessment of small probabilities, and we have studied the shift in subjects' weighting of attributes in response to changes in the range of value on that attribute. Theoretical work has look at the role expertise plays in the judgment of probability, methods of assessing group value judgments, the tradeoff between modeling error and judgmental error in multiattribute utility measurement, the literature on subjective weighting and a method for testing subjective versus objective weighting of dimensions, and reliability of procedures for the elicitation of multiattribute utility judgments.

II.A. Elicitation and Quantification of Uncertainty

II.A.1 Group assessment of uncertainty: Human interaction versus mathematical models.

Research has shown that the probabilitistic judgments required as inputs to decision analyses are usually more valid when made by groups than when made by individuals (Seaver, 1976). Yet both the theory of and the assessment techniques for subjective probability have been primarily oriented toward quantifying the uncertainty of a single individual. This does not mean, however, that the theory is limited to the single person case. As Savage (1954) points out, extension to the multi-person case are often made. The "individual" simply represents a *family, corporation, or nation*; a sort of "super individual." But we are still left with a problem; the group acting as N individuals.

With this problem in mind, Seaver (1978) investigated both behavioral and mathematical methods for deriving group probability estimates from the individual estimates of group members. Estimates of both discrete (reported in Edwards, John and Stillwell, 1977) and continuous events were explored. Behavioral techniques compared were: Delphi (Dalkey and Helmer, 1963), Nominal Group Technique (Van de Ven and Delbecq, 1971), a consensus technique in which groups were simply to reach a consensus in whatever way they choose, and a mixed technique which, like the Nominal Group Technique, had individuals make judgments and present them to the group, but allowed only presentation of specific reasons for the judgment without open discussion. Mathematical aggregation models compared for continuous events include the linear and conjugate models for each of three weighting procedures: equal weights, DeGroot weights (DeGroot, 1974), and self-rating weights.

The results of this study were essentially negative, a fact that both interesting and important for applications. Seaver found that, although significant, quadratic scoring rule differences among behavioral techniques were slight. Thus, the effort expended in bringing experts together is generally unjustified. Simple procedures, such as equally weighting the individual judgments and taking a simple linear average, produce group assessments that are as good as or better than those produced by more complex aggregation rules or behavioral techniques.

As with discrete events, group interaction judgments were more extreme (in terms both of mean densities and of smaller interquartile ranges) after group interaction than before. This finding was accompanied by a decrement in the calibration of those judgments. Subjects seem to take the agreement of other group members as information independent of their own, and adjust their probability judgment upward.

II.A.2 The use of marker events for the assessment of small probabilities.

Both experimental and applied work have shown that problems arise in the subjective assessment of the likelihood of highly unlikely events. Fujii, Seaver and Edwards, (1977) found a distinct bias of judgments away from the extremes of the probability range. Subjects consistently avoided estimates near 0.0 and 1.0. Obviously, this situation presents a problem for decision analysis when the outcome is sensitive to that bias and we have been investigating methods whose purpose is to promote the veridicality of small probability judgment.

Work by Lichtenstein, Slovic, Fischhoff, Layman, and Coombs (1978) seemed to indicate that subjects were quite good at judging which of two events was the more likely. This finding suggested that probability assessors could make a series of comparisons of event pairs or a simultaneous comparison of the event with unknown probability with a list of events with known probability and thereby arrive at an estimate of the unknown probability. This estimate would not involve direct numerical estimates and thus would avoid the bias to which these judgments were subject.

Several studies were designed to evaluate this elicitation technique. Edwards, John and Stillwell (1977) describe these studies and report the general conclusion that, at least as so far designed, the new response mode did not result in marked improvement in resulting judgments. Correlations for individual subjects between the true probabilities and those estimated by subjects using the procedure tended to be low positive, on the order of .1 to .2.

Based on these results it was suggested that for the skill of the probabilistic judge to be effective in realizing improved estimates some form of aggregation over individuals must occur. The responses found by Slovic et al. were across subject rather than within and the quality of individual subjects' responses were not studied. Slovic, et al. found that directional responses were correct for a large proportion of subjects, but not all persons were correct and most ratio judgments were too conservative.

The above findings suggested a sort of majority rule principle for the paired comparisons of probabilities as a possible solution to the small probabilities elicitation problem. To test this possibility the following study was designed. Subjects were run in groups of five. All five subjects were shown pairs of individual events of the following type: "A secretary given that the person was employed in California in 1976." This is a condensation of "Select an adult at random from the U.S. population. It turns out that the person chosen was employed in California in 1976. Given that information, what is the probability the person was employed as a Secretary?" Each subject was then asked to determine which of the pair he/she felt was more likely. A branching structure was used as in the individual studies but the branch taken was determined by at least three of the five subjects in the group selecting the same event as the most likely. Choices were forced such that for each pair, one must be selected. After five such choices the group had confined the estimate to a very narrow range of probability (.05 in log distance) and the midpoint of this range was taken as the group's judgment.

Result of this study raise serious doubts as to the viability of the marker event technique for the elicitation of small probabilities. Groups of individuals performed only slightly better than did individual subjects. Correlations for groups between the true probabilities and those elicited were for all groups nonsignificant and in some cases slightly negative.

Again, we find ourselves in the position of needing to explain the seeming divergence of our results from those of Slovic et al. And, again, we must look to the differences in the task. Slovic et al found that when the ratio of likelihoods of two events was greater than 2:1, over eighty percent of subjects could correctly judge the direction of the response. On the other hand, in order for the marker event technique to be useful, the subject must necessarily make judgments that are close to 1:1. The results of Slovic et al. show that subjects asked to make judgments less than 2:1 often times showed well over fifty percent of those responses to be in the wrong direction. And for the majority of pairs, judgments made with the true ratio of likelihoods less than 2:1 did not markedly differ from fifty percent.

We have come to believe that the marker event technique for small probability assessment is not viable. The nature of the required task is such that subjects lack the sensitivity required for the judgments. At the same time we feel that the results of Slovic et al. are misleading ability to make relative likelihood judgments about human for very low probabilities. In a gross sense subjects performed this task moderately well, but certainly not well enough so that some aggregated form of the judgments could result in a useful assessment technique.

II.A.3 Combining probabilitistic judgments in the presence of differing levels of information.

Often those with special knowledge of a substantive area are asked to work together to provide probabilistic input to a decision analysis. Their special knowledge is thought to aid them in coming to "better" estimates of the needed unknown quantities. The most common situation is where more than one judge is involved, and each judge has a different level of information or knowledge of the process about which they are trying to make judgments.

Gabrielli and Edwards (1978) looked at this situation experimentally. In a book-bag-and-poker-chip task subjects were given controlled amounts of information. One subject, said to be the high level expert, was given more diagnostic information (d'=1.15) about an underlying data generating process. A second subject, the low level expert, was given less diagnostic information (d'=.41) about that same data generating process. First, each subject made judgments individually from their own sample. Subjects were then combined into two person groups and asked to arrive at a joint judgment that took into account both bits of information. They were not allowed to exchange information about their individual sample except to express the strength of their individual judgment. The independence of the two bits of information was stressed.

Individual judgments were consistent with past research (Domas, Goodman, and Peterson, 1972; Stillwell, Seaver and Edwards, 1977) in likelihood ratio judgment. The high expertise subjects tended to make conservative estimates while the low expertise subjects were radical. The correlations for individual subjects between their twenty responses and those calculated from Bayes theorem, the normative rule, were for the majority of subjects above .75.

For the group judgments, three sets of numbers were compared: 1) The actual group response was compared with veridical responses and 2) with a number calculated by multiplying individual judgments in the appropriate Bayesian manner. In addition, 3) the number calculated from the individual judgments was compared with the veridical response. Median correlations, slopes and intercepts were respectively, .79, .45 and .16 for 1), .54, .23 and .48 for 2) and .53, .50 and .79 for 3).

The fact that actual group responses correlate fairly highly with products of individually estimated likelihood ratios means no more than that the two members of each group were allowed to tell each other what their individual estimates had been, and were instructed to take those estimates as veridical for the purpose of arriving at group assessments. But the low slope shows that they did not arrive at these final assessments by the normatively appropriate procedure. Direct observation showed that some pairs, but not all, arrived at final estimates by simply averaging their individual estimates.

To check how effective a theory, simple averaging might be for explaining the group assessments, we also calculated regressions between actual group responses and the mean of the logs, rather than the sum of the logs, of the individual responses. The median slope for that regression is .89. Similarly, we calculated the same regression between actual group responses and correct Bayesian numbers; that median is .47. These regressions are too low to permit the conclusion that sujbects systematically averaged their estimates. We are left with the conclusion that the subjects found the task confusing, and adopted ill-specified and confused strategies for determining their responses.

A third comparison was that of group responses with a simple averaging model. This tested the notion that subjects simply took an average of the two numbers that represented their individual judgments as their group judgment. Finally, the fourth test examined the geometric mean of the two individual judgments as predictor of the group response.

Regression analyses were performed using the Bayes model on veridical likelihood ratios, the Bayes model on individual responses, the geometric mean model and the arithmetic mean model as predictors of group response. The results of these four analyses suggest that subjects were averaging the individual judgments to come to the group response. Average group correlations of .46, .64, and .91 and average group slopes of .37, .44, .78 and 1.16 were found between response and Bayes model on the veridical, Bayes model on individual responses, the geometric mean model and the arithmetic mean model respectively. Clearly, considering the four models investigated here, the arithmetic mean model best describes subjects' behavior.

There is an interesting point to be noted concerning the subjects' perception of the experimental task. Using the arithmetic mean as the group response as described above would in fact be the normative rule if the data seen by the two subjects had been totally dependent except for random error. Thus, the fact that subjects did average the individual responses suggests that they perceived the two individual judgments to represent redundant information with an error component. In contrast, the subjects in Seaver's (1978) experiment treated the judgments of the others in the group as independent information sources when it seems quite likely that they were redundant. They seemed to give too much importance to group agreement.

These findings are linked with those of a previous report (Eils, Seaver, and Edwards, 1977) showing that if an individual subject is asked to estimate mean log likelihood ratios in a Bayesian inference task, the resulting judgments are extremely close to veridical. Apparently, averaging is a congenial method of aggregating information, whether that information is accumulated by looking at successive items of data inside one head, or is instead obtained as a result of group interaction. If, for example, the Gabrielli-Edwards experiment had used mean log likelihood ratios as its response mode, both individual and group responses might well have been extremely close to the veridical value.

II.B. Multiattribute Utility Assessment

The current contract called for simulation research on the trade-off between modeling error and judgmental error in multiattribute utility measurement (MAUM) and behavioral experimentation on the topic of group utility assessment.

Our work on these two problems is reported in two technical reports (Barron, 1978; Eils & John, 1978). Laboratory research designed under the previous contract and completed under the current contract on an important topic -- the reliability and validity of utility assessment procedures -- is reviewed below. No further report is planned. Research on the problem of importance weighting in additive, riskless MAUM was also conducted and is reported in three technical reports. John and Edwards (1978a) reviewed the literature on importance weighting, John and Edwards (1978b) performed a criterion validation experiment on various weighting techniques, and Gabrielli, von Winterfeldt, and Edwards, (1978) investigated the sensitivity of subjective estimates of attribute importance to value ranges.

II.B.1. Additive vs Multiplicative MAU model -- Error Tradeoffs

Barron (1978) attacks the problem of judgmental error and modeling error for riskless and risky MAU functions. A new procedure for eliciting multiplicative (or additive) MAU functions, holistic orthogonal parameter estimation (HOPE), is presented. Three types of error associated with all MAUM procedures are identified; two are judgmental and one corresponds to model mis-specification . Simulation data are reported. A methodology for detecting judgmental errors (within the HOPE paradigm) is explained. Finally, various considerations in the application of HOPE are discussed.

The HOPE procedure consists of three steps: preparation, holistic judgment of an orthogonal array of alternatives, and derivation of utility and weight parameters from holistic responses. During preparation, attributes are determined, value ranges are estimated, and preferential and utility independence assumptions are checked. These tasks are performed in the manner suggested in Keeney and Raiffa (1976). Next, an orthogonal array (see Addelman, 1962) of alternatives is constructed, and holistic ratings of worth are collected. The logic of this step is identical to that of functional measurement (Anderson, 1974, 1977); however, HOPE has the advantage of requiring far fewer holistic assessments. For example, a decision problem with five attributes, each defined on four levels, requires only 16 holistic judgments. One replication of a functional measurement complete factorial design for the same problem would require 625 (5^4) judgments. In the riskless case, these holistic assessments are simply ratings (magnitude estimates), while standard multi-arrtibute lottery procedures (basic reference lottery tickets) are utilized in the risky case. The arithmetic of utility and weight derivation is explained in detail in Barron (1978) and will not be repeated here. The analysis allows for the diagnosis of the MAU model as either multiplicative or additive.

Barron (1978) identified three sources of error in utility elicitation. Modeling errors result from the use of an inadequate functional form (and corresponding assessment procedure). Common examples include substituting additive models for multiplicative, and riskless assessments for risky ones. Judgment errors are of two types: gross biases and random deviations. Gross biases, such as the so called "certainty effect" (Kahneman and Tversky, in press), may cause severe inconsistencies in the elicited utility structure. Random judgmental errors can be attributed to many sources, including lack of attention in the subject (caused by either boredom or fatigue), imprecision in the response mode, etc. Data are reported from a simulation study (for details, see Barron and Person, in press) suggesting that model misspecification is a much more severe problem than random judgmental error. Although errors in multiplicative models are relatively constant as random error become more severe, the accuracy of additive functional forms drops drastically as random error is magnified. This suggests that a blanket acceptance of additive models, with no checks as to their appropriateness, is ill-advised.

Barron argues that the ability of the HOPE procedure to distinguish additive models from multiplicative ones is an important asset, not shared by the SMART procedure. Unfortunately, the issue of systematic biases in judgment, common to all model forms and assessment procedures, is not addressed in the simulation.

Barron's simulation findings are useful, but further research is clearly warranted. The results are clearly dependent upon the distributions of random error used in creating "judgmental errors". Of course, the distributions chosen by Barron were simply assumed, as little or no research on the topic exists. With different distributions of random error, it is plausible that judgmental errors would become a greater concern relative to model misspecification. Whether Barron's results hold for random errors in judgments required by the more traditional decomposition approaches is unknown. Furthermore, all of these results are derived using two hypothetical multiplicative models. Just how realistic these models are is not addressed in the paper, and is very difficult to assess. Like all simulation findings, these must be tempered by awareness of the assumptions and defining problem characteristics decided upon by the simulator(s).

Barron offers a methodology for the detection of errors in holistic judgments. Two orthogonal designs are specified (with "minimal overlap"), and holistic judgments are collected over each. One approach is to pool the data, and estimate the parameters with increased degrees of freedom. The subjective holistic estimates are than compared to those specified by the HOPE mode. Another approach is to estimate the model parameters twice (once for each set of judgments), and then to apply a "jacknife validation" logic. That is, elements in one orthogonal design are compared to the composites generated by the HOPE model derived from the other design. With either of the above approaches, large discrepancies between HOPE model composites and subjective holistic estimates are an indication of judgmental error. In applied settings, these discrepant holistic estimates may be reassessed.

The HOPE methodology for assessing MAU functions is a useful addition to decision analysts' "bag of trick." It combines many of the best features of the decomposition approach; discrimination between additive and multiplicative forms is possible and only a small number of judgments are required. In addition, Barron claims that holistic estimates are more easily collected (via questionnaire) than judgments required by standard decomposition procedures. The well known finding that holistic assessments are generally poor (see Slovic and Lichtenstein, 1971) is assumed not to apply to trained experts. Barron asserts that experts, familiar with the task of combining large amounts of information, can provide accurate holistic estimates. Of course, there is much evidence suggesting that some professional experts (clinical psychologists, for example) consistently make poor holistic evaluations. (Goldberg, 1965, 1968, 1970, 1971). In practice, the quality of holistic assessments will vary as a function of the expertise of the decision maker, and the complexity of the decision problem. The MAU analysis performed by Edwards (1978) on the problem of choosing an alternative busing plan for the Los Angeles Unified School District required well over 100 dimensions. The efficacy of the HOPE procedure in such a case is obviously limited.

Further research on the HOPE precedure is currently planned. A study has been designed to assess the criterion validity of the weight estimates derived from the HOPE procedure. Formally, the weight estimates are obtained analytically as a special case of the general linear model; thus, the procedure is formally equivalent to many other well known approaches to obtaining statistical weights (e.g., regression, ANOVA, fractional replication design ANOVA). Using the multiple-cue paradigm of John and Edwards (1978), HOPE weights will be compared to other statistical weights, as well as direct subjective weight estimates. In another study, the riskless version of HOPE will be compared to a number of other MAU assessment techniques in a realistic decision setting. MAU models of credit card applicant worthiness will be elicited from credit officers from a major California bank. All procedures will be compared against an evaluation model based on a standard statistical analysis of a large data base. Thus, the criterion validity of HOPE, relative to other MAU methods, will be assessed. This study will also investigate the undesirable possibility that HOPE procedures are not as useful in helping decision makers "get their heads straight" as the more familiar decomposition procedures.

II.B.2. A Criterion Validation of Group MAUM

Eils and John (1978) report a behavioral investigation of group MAUM based on a doctoral dissertation by Eils (1977). (For an earlier summary of some of this work, see Edwards, John, and Stillwell, 1978). Twenty-four groups, each composed of three college students, produced consensus evaluations of ten hypothetical applicants for bank credit cards. Half of the groups obtained consensus evaluations via Edwards' SMART procedure, while the other half performed holistic assessments. Also, half of the groups used a formal group communication strategy, while the other half were given no instructions regarding the dynamics of group interaction. The quality of group assessments was defined to be the Pearson product-moment correlation between the group's assessments and those from a configural (nonlinear) model used by Security Pacific Bank based on a standard discriminant analysis utilizing a large data base.

The SMART decomposition methodology significantly improved the quality of collective decisions, as did, to a lesser extent, the communication strategy. The correspondence between the bank criterion and group decisions reached via SMART was improved when the ten attribute weights were obtained from the simplified assessment schemes of rating, ranking, or setting all equal to a constant.

The bank's formalized process of evaluating applicants for revolving credit loans reflects, with some degree of accuracy, the nature of the complex relationship between applicant characteristics and subsequent loan performance. Information bearing on this complex relationship is a part of individuals' past experience (otherwise the choice entities would appear equally attractive and evaluation would be impossible). Thus the degree to which group decisions correspond to the bank's systematic and complex evaluation provides a measure of the match between the collective decision elicited and group members' experience. It is argued that the advanced behavioral technologies explored by this research are valid in the sense that they elicit a more nearly complete representation of individuals' past experience. Further research within this paradigm, using professional bankers, was discussed in II.B.1.

While the simple multi-attribute utility analysis (SMART) was developed for use by individuals, these experimental results suggest that the technique is readily adaptable to the group task setting. The success of the group decision technology lies in its ability to focus attention to individual value relevant factors. As reflected by the improved correlations obtained with simplified weighting schemes, the precise specification of weight parameters was of little value in the present decision task. In the reported study, the assessment of ratio weights merely exhausted more group time and energy as well as lowered the quality of the decision product. Whether these findings will be replicated with expert credit judges is an interesting question.

II.B.3. MAU reliability and validity

One of the gaps in the transfer of decision analysis from abstract theory to useful application is reflected in the paucity of attempts to assess the reliability and validity of elicitation procedures. Fujii-Eustace and Edwards designed a behavioral study (reported in Edwards, John, and Stillwell, 1977) comparing simple utility models (riskless, additive) to more complex ones (risky, multiplicative). They elicited six subjects' two-attribute utility functions for separable (market baskets of tea and ice cream) and inseparable (amount of leanness of ground beef) commodity bundles. All assessments were performed twice and the four single-attribute utility functions were determined over a fixed range for gains added to a small initial bonus and losses substracted from a larger initial bonus. The subjects were all paid a fixed amount, there were no real transactions of stimulus commodities.

There are three major findings in the study. First, test-retest reliability was extremely high. No differences were evidenced among the two risky procedures (certainty equivalents or basic reference lottery ticket) and the riskless rating scale procedure, for either the single-attribute utilities or composite multi-attribute utilities. In addition, there was high convergence among all of the multi-attribute utility models. All pairs of riskless/risky and additive/multiplicative models corresponded equally well. Unfortunately, the tests of reliability and convergent validity are based on correlational analyses. Since the monotonicity restriction virtually guarantees high correlations, these findings should be interpreted cautiously. In particular, we will discuss below how the high reliability of the single-dimension utilities is almost totally artifactual.

Secondly, all of the utility models predicted holistic ratings of multi-attribute commodity bundles and dichotomous choices between lotteries of multi-attribute commodity bundles extremely well. Although the risky certainty equivalents procedure tended to predict holistic ratings best and the risky basic reference lottery ticket method tended to predict the dichotomous lottery choices worst, all of the differences were quite small.

Lastly, an attempt was made to diagnose each and every single-attribute utility curve as either convex, concave, or linear. Presumably, the risky assessment procedures, which take attitude towards risk into account, should produce more nonlinear utility functions than the riskless rating scale procedure. Also, a critical comparison of classical utility theory and prospect theory (kahneman and Tversky, in press) is afforded. Classical utility theory would predict that the shape of the utility curve is constant (and probably concave) over the "gains" and "losses" assessments, whereas prospect theory asserts that the curves are more likely to be concave for "gains" and convex for "lossess". This analysis was inconclusive for three reasons: non-monotonicity of several utility curves, low test-retest reliability of the diagnoses of the curves' shapes, and rather extensive individual differences among subjects. For the commodities and ranges employed, non-monotonictiy is simply inexplicable. It is perhaps noteworthy that non-monotonicity was evidenced only in the basic reference lottery ticker procedure; rating scales and certainty equivalents force the curves to be monotonic. The rather low test-retest reliability of curve shape suggests that the previously mentioned high test-retest correlations for single-attribute utilities are artifactual. In all likelihood, the shape diagnoses obtained are spurious deviations from linearity (which might be expected over the rather restricted value ranges assessed). Finally, the consistencies (in group data) in shape relating to gains vs losses and risky vs riskless were not consistent across subjects.

No more than two subject exhibited the same patterns of curve shape across these two factors of interest.

The results of this study support a small but growing body of research that suggests that simple additive, riskless utility assessment procedures are adequate for a large class of decision problems (Fischer, 1972, 1976, 1977, von Winterfeldt & Edwards, 1973). Edwards' (1977) SMART procedure for evaluating MAU function, using simple magnitude estimates (ratings) of single-attribute utility and ratio estimates of weights, yielded models that closely correspond to the complex multiplicative models and risky assessment procedures advocated by Keeney and Raiffa (1976).

Of course, further research is warranted. One could argue that the high correspondence of additive and multiplicative models is partially due to a high degree of independence between the two attributes in each of the commodity bundle stimuli. Even in the inseparable case (amount and leanness of of ground beef), there is little reason, a priori, to expect any value dependence. Likewise, the rather trivial value ranges used, along with the absense of any real transactions virtually assured that attitude towards risk would be irrelevant. The usual findings of risk aversion, both single- and multi-attribute, are less likely to occur when the stimuli are both imaginary and trivial (cell 4 of Table 1). Any critical comparison of prospect theory and classical utility theory demands stimuli and value ranges which produce significant levels of non-neutral attitude towards risk.

The ideal experimental setting, using non-trivial stimuli and real transactions (cell 1 of Table 1) is not practical. Handling substantial gains or losses (even from an initial bonus) is an obvious problem. Although rigged gambling devices are a possibility, the use of such deception would have to be approved by a human subject committee. The solution, of course, is to use either real transactions for less significant stimuli (cell 2 of Table 1) or imaginary transactions for more important stimuli (cell 3 of Table 1). Research on the topic of atitude toward risk and value independence is currently planned in the domains of cells 2 and 3.

		Stimuli	
		Significant	Insignificant
Transactions	Real	1	2
	Imaginary	3	4

Table 1. Experimental settings relevant to studying attitudes toward risk and valve independence.

II.B.4. A Review of Importance Weighting for Additive, Riskless MAU Models

John & Edwards (1978a) identified three approaches to the weight estimation problem: no estimation (i.e., equal weights), direct subjective estimation, and indirect holistic estimation. A rather extensive literature on equal weighting, both pro and con, is summarized and discussed. Several methods for directly assessing importance weights are catalogued, including ranking, fractionation, subjective estimate methods, and paired comparison procedures. Indirect holistic methods, including unbiased and biased regression analyses, the ANOVA and fractional ANOVA paradigm, and the Keeney-Raiffa techniques of pricing out and trading off to the most important dimension, are all explained with particular emphasis on their common relationship to the general linear mode.

The validation literature on importance weighting techniques is characterized by three notions of weight correspondence: direct comparisons of weights elicted via different methods, comparisons between composites formed from elicited weights and holistic evaluations, and comparisons between different composites formed from weights elicited via different methods. More recent literature utilizing a criterion related notion of weight validity is also reviewed. Most, if not all of the evidence presented by Slovic and Lichtenstein (1971) suggesting that subjective weights are poor is reconciled. In addition virtually all of the more useful recent work on this topic suggests that the concept of attribute importance is a psychologically meaningful one. That is, subjects usually give responses to direct subjective assessments of importance weights that are both consistent (high convergent validity) and accurate (high criterion validity). There is some evidence indicating that the psychological concept of importance is more closely related to the statistical notion of attribute validity (correlation) than to either the least-squares regression weights (Schmitt, 1978) or the "proportion of variance accounted for" (Brehmer & Qvarnstrom, 1976; Schmitt, Coyle & Saari 1977; Slovic, 1969).

When problems with multi-collinearity and cross validation are taken into account, few discrepancies between subjective weights and statistical indices of importance (derived from holistic evaluations) were observed. Bases on recent research, the conclusions reached by Slovic and Lichtenstein (1971) are no longer justifiable, if they ever were. Unfortunately, their conclusions have been generally accepted and widely quoted (see Nisbett & Wilson, 1977; Hammon, McClelland, and Mumpower, 1978). Partially because of Slovic and Lichtenstein's negative summary, there has been a shortage of research on subjective weights as compared to that on statistical weights (this view agrees with that of Schmitt and Levine, 1977). Finally, this reappraisal of the weighting literature is highly relevant to the current debate over the extent to which people are aware of their own cognitive processes. Serious doubt is cast over that portion of Nisbett and Wilson's (1977) conclusions on this issue which are based on Slovic and Lichtenstein's review.

II.B.5. A Criterion Validation of MAU Weighting Techniques

John and Edwards (1978b) proposed a research paradigm for comparing weight estimates to empirically derived "true" weight, thus obtaining a measure of the criterion validity weight assessment techniques. In the approach, also proposed and tested by Schmitt (1978), subjects are first taught a MAU model via multiple cue probability learning and outcome feedback. Then, various assessments of the importance weight parameters for the model attributes are obtained. The composites formed from these weight estimates are then compared to composites formed from "optimal" statistical weights derived from the outcome feedback.

Two experiments were conducted to assess the validity of several weight assessment techniques. In the first experiment, a four attribute MAU model, with zero environmental correlations among attributes, was taught to nine subjects. The regression, rank, and ratio weight estimates all resulted in composites which closely matched those of the true model; most subjects' weighting schemes were a great improvement over either equal or extreme weighting. For three of the nine subjects, the rank and ratio assessments produced lower matching than did the regression weight estimates.

In the second experiment, a total of eight subjects were taught one of two four-attribute MAU models, each involving substantial attribute intercorrelations. Both of these models were less explicit (more error variance) than the one taught in Experiment I. A total of eight methods were employed in assessing subjects' importance weights: OLS and ridge regression on holistic evaluations, OLS and ridge regression using ratio weight estimates as validity coefficient estimates, direct subjective ranking and ratio estimation, and the two Keeney-Raiffa techniques of pricing out and trading off to the most important dimension. For the model involving one large positive correlation between two of the attributes, all eight weight assessment methodologies produced equally good composites; all composites derived from subjects weights corresponded to the "true" model composites better than simple heuristic rules such as equal weighting and extreme weighting. For the model involving two rather large negative intercorrelations among attributes, the results are inconclusive. Althought the statistical weights were superior for one subject who seemed to have learned the model well, the direct assessment and Keeney-Raiffa weights were superior for two of the subjects who did not learn the model so well. Only one subject produced valid weights across all eight assessment techniques.

The present research and findings are interesting from both an applied and theoretical perspective. For the applied decision analyst, the work by Schmitt (1978) and that reported in John and Edwards (1978b) contributes strong evidence for the assertion that the additive MAU model is a valid perscriptive tool. The evidence that people can indeed provide direct subjective estimates of importance weighting is an important component to this assertion. In most interesting problems, such as choosing a school desegregation plan or siting a nuclear power plant, a large alternative set is not readily known a priori. In such applied situations, the feasibility of most indirect holistic approaches to deriving importance weights is in doubt. Even if a reasonable large set of alternatives could be generated, in most cases the number of dimensions involved makes the task of holistic evaluation of alternatives extremely difficult, if not impossible.

Although decision analysis is often helpful in "getting your head straight", most analysts would like to think of themselves as more than therapists. The overwhelming belief among most decision analysts is that these methods elicit parameter estimates of preference models that result in a normative choice structure. That is, decision analysts believe that their clients should behave in the manner suggested as optimal by the elicited choice structure. Although the stimuli used in the present study (diamonds defined on four dimensions) and in Schmitt's (1978) study (graduate applicants defined on four attributes) are simplistic, and the acquistion of information about attribute importance is contrived (feedback learning), the results suggest that attribute importance is a valid psychological construct. That people can make accurate estimates of importance weights in the laboratory setting is certainly a necessary condition for their being able to do so in the more complex and emotional settings usually faced by a decision analyst and his clients.

From a theoretical perspective, the results from the current study, and Schmitt's (1978) study challenge Nisbitt and Wilson's speculations concerning importance weights and prospective introspection. Subjects were able to provide importance weights predictive of their own holistic evaluations in an experimental setting for which there were no stored "rules" for determining judgments. The diamond appraisal policies in the present study were learned indirectly, without the intervention of verbal descriptions or formal linguistic rule. Subjects demonstrated an awareness both of their own rules for making diamond appraisal, and of the criterion diamond model used to generate the outcome feedback.

The present study suggests that an important future variable in weighting research is the intercorrelation matrix of attributes. Although the "true" criterion model is more difficult to determine when attributes are intercorrelated, the application of biased regression techniques makes the task a manageable one. The results of the present study were moderately encouraging for the novel hybrid weighting approach suggested by Newman (1977); further research is needed however.

A possibly important intervening variable in the assessment of importance weights is the amount of exposure subjects have to the "true" MAU model. Also, the explicitness of the MAU model is another potential intervening variable. If the overall utility of the stimuli are not predicted well by the attributes considered (high error variance), subjective weights may not be so accurate. The level of experience of the decision maker, and the strength of the relationship between attributes and overall utility are concrete variables, often highly descriptive of specific applied settings. The first variable relates to the notion of decision maker expertise, while the second is a function of the defining characteristics of the decision problem. Future research on weighting should systematically explore the effects of the number of trials of feedback learning and proportion of error variance in the true model on subjective estimates of attribute importance. The problem of group assessment of importance weights is yet an additional topic for future research that has heretofore received little attention.

II.B.6. Sensitivity of importance weights to value ranges.

One of the implicit assumptions of any additive MAU model is that importance weights are dependent upon the value ranges of the attributes. Although recent versions of Edwards' SMART procedure suggest that weights be elicited with the value ranges of the individual attributes in mind, the original version required that importance weights be elicited prior to a specification of the alternatives (Edwards, 1971). Otway and Edwards (1977) further recommend a transformation procedure by which weights elicited with one set of ranges in mind can be adjusted to accommodate the actual set of value ranges encountered. In two experiments, Gabrielli and von Winterfeldt (1978) empirically investigated the assumption that subjects' importance weights are sensitive to the value ranges of the attributes.

In the first experiment, single-attribute utility functions and importance weights were elicited from twenty-four subjects for two different decision problems: Liquified natural gas plant location sites and apartments. Each decision problem consisted of five alternatives defined over six dimensions. Each subjects' single-attribute utilities and weights were assessed twice. The value range on one of the dimensions (environmental damage in millions of dollars for the LNG problem and break-ins per year in the apartment problem) was either increased of decreased between the two assessments. Since the utility functions of both of the altered attributes are monotonic, MAU model axioms require that the weights on these dimensions be greater when the value ranges are greater. Results indicated that subjects' weights were not ordinally consistent with the value ranges presented. For those subjects who were ordinally consistent, the expected weight changes were substantially less than would be expected from the elicited single-attribute utility curves.

There are three possible interpretations of this experiment. One could conclude that the elicited importance weights are not sensitive to the value ranges of the attributes, and are thus incorrect. Subjects may normally consider implicit value ranges based on their own experience. This interpretation would suggest that the psychological concept of importance weight is independent of the value ranges of the attributes. A second possibility is that the subjects' utility functions over the altered attributes were misestimated, and that they are relatively flat over the altered range. If this were the case, one would expect little or no change in the weights. The deviations observed would be attributable to random response error. A third explanation is that both the single attribute utilities and the importance weights were correctly determined, and that the observed inconsistencies are nothing more than random response error. "Expected weights", calculated from the importance weights from the first session, and both sets of single attribute utilities, are typically within .05 (weights normalized to sum to one) of the first session weights. That is, even if the utility functions are assumed to be absolutely correct, little difference in the two sets of weights could be expected.

A second experiment was conducted to overcome the interpretation problems manifest in the first. Sixty-nine subjects were shown "utility profiles" of three cars, defined on the two dimensions of gas mileage and weight. In addition to providing the location measures, the experimenter told the subject that gas mileage should be considered twice as important as weight. The subject was then told to imagine that all of the values for the gas mileage dimension had doubled. Half of the subjects were required to make a forced choice between increasing or decreasing the importance weight ratio, while the other half were given the additional response option of no change.

Since the location measures given for both ranges of gas mileage indicate that the experimenter is linear in utility across the entire value range from 10 to 60 miles per gallon, one would expect that subject would provide new weights whose ratio was 4 to 1, or twice that given originally. Of the 34 subjects allowed to specify no change, over two-thirds did so. Eight of the ten remaining subjects correctly specified an increase in the weight ratio. In the forced choice condition, only about one-third of the 35 subjects correctly specified an increase. Apparently, subjects could not even specify importance weight ratios ordinally consistent with value ranges in a two-attribute linear utility decision setting.

It should be emphasized that no subjects' utilities or weights were actually assessed in this second experiment. The numbers provided by the experimenter as utilities and weights precisely determine a "correct" response for the experimenter's adjusted weight ratio. In essence, the subject has been asked to work a hypothetical MAU problem, similar to an exercise in a textbook. The task resembles one of puzzle solving much more than one in which a subjective quantity (importance weight) is elicited. There is a distinct difference between subjects giving importance weights which are sensitive to value ranges and subjects having an intuition for the analytical properities of MAU theory. It is questionable whether the first issue was addressed in this experiment at all.

Although the results of Gabrielli and von Winterfeldt's (1978) investigation are suggestive, further empirical research is clearly warranted. The finding that "importance" is a range insensitive concept would pose a major problem to the SMART Procedure. Two alternative solutions would be possible: In the first the term importance would be given up altogether and substituted by cross attribute relative value or indifference judgments (Kenney and Raiffa, 1976); in the second importance judgments would be made independently of ranges and ranges would be defined to cover a "plausible" set of alternatives rather than the available set.

The first solution is simple and can easily be implemented with only minor rewording of the weighting procedure. The second solution requires elicitation of accurate "implicit value ranges" for each attribute. Of course, the interpretation of the MAU model and the role of weights as rescaling factors remains the same. The decision maker's implicit value ranges are simply substituted for the actual value ranges of the available alternatives. Whether either of these alterations of SMART are necessary is the topic of further research.

References

Addelman, S., Orthogonal main effect plans for asymmetrical factorial experiments. Technometrics, 1962, 4, 21-46.

Anderson, N.H., Information integration theory: a brief survey. In D.H. Krantz, R.C. Atkinson, R.D. Luce, and P. Suppes (Eds.). Contemporary developments in mathematical psychology (Vol.2). San Francisco: Freeman, 1974.

Anderson, N.H., Note on functional measurement and data analysis. Perception and Psychophysics, 1977, 21 201-215.

Barron, H.F., Validation and error in multiplicative utility functions (SSRI Tech. Rep. 78-2). Los Angeles: University of Southern California, Social Science Research Institute, December, 1978.

Barron, H.F. and Person, H.B., Assessment of multiplicative utility functions via holistic judgments. Organizational Behavior and Human Performance, in press.

Brehmenr, B., and Qvarnstorm, G. Information integration and subjective weights in multiple-cue judgments. Organizational Behavior and Human Performance, 1976, 17, 118-126.

Dalkey, N., and Helmer, O., An experimental application of the Delphi method to the use of experts. Management Science, 1963, 9, 458-467.

DeGroot, M., Reaching a consensus. Journal of the American Statistical Association, 1974, 69, 118-121.

Domas, P.A., Goodman, B.C., and Peterson, C.R., Bayes theorem: response scales and feedback (Tech. Rep. 037230-5-T). Ann Arbor, Mich.: University of Michigan, Engineering Psychology Laboratory, 1972.

Edwards, W., Social utilities. The Engineering Economist, Summer Symposium Series, 1971, 119-129.

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 of social decision-making. IEEE Transactions on Systems, Man, and Cybernetics, 1977, SMC-7, 326-340.

Edwards, W., Multiattribute utility measurement in a highly political context: evaluating desegregation plans in Los Angeles. Unpublished manuscript, 1978. (Available from Social Science Research Institute, University of Southern California, Los Angeles

Edwards, W., John, R.S., and 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., and 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.

Eils, L.C.III The effects of communication strategy and decision technology on the process and product of decision making groups. Unpublished doctoral dissertation, University of Southern California, June, 1977.

Eils, L.C.,III, and John, R.S. Multi-attribute utility analysis and communication strategy in group decision making: a criterion related validity experiment (SSRI Tech. Rep. 78-4). Los Angeles: University of Southern California, Social Science Research Institute, December, 1978.

Eils, L.C., III, Seaver, D.A., and Edwards, W., Developing the technology of probabilistic inference: aggregating by averaging reduces conservatism (SSRI Tech. Rep. 77-3). Los Angeles; University of Southern California, Social Science Research Institute, August, 1977.

Fischer, G. Four methods for assessing multi-attribute utilities: an experimental validation (Tech. Rep. 037230-6-1). Ann Arbor, Mich.: University of Michigan, Engineering Psychology Laboratory, June, 1972.

Fischer, G.W., Convergent validation of decomposed multiattribute utility assessment procedures for risky and riskless choice. Organizational Behavior and Human Performance, 1977, 18, 295-315.

Fischer, G.W., Multidimensional utility models for risky and riskless decisions. Organizational Behavior and Human Performance, 1976, 17, 127-146.

Fujii, T., Seaver, D.A., and Edwards, W., New and old biases in subjective probability distributions: do they exist and are they affected by elicitation procedures? (SSRI Tech. Rep. 77-4). Los Angeles: University of Southern California, Social Science Research Institute, 1977.

Gabrielli, W., and Edwards, W. Expertise as sample diagnosticity: combining judgments in Bayesian inference task (SSRI Tech. Rep. 78-8). Los Angeles: University of Southern California, Social Science Research Institute, December, 1978.

Gabrielli, W., and von Winterfeldt, D., Are importance weights sensitive to the range of alternatives in multiattribute utility measurement? (SSRI Tech. Rep. 78-7). Los Angeles: University of Southern California, Social Science Research Institute, December, 1978.

Goldberg, L.R., Diagnosticians vs. diagnostic signs: the diagnosis of psychosis vs neurosis from the MMPI. Psychological Monographs, 1965, 79 (9, Whole No. 602).

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

Goldberg, L.R., Man versus model of man: a rational plus some evidence, for a method of improving on clinical inferences. Psychological Bulletin, 1970, 73, 422-432.

Goldberg, L.R., Five models of clinical judgment: An empirical comparsion between linear and nonlinear representations of the human inference process. Organizational Behavior and Human Performance, 1971, 6, 458-479

Hammond, K.R., McClelland, G.H. and Mumpower, J., The Colorado report on the integration of approaches to judgment and decision making (Rep. No. 213). Boulder, Co.: University of Colorado, Institute of Behavioral Science, October, 1978.

John, R.S. and Edwards, W., Importance weight assessment for additive, riskless preference functions: a review (SSRI Tech. Rep. 78-3). Los Angeles: University of Southern California, Social Science Research Institute, December, 1978a.

John, R.S. and Edwards, W., Subjective versus statistical importance weights: a criterion validation (SSRI Tech. Rep. 78-5). Los Angeles: University of Southern California, Social Science Research Institute, December, 1978b.

Kahneman, D. and Tversky, A., Prospect theory: an analysis of decision under risk. Econometrica, in press.

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

Lichtenstein. S., Slovic, P., Fischhoff, B., Layman, M., and Combs, B., Judged frequency of lethal events. Journal of Experimental Psychology: Human Learning and Memory, 1978, 4, 551-578.

Newman, J.R., Differential weighting in mulit-attribute utility measurement: When it should not and when it does make a difference. Organizational Behavior and Human Performance, 1977, 20, 321-324.

Newman, J.R., Reflections on Prediction models and decision analysis models Unpublished manuscript, 1977. (Available from Social Science Research Institute, University of Southern California, Los Angeles).

Nisbett, R.E., and Wilson, T.D., Telling more than we can know: Verbal reports on mental processes. Psychological Review, 1977, 84, 231-259.

Otway, H., and Edwards, W., Application of a simple multiattribute rating technique to evaluation of nuclear waste disposal sites: a demonstration (Research Memorandum RM-77-31). Laxenburg, Austria: International Institute for Applied Systems Analysis, June, 1977.

Savage, L.J., The Foundations of Statistics. New York: Wiley, 1954.

Schmitt, N., Comparison of subjective and objective weighting strategies in changing task situations. Organizational Behavior and Human Performance, 1978, 21, 171-188.

Schmitt, N., Coyle, B.W., and Saari, B.B., Types of task information feedback in multiple-cue probability learning. Organizational Behavior and Human Performance, 1977, 18, 316-328.

Schmitt, N., and Levine, R.L., Statistical and subjective weights: some problems and proposals. Organizational Behavior and Human Performance, 1977, 20, 15-30.

Seaver, D.A., Assessment of group preference and group uncertainty for decision making (SSRI Tech. Rep. 76-4). Los Angeles: University of Southern California, Social Science Research Institute, June, 1976.

Seaver, D.A., Assessing probability with multiple individuals: group interaction versus mathematical aggregation (SSRI Tech. Rep. 78-6). Los Angeles: University of Southern California, Social Science Research Institute, December, 1978.

Slovic, P., Analyzing the expert judge: A descriptive study for stockbroker's decision processes. Journal of Applied Psychology, 1969, 53, 255-263.

Slovic, P., and Lichtenstein, S., Comparison of Bayesian and regression approaches to the study of information processing in judgments. Organizational Behavior and Human Performance, 1971, 6, 649-744.

Stillwell, W., Seaver, D.A., and Edwards, W., The effects of response scales on likelihood ratio judgments (SSRI Tech Rep. 77-5). Los Angeles: University of Southern California, August, 1977.

Van de Ven, A. and Delbecq, A., Nominal versus interacting group processes for committee decisionmaking effectiveness. Academy of Management Journal, 1971, 14, 203-212.

von Winterfeldt, D., and Edwards, W. Evaluation of complex stimuli using multiattribute utility procedures (Tech. Rep. 011313-2-T). Ann Arbor, Mich.: University of Michigan, Engineering Psychology Laboratory, October, 1973.

CONTRACT DISTRIBUTION LIST
(Unclassified Technical Reports)

Director Advanced Research Projects Agency Attention: Program Management Office 1400 Wilson Boulevard Arlington, Virginia 22209	2 copies
Office of Naval Research Attention: Code 455 800 North Quincy Street Arlington, Virginia 22217	3 copies
Defense Documentation Center Attention: DDC-TC Cameron Station Alexandria, Virginia 22314	12 copies
DCASMA Baltimore Office Attention: Mr. K. Gerasim 300 East Joppa Road Towson, Maryland 21204	1 copy
Director Naval Research Laboratory Attention: Code 2627 Washington, D.C. 20375	6 copies

Revised August 1978

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

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
Camp H. M. Smith, Hawaii 96861

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 20[illegible]
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

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

Superintendent
Naval Postgraduate School
Attention: R. J. Roland, (Code 52Rl)
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 54
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

Deputy Under Secretary of the Army
(Operations Research)
The Pentagon, Room [illegible]
Washington, DC [illegible]

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
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

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

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

<u>Department of the Air Force</u>

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 [illegible]

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
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, Virginia 22101

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 001922-T-Final	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Research on the Technology of Inference and Decision		5. TYPE OF REPORT & PERIOD COVERED Final Technical 10/77 -- 12/78
		6. PERFORMING ORG. REPORT NUMBER SSRI 79-1
7. AUTHOR(s) Edwards, Ward, John, Richard, and Stillwell, William		8. CONTRACT OR GRANT NUMBER(s) Prime Contract N00014-76-C-0074 Subcontract 76-030-0715
9. PERFORMING ORGANIZATION NAME AND ADDRESS Social Science Research Institute University of Southern California Los Angeles, CA 90007		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Advanced Research Projects Agency 1400 Wilson Boulevard Arlington, Virginia 22209		12. REPORT DATE Janu[redacted]979
		13. NUMBER OF PAGES 33
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) Decisions & Designs, Inc. 8400 Westpark Drive, Suite 600 McLean, VA 22101		15. SECURITY CLASS. (of this report) unclassified
		15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

(9) Final rept. Oct 77 - Dec 78,

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

(12) 46p.

18. SUPPLEMENTARY NOTES

(10) Ward/Edwards, Richard S./John, William/Stillwell

19. KEY WORDS

multiattribute utility
multiple cue probability learning
log likelihood ratio
validation
Bayes theorem
Delphi technique
Nominal groups technique
Degroot weights
Paramutual model

20. ABSTRACT

This reports summarizes fifteen months of research on the technology of inference and deicision. Theoretical and experimental work on two major topics; assessment of subjective probabilities and multiattribute utility theory, is discussed. Experimental work showed that individuals bringing orthogonal information to a group probabilistic inference task tend to average that information rather than multiplying as the normative rule indicates. Other

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)

390 664

experimental work indicated that for both continuous and descrete variables simple averaging of individual's probability judgments to form a group judgment did not differ significantly from behavioral interaction in final quality of the judgments as evaluated by a quadratic scoring rule. Several experiments in the area of assessing small probabilities serve to show that judgments of these small numbers are of strickingly low quality. Use of both the marker events technique and several computer program aids have not substantially improved these judgments.

Our examination of the tradeoff between modeling error and judgmental error indicates that three forms of error effect utility elicitation: model, specification, judgmental bias and random judgmental error. In a simulation study the relative effects and the tradeoffs between these three are examined. In an experimental study the sensitivity of subjects to the range of alternatives is examined and it was found that subjects are not appropriately sensitive to that range.

The validity and reliability of multiattribute utility procedures were examined in four studies. The first tested the underlying assumption of multiattribute utility elicitation, i.e., "divide and conquer". Using a bank credit scoring model as a validation criterion, it was found that use of a multiattribute procedure led to group judgments that more highly correlated with the output of the bank model than did unaided holistic judgments. The second study examined several multiattribute utility elicitation procedures in a test-retest reliability examination. Results indicated, as expected, that the more simple procedures tended to provided more reliable judgments. The third study reviewed the existing literature on subjective and objective weighting of importance dimensions. Strong evidence was found that subjective weights were valid in both a descriptive sense (policy description) and a normative sense (both when compared with statistical weights derived from the subjects holistic judgments and a criterion when one was available). The fourth study was an empirical validation of a criterion establishing technique. We have developed a variation of the multiple cue probability learning technique that allows us to teach subjects a utility function. Validation of utility elicitation procedures is therefore possible. We have tested this procedure and found it both workable and useful.