NUREG/CR-2255
SAND81-7140
AN, RX
Printed May 1982
CONTRACTOR REPORT

Expert Estimation of Human Error Probabilities in Nuclear Power Plant Operations: A Review of Probability Assessment and Scaling

William G. Stillwell, David A. Seaver, Jeffrey P. Schwartz
Decision Science Consortium, Inc.
7700 Leesburg Pike, Suite 421
Falls Church, VA 22043

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550
for the United States Department of Energy
under Contract DE-AC04-76DP00789

Prepared for
U. S. NUCLEAR REGULATORY COMMISSION

NOTICE

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, or any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for any third party's use, or the results of such use, of any information, apparatus product or process disclosed in this report, or represents that its use by such third party would not infringe privately owned rights.

Available from
GPO Sales Program
Division of Technical Information and Document Control
U.S. Nuclear Regulatory Commission
Washington, D.C. 20555

and

National Technical Information Service
Springfield, Virginia 22161

TABLE OF CONTENTS

ABSTRACT

This report reviews probability assessment and psychological scaling techniques that could be used to estimate human error probabilities (HEPs) in nuclear power plant operations. The techniques rely on expert opinion and can be used to estimate HEPs where data do not exist or are inadequate. These techniques have been used in various other contexts and have been shown to produce reasonably accurate probabilities. Some problems do exist, and limitations are discussed. Additional topics covered include methods for combining estimates from multiple experts, the effects of training on probability estimates, and some ideas on structuring the relationship between performance shaping factors and HEPs. Preliminary recommendations are provided along with cautions regarding the costs of implementing the recommendations. Additional research is required before definitive recommendations can be made.

ACKNOWLEDGEMENTS

This work was supported by the U.S. Nuclear Regulatory Commission, Office of Nuclear Regulatory Research, with Mr. James W. Pittman, Human Engineering Section, Human Factors Branch, Division of Facility Operations, as program manager. It is part of a larger effort directed by Sandia National Laboratories on human reliability analysis in nuclear power plant operations. Thorough and penetrating comments on earlier drafts of this report by Sandia personnel have been particularly helpful in preparing this report. The technical monitor, Dr. Louise Weston of the Statistics, Computing, and Human Factors Division, Sandia National Laboratories, has been particularly helpful.

1.0 INTRODUCTION

This report reviews the existing psychological judgment and scaling literature which may be relevant for the use of expert judgment in estimating human error probabilities (HEPs) in nuclear power plant operations. A good deal of developmental work has been done which aids the expert in identifying factors related to the magnitude of an HEP as well as pitfalls in the HEP judgment process (see, for example, Swain 1967, 1971, 1978; Swain and Guttmann, 1975, 1980). Now, a set of actual scaling techniques needs to be developed that can transform expert judgment into the needed HEPs under varying circumstances in an efficient and valid manner.

Three bodies of literature are of particular relevance to this topic: (a) probabilitiy assessment (for reviews see Einhorn and Hogarth, 1981; Slovic, Fischhoff, and Lichtenstein, 1977; Spetzler and Stael von Holstein, 1975; Wallsten and Budescu, 1980), (b) psychological scaling and measurement theory (Cliff, 1973; Roberts, 1979; Torgerson, 1958), and (c) human reliability analysis (Meister, 1971; Swain, 1978; Swain and Guttmann, 1980). This review will concentrate on (a) and (b) and their implications for (c).

There is considerable overlap between the scaling literature and the literature on probability assessment. Many of the theoretical and experimental approaches to probability assessment have been undertaken by experimental psychologists (for instance, Edwards, 1954), measurement theorists (Fine and Kaplan, 1977; Krantz, Luce, Suppes, and Tversky, 1971), or those primarily interested in psychological scaling (e.g., Roberts, 1979). There is, however, much that is unique to each, and we will structure this review in a manner appropriate to these unique contributions. A glossary defining terms frequently used in the literature is appended to this report.

The review will be in three parts. The first part (Section 2.0) will concentrate on the literature on probability assessment. The next section (3.0) of the review will cover relevant work from psychological scaling and measurement theory. As previously stated, there is some overlap with probability assessment literature, but there are a number of scaling methods and judgment paradigms that have been poorly explored as probability elicitation techniques. For example, only recently has some preliminary testing been done utilizing paired comparison judgment of event probabilities (Edwards, John, and Stillwell, 1977, 1979; Lichtenstein, Slovic, Fischhoff, Layman, and Combs, 1978), although preliminary development of a technique was accomplished some time ago (Blanchard, Mitchell, and Smith, 1966; Thurstone, 1927; Torgerson, 1958).

Finally, Section 4.0 will discuss some of the implications of the literature, and directions we might go in developing a family of techniques for HEPs. We also address some likely problems, special considerations in judgment situations that would affect elicitation, and suggestions for a measure of the strength or quality of the final judgments elicited.

2.0 SUBJECTIVE PROBABILITY ASSESSMENT

This section reviews the use of expert judgment to estimate probabilities, the various procedures used to assess subjective probabilities, and some special problems and aids for such assessments. The first subsection discusses and justifies the use of experts to estimate subjective probabilities. Section 2.2 then reviews various assessment procedures that have been developed, tested, and applied. The role of multiple experts and derivation of probabilities from their judgments are reviewed in Section 2.3. Section 2.4 discusses biases that often occur in subjective probability assessments and Section 2.5 describes the role of training. In Section 2.6, we then summarize the major points in the previous subsections relevant to the estimation of HEPs.

2.1 Use of Expert Judgment for Assessing Probabilities

We are motivated to examine the role of expertise in human error probability judgment for three reasons. First, we wish to ascertain whether there is a demonstrated ability of experts to provide good estimates of unknown probabilities. Second, we want to explore the contribution of using experts who bring different types of expertise to the judgment task. And, third, if experts are better at probability estimation than are nonexperts, that fact would be a motivating argument for training as a method for improving probabilistic judgment.

Research on probability elicitation has examined the role of expertise in the judgment of probabilities without specifically stating what constitutes an expert. Instead, individuals whose expertise is readily agreed upon, but not detailed, are used as judges to study the quality of expert judgment of probability. Four substantive areas have been studied in sufficient detail that general inferences can be made. They are (1) military and intelligence, (2) business, (3) medicine, and (4) weather prediction. There is conflicting evidence as to the quality of experts' probability estimates, with evidence generally varying according to the substantive area.

The evidence from area 1 is mostly classified and, therefore, inaccessable for present purposes. What is accessable, while based on the use of expert subjects, is based on hypothetical problems, which may limit its generality. The general conclusions that can be drawn from it are (1) military and intelligence experts prefer to respond in numerical form when expressing uncertainty, (2) a large reduction in miscommunication has been shown from the use of probabilities to communicate uncertainty (Kelly and Peterson, 1970), (3) the reliability intrasubject product moment correlation for repeated judgments of the expert judgments has been good with a mean of .79 and a range of .42-.97 (Johnson, 1977), and (4) satisfactory use of subjective probability estimates has been made by the CIA in an applied setting, although the classified nature of the work makes the available details somewhat sketchy

(Zlotnick, 1972). What we do know is that Bayesian methods of probability estimation and revision have and are being used to solve traditional intelligence processing problems, for example, aggregation of information from diverse sources, systematizing communication, and information reduction.

The areas of 2, 3, and 4 are more interesting, and somewhat in conflict. Most of the negative evidence for experts' ability to report subjective probabilities comes from the business literature. A large number of studies have shown that bankers (Stael von Holstein, 1972), security analysts (Bartos, 1969, cited by Winkler, 1972), and stock analysts (Stael von Holstein, 1972) often cannot out perform even a simple strategy in which a uniform distribution (essentially a "no information" strategy, for example, each of n events is assigned a probability of 1/n) is used to predict the performance of some random variable. One study did, however, show that bankers were able to predict the fluctuation of the interest rate on certificates of deposit for a nine month period (Kabus, 1976). One problem with many of these studies, however, is that the events being predicted (e.g., stock prices) are inherently unpredictable (nearly random).

Evidence in the medical field is mixed, but somewhat positive. Ludke, Stauss, and Gustafson (1977) asked nurses and senior nursing students to estimate distributions of familiar physiological variables (for example, weight of an American baby at birth; systolic blood pressure of American males, ages 18-79; hematocrit level of American males, 18-79) using several assessment methods. All methods proved both reliable (average test-retest correlation across methods = .986) and accurate (across the probability distribution, divided into ten sections for computational ease the average difference in area for an individual section = .037). Thus, these experts could be said to be extremely good at providing subjective probability distributions, no matter what the elicitation method. Methods and results will be discussed in more detail in Section 2.2.

There is, however, some tendency to overestimate the probability of events with severe consequences (Lusted, 1977). This may well be explained by a desire to be sure such severe consequences are avoided, thus focusing special attention on these events. Still, medical evidence generally indicates that experts are able to provide subjective estimates of probabilities that agree with relative frequency estimates for the same quantities. Lusted (1977) gives calibration curves* for

*A calibration curve plots assessed probabilities on the abscissa versus relative frequencies on the ordinate. Perfect calibration would result in a calibration curve that is a straight line from (0,0) to (1,1) indicating, for example, that for events assigned a subjective probability of .70, 70 percent occur.

three events (skull fracture, pneumonia, and extremity fractures) that show what appears to be extremely well-calibrated responses (although validity coefficients and raw data are not reported). DeSmet, Fryback, and Thornbury (1979) do show data that permit calculation ot a validity coefficient for skull fracture estimates that is quite good ($r = .88$).

In seeming contradiction to the equivocal evidence presented above, weather forecasters are remarkably good at providing subjective probability judgments. Average deviations from perfect calibration for precipitation forecasts ranged from .028 to .044 in two studies (Murphy and Winkler, 1974, 1977b). Additional results are reviewed by Wallsten and Budescu (1980) and Murphy and Winkler (1977a). It is interesting to note that one factor in evaluating forecasters for promotion is their forecasting performance as measured by the Brier* score, thus ensuring high motivation as well as extensive experience with making probabilistic predictions. On tasks other than predicting precipitation for which forecasters do not routinely make probabilistic predictions, their subjective probability estimates are more in line with those of other experts (Stael von Holstein, 1971b).

Our second motivation for examining the role of expertise in probability judgment comes from our desire to determine the optimal composite structure of expert groups. It seems likely that experts with different training and substantive expertise, who all have sufficient knowledge relevant to the events considered, would bring more information to the estimation problem than would experts of very similar backgrounds. If this is assumed to be true, the question still remains as to whether those with expertise from different areas are able to combine their differing information to arrive at better estimates of some unknown quantity. In a review of the literature comparing individual and group judgment, Seaver (1976) states among his major conclusions that "...a larger diversity of individual opinion among group members will lead to greater superiority of the group judgment over individual judgments," and cites several sources as evidence for his conclusion. One rationale for this conclusion comes from psychological test theory (Gulliksen, 1950; Nunnally, 1967) which proves an increase in quality (validity as measured by the correlation between the average individual response and the true value) of group judgment will occur as the heterogeneity of the group increases.

*The Brier score for a precipitation forecast of probability r is

$$s(r) = (1-r_j)^2$$

where r_j is the probability assigned to the event that actually occurs (precipitation, no precipitation). The Brier score is a strictly proper scoring rule which means that the expected score is minimized when the forecast, r, is equal to the forecaster's true subjective probability of precipitation, p.

This recommendation should be approached with caution, however. An assumption underlying the conclusion of the superiority of heterogeneous over homogeneous groups is that more total information will translate into better group judgments. The possibility that group members with "poor" information will sway the group away from quality estimates or will adversely affect the average is not at issue. Nor is whether an extreme group or individual, by their extremity, demonstrates either more or less information, different information, or possible bias. Outlying judgments should be identified and examined. This point will be discussed in more detail in the section on the use of multiple experts.

And, finally, our third motivation for examination of experts is to compare them with nonexperts in order to infer the role training might have in probability judgment. The evidence that experts perform better at probability estimation than do nonexperts is not extensive. While several studies have failed to find strong differences between the probabilistic assessment abilities of experts (e.g., doctors) and those of lesser expertise (medical students or even lay persons), those that have found differences report the crucial factor to be statistical rather than substantive expertise (Stael von Holstein, 1971a, 1972; Winkler, 1967). These findings will be discussed in detail in Section 2.5. In the judgment of HEPs, experts in the complex human role in nuclear power plant operation would make the appropriate assessments. Thus, the evidence suggests that when some training is possible, the effort would best be spent to give the experts some understanding of the nature of probabilistic thought.

2.2 Probabilistic Assessment Techniques

For a number of different elicitation problems, the method of eliciting an uncertain quantity has been shown to affect the quality of the resulting judgment. This is just a small part of the much larger psychological finding that the same question, when asked in two different ways, will often result in two different answers. Because of both the vast breadth of this literature and its diffuse relationship to probability judgment, we will confine our review here primarily to literature directly relevant to the judgment of subjective probabilities. Additional discussion of this topic related to psychological scaling in general is included in Section 4.0.

What is the best technique for eliciting subjective probabilities? Literature on the subject is confounded by problems of artifactual results; lack of generality of task, stimuli, and subjects (most of the studies discussed in the following paragraph used college students or paid volunteers as subjects); as well as a lack of applied relevance. Often the conclusions are only of use when the nature of the probability being sought is known in detail, certainly not the case in most applied contexts.

Still, a large number of studies have directly compared elicitation techniques for generating estimates of probabilities and some general conclusions are warranted. Among direct methods for discrete quantity estimation (i.e., simply asking the assessor for the relevant number usually in probability or in odds), written responses are more consistent (less variation over time) than are verbal responses, and also tend to be more extreme (Domas, Goodman, and Peterson, 1972, Goodman, 1973). Goodman (1973) directly compared a number of response modes using the data from previously unpublished studies conducted at the University of Michigan. Two of the studies used scenarios of a ten-years-in-the-future political-military situation. Subjects (college students) were given a complex summary of the history of the world ten years hence. They were asked to judge the likelihoods of alternative hypotheses such as "Russia and China are about to attack North America," using as data hypothetical intelligence reports. For example:

> "At 0630 this morning, two full squadrons of conventional submarines sailed from Vladivostok. They steamed in a southerly direction until they were clear of the harbor, and then submerged. Evaluation: probably a routine exercise although this is an unusually large force."

Thus, there were no normative answers against which the estimates could be compared.

The third and fourth studies (later published as Domas, Goodman, and Peterson, 1972) used variations of an abstract task where subjects are asked to say which of two distributions was more likely to have produced a given sample of data. In the third study, subjects were asked to act as analysts for a hypothetical commercial shipping firm. Using data as to the location of competitor ships, gross tonnage, percent of capacity and fuel taken on at port of departure (each of these was given a carefully defined, independent relationship to probability), subjects estimated the likelihood that the competitor ship was bound for port A versus port B. In the fourth study, the task was to judge which of two parent populations produced a sample of seven inch, blue and yellow sticks. The relative length of the two colors was the random variable.

For each of the studies, correlations between the true probabilities, where they could be calculated, and the subjects' responses were high (range .91-.991) and the intercepts of regressions of judgments on true were near 0.0. Thus, the slope of the regression line is an important descriptor of the relative magnitude of responses. A value of 1.0 is optimal when responses are compared with true values. When responses are being compared directly with other responses the magnitude represents the average relative size of the estimates. Study three compares verbal and written with true values, with verbal estimates always smaller (average slopes .85 versus 1.3 and .88 versus 1.1 for verbal and written responses respectively).

Reliability (consistency within judges) figures for the Goodman studies suggest that written responses tend to be more reliable than are verbal, although all were quite reliable. Comparable average reliability coefficients (test-retest) were .97, .66, and .95 (ranges not reported) for verbal estimates and .92, .89, and .97 for written for studies one, two, and three, respectively.

Goodman (1973) also examined the impact of scale spacing, logarithmic versus linear, as did Stillwell, Seaver, and Edwards (1977). Stillwell et al. asked college students to assess the relative likelihood of two-colored seven inch sticks being produced by two normal distributions. Sticks from one population had a mean length of one color of x inches. Sticks from the second population had a mean length of the second color of x inches. Population variances were constant and the value of x was varied. In a second experiment, the subjects assessed the likelihood that a lake was polluted, given the bacteria count from a sample of lake water. The distributions of polluted and nonpolluted lakes for this bacteria were described as possible alternative data generators. The overall study was to examine the effects of scale spacing and the magnitude of the upper endpoint.

Both Goodman (1973) and Stillwell et al. (1977) found strong effects for spacing. Goodman (1973) found the average slope of logarithmic responses regressed on linear responses was 1.4 Stillwell et al. found average slopes of .2 to .44 for linear scales and .40 to .62 for logarithmic scales in their first study. In the second study, across eight different conditions, the average slope for logarithmic responses was .14 higher than for linear responses. Stillwell et al. also found that when responses were made in odds the upper endpoint of the scale affected responses, even though subjects were encouraged to use larger responses if they were appropriate. Average slopes were .43, .44, and .75 for endpoints of 10C:1, 1,000:1, and 10,000:1, respectively, in the first study; and, in the second, average slopes were .41 and .62 for endpoints of 100:1 and 10,000.1, respectively.

A pair of studies examined responses in probabilities versus odds (Fujii, 1967; Phillips and Edwards, 1966). Odds responses are the ratio of the probability of the occurrence of an event or truth of hypothesis to the probability of its nonoccurrence or falsity, i.e., odds equal p/(1-p). Each study used a book-bag and poker chip paradigm where the problem was to estimate the relative likelihood that a given sample of blue and red poker chips came from a bag that held predominately blue chips or one that held predominately red chips. The exact composition of the two bags was specified so a correct answer could be calculated, against which the subjects' responses could be compared. Each of these studies found subjects who responded in probabilities to give smaller answers than did those who responded in odds or likelihood ratios. (In these studies, odds are the ratio of the probability of one hypothesis, e.g., the bag holds predominantly blue chips, to the probability of the alternative hypothesis given a datum, e.g., a chip has been observed. Likelihood ratios are the ratio of the probability of a particular datum given one hypothesis to its probability given the alternative hypothesis.) Data analysis for each of these experiments consisted of plots comparing the response values to those calculated by Bayes' Theorem, the normative rule. In all cases, the responses for subjects responding in odds or likelihood ratios were closer to true values than were responses in probabilities. Congruent with these findings are those of most of the studies discussed above. Where objective probabilities could be determined, the larger responses were also closer to the true values (Domas et al., 1972; Fujii, 1967; Phillips and Edwards, 1966; Stillwell et al., 1977).

The finding that likelihood ratio or odds responses are generally better than direct estimates of probabilities suggests another assessment approach. If assessors are better able to make judgments of ratios of event likelihoods than they are absolute value estimates, it seems logical that they will be better still at making simple judgments of the directional relationship, i.e., more likely than or less likely than. This type of judgment is known as paired comparison.

Early work in psychological scaling developed the theoretical foundation to build subjective scales from paired comparison judgment (Thurstone, 1927, 1931; Torgerson, 1958). But only recently have these procedures been used for subjective probability assessment (Edwards et al., 1977, 1979; Lichtenstein et al., 1978; Rigby and Edelman, 1968). Initial studies have shown positive results. The study by Lichtenstein et al. (1978) found that assessors were generally able to correctly judge the direction of the relationship between event pairs, although the estimates of the magnitude were smaller than were the true values. Hunns and Daniels (undated) show in a very limited test that a direct application of Thurstone's Case V of the law of comparative judgment produced an estimate of a single event probability that was remarkably close to the value calculated from relative frequency data.

The law of comparative judgment is a model relating the proportion of times one of a pair of stimuli is rated greater than the other on a given attribute to the scale values. The model is derived from three postulates: (1) each stimulus gives a differential impact on the psychological scale of interest; (2) the impact of the stimulus is not constant, but will vary in terms of a specific statistical process; and (3) the mean and standard deviation of the distributions of aggregate responses across judges can be used to represent the stimulus scale value and discriminal dispersion, respectively. Cases I and II of Thurstone's law are merely different ways of stating the complete form of the law, which is not solvable. Cases III, IV, and V are derived from sets of simplifying assumptions which provide solvable and, therefore, usable forms of the law.

Other attempts to use this form of judgment to produce scales of event probabilities have been less fruitful. Edwards et al. (1977, 1979) used a series of paired comparison judgments to sequentially narrow the range of the probability of an event by having assessors compare its likelihood with that of a series of other events with known probabilities. They found that rank order correlations between the probabilities derived in this manner and true probabilities of the events were small to moderate. It should be noted, however, that the range of true values was small (.005-.02) and, therefore, these correlations are suspect. These studies indicate some potential for paired comparison judgments as the basis for estimating probabilities, but also suggest that the paired comparison techniques that have been tested to date may not be adequate and need to be revised.

Additional studies indicate that consistency checks should be included in eliciting probability estimates. Beach (1974) found a surprising lack of consistency between probabilities derived from direct estimates and those from an indirect procedure (one in which the response is something other than a probability, likelihood, or odds that is then converted into a probability), as to a lessor degree did DuCharme and Donnell (1973). On the other hand, the findings of Domas et al. (1972) show that inconsistency, when pointed out to the estimator, is an extremely useful way to improve the quality of the final judgment. The way in which the comparison of the results of two elicitation methods, usually one direct and one indirect, aids the estimator is not well understood, but the finding is clear, as is the implication for a combined probability assessment methodology.

As in the case of discrete probability estimation, much research has examined the impact of elicitation techniques on the judgment of probability distributions. Alpert and Raiffa (1969), in testing four response modes, found only small differences in the number of true values falling outside the extremes of the requested ranges (known as "surprises") resulting from techniques which asked for the median, interquartile range, and either the .01 and .99 values, the .001 and .999 values, the "minimum" and "maximum" values, or an "astonishingly low" and "astonishingly high" value. Selvidge (1975) found that responses were better when seven fractiles* were assessed rather than five.

In another study of elicitation techniques, Seaver, von Winterfeldt, and Edwards (1978) examined the proportion of "surprises" resulting from two assessment factors, the way in which uncertainty was specified (odds, odds on a logarithmic scale, or probability) and the type of response required (uncertainty measures or values of the variable). Elicitation methods requiring values of the unknown quantity as responses used questions such as "What is the number of people such

*A fractile is the value of a random variable that corresponds to a specified point on the cumulative probability distribution. For example, the .50 fractile is the median.

that your odds are 3:1 that the true population of Canada is less than that number?" For the probability group substitute, "probability is .75" for "odds of 3:1." For questions requiring uncertainty measure as responses, the question might be "What is your probability that the population of Canada is less than 130 million people?"

Their results showed that subjects responding in probabilities or odds were much better than those responding with fractiles, fractiles expressed in odds, or log odds. Not only were the probability and odds responses better as measured by a proper scoring rule, but they were also much better in terms of the proportion of surprises or "surprise index." The probability and odds procedures resulted in 4.7 percent and 4.5 percent (compared to the correct five percent) of responses outside the interval assessed to contain 95 percent of the values as opposed to 19.9, 24.2, and 34.2 percent for the log odds, odds fractile, and probability fractile procedures, respectively. For the interquartile (IQ) range (where the correct proportion would be .50), differences between procedures were small with one exception. The log odds procedure resulted in less accurate assessments than the other procedures. Proportions of true values outside the IQ range were 43.0, 52.9, 46.8, 57.9, and 69.1 for the probability, odds, odds fractile, probability fractile, and log odds procedures, respectively.

Finally, a study by Barclay and Peterson (1973) compared the tertile method (i.e., the judge is asked to provide the fractiles .33 and .67) with a "point" method in which the assessor is asked to give the modal value of the uncertain quantity, and then two values, one above and one below the mode, each of which is half as likely to occur as is the modal value (i.e., points for which the probability density function is half as high as at the mode). They found the tertile method resulted in better assessments, although both resulted in too many estimates outside the assessed values. With the tertile method, 77 percent of the true values fell outside the two tertiles, where 67 percent should occur. The point method resulted in 61 percent of the true values being outside the two values assessed, where only 25 percent should occur.

Two additional studies addressed response mode effects. John and Edwards (1977) used several different distributions (normal, bimodal, skewed, beta) to generate samples of individual stimuli and asked subjects to assess the distribution that produced each sample. They tested three response modes, a fractile method, a probability method (each of these are the same used in Seaver et al, 1978), and a method in which assessors simply drew the distribution on graph paper and labeled the axes. Distributions were compared in terms of the maximum deviation between the assessed and the true density functions. The graph method was generally found to be better, but the probability procedure was better in a few cases where the true distribution was unusual (bimodal, for example).

The second study (Ludke, Stauss, and Gustafson, 1977) used three criteria to compare elicitation techniques: accuracy, reliability, and acceptability. Questions involved physiological variables as discussed in Section 2.1. The distribution of each variable, established from actuarial records, was broken into ten intervals. The intervals were of equal length in terms of the variable being measured, and covered the entire range of the variable. All subjects, nurses and senior nursing students, were familiar with each of the uncertain quantities and had both "hands-on" experience and textbook knowledge regarding the topics of the questions. Each nurse was randomly assigned to one method, providing 36 nurses per method. Each subject participated in two sessions, three weeks apart.

Five methods were examined. In the first, called the probability method, the subjects were asked to give the probability associated with each of the ten intervals of the uncertain variable. For example, one probability question was "Given a population of 1,000 American males, ages 18-79, how many would have systolic blood pressure readings (in millimeters of mercury) in each of the following ten intervals?" The probability instructions also asked the subjects to normalize their responses and write them on a linear scale provided on the answer sheet.

For the log odds method, subjects responded to questions similar to "How likely is it that an American male, age 18-79, will have systolic blood pressure reading in each of the following ten intervals?" Instead of answering the question directly, subjects were instructed to rank the intervals from the one that contained the most males to that with the least. Then they were asked to estimate the ratio of males in each interval to that of the interval ranked directly below it. They recorded the ratios on a logarithmic scale. The log-log method was the same as the log method except that the subjects recorded their answers on a double logarithmic spaced scale. The fourth method, called the ranking method, was developed by Smith (1967).

Subjects performed the same interval ranking task as for the log odds method. In addition, they assigned probability estimates to both the first and last ranked intervals, and provided estimates of the rank order of the first order differences. First order differences were determined with questions such as "Now that you have ranked the intervals, consider each pair of intervals formed by taking an interval and one of the adjacent ranked intervals. There are nine of these pairs, 1 and 2, 2 and 3, and so on up to 9 and 10. Now judge which of the differences in likelihood is largest, second largest, and so on."

Using mathematical ideas taken from Kendall (1962) about the average distance between ranks of random normal deviates, it can be shown that the expected values of the relationships between the first order differences in descending order are:

$$(1)\qquad \begin{cases} \frac{1}{n}\left(\frac{1}{n} + \frac{1}{n-1} + \dots + \frac{1}{2} + \frac{1}{1}\right) \\ \frac{1}{n}\left(\frac{1}{n} + \frac{1}{n-1} + \dots + \frac{1}{2}\right) \\ \cdot \\ \cdot \\ \cdot \\ \frac{1}{n} \cdot \frac{1}{n} \end{cases}$$

The rationale underlying this is that if a magnitude is divided randomly into n parts, these formulas give the expected values of the parts in descending order.

This set of relationships is used to turn the ranking judgments into probability estimates for the distribution areas.

Smith (1967) gives the following example of the use of the technique. Assume that an expert believes that there is some possibility that the percent of market which a product will capture may be anywhere in the range from zero to 100%. The "expert" may reason that the following rankings (in order of ascending probability) should be assigned to the various possible intervals:

(2)	Interval:	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
	Ranking:	1	2	7	8	10	9	6	5	4	3

and that the differences (in ascending order) should be ranked as follows:

(3)	Interval:	0-10	10-20	90-100	80-90	70-80	60-70	20-30	30-40	50-60	40-50
	Ranking:	1	2	3	4	5	6	7	8	9	10
	Ranking of Differences:	1	3	2	9	6	5	4	8	7	

Further, suppose the expert feels that there is only a probability of 0.005 that the product will capture less than ten percent of the market, and that the best estimate of the probability that the percent of the market captured will be between 40 and 50 percent is 0.20.

Since we have nine differences, the expected proportions become as given by (1). Rearranging according to the ranking of differences in (3), we obtain:

(4) Proportions: .0123,.0421,.0262,.3143,.1106,.0829,.0606,.2032,.1477

The total range of values being .9998, and the range of expected probabilities being 0.20 - 0.005 = 0.195, we multiply the values of (4) by 0.195/.9998 = .1950 to get as the differences:

(5) Differences: .0024,.0082,.0051,.0613,.0216,.0162,.0118,.0396,.0288

Finally, accumulating from 0.005, we get:

(6) .0050, .0074, .0156, .0207, .0820, .1036, .1198, .1316, .1712, .2000.

The probabilities in (6) total to only 0.8569 instead of 1.0000. Therefore, to get our final relative probabilities, we multiply each number in (6) by 1/.8569 = 1.167. The results are:

(7) .0058, .0086, .0182, .0242, .0957, .1209, .1398, .1536, .1998, .2334.

From (2), we have that these relative probabilities should be associated with the various intervals as:

Interval:	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Ranking:	1	2	7	8	10	9	6	5	4	3
(8) Relative Probability:	.0058	.0086	.1398	.1536	.2334	.1998	.1209	.0957	.0242	.0182

With the fifth method, bisection, respondents were asked to give a value of the uncertain quantity such that they believed that it was equally likely that a randomly selected individual will have a value above or below the given value. The respondent was then asked to assume that he or she knew that the selected individual had a value above (below) the value dividing the distribution in half. They were then asked for a value that divided the halves in half. This procedure was followed until the distribution was divided into eight parts. The experimenters then interpolated to get probabilities for the ten areas comparable to those from the other techniques.

The accuracy of the various methods was assessed by computing an error score by summing the absolute differences between the estimated and actuarial values across ten question intervals. This error score was computed within assessment method and averaged across subject and assessment session. Reliability of the methods was compared by taking test-retest correlations between the two sessions across subjects and within distributions. As with the accuracy calculations, the reliability correlations were calculated comparing the ten intervals of the probability distributions assessed in session 1 with those assessed in session 2. The acceptability of the methods was simply the rating by each assessor for each method.

While the subjects found the probability and bisection methods to be more acceptable than the other methods, the ranking method proved to be the best in terms of accuracy. In addition to its accuracy in reproducing the true distributions, the ranking method was also equal to or better than the other methods in terms of reliability. Table 1 shows the accuracy and reliability results.

TABLE 1

Results: Ludke *et al*. (1977)

	Bisection	Probability	Ranking	Log Odds	Log-Log Odds
Reliability (Average Test-Retest Correlation)	.981	.983	.999	.987	.980
Accuracy (Average Error Per Section of Distribution)	.091	.071	.060	.070	.081

	Ranking, Log Odds, and Log-Log Odds Together	Bisection and Probability Together
Acceptability		
Easy to Use	31%	48%
Tiring and Boring	54%	36%
No Confidence in Estimates	71%	41%
Willing to Give Estimates for Actual Use	29%	50%

While the ranking method proved more accurate in Ludke et al. (1977), it was criticized earlier by Morrison (1967) and Green (1967) on practical grounds. They, for instance, point out that for a distribution broken into ten parts, the judge who used paired comparisons to arrive at a ranking would be required to make (9!)/(7!)(2!) = 36 paired comparison judgments of first order differences, as opposed to a few judgments for other procedures. This is in addition to the initial paired comparisons of sections of the distributions needed to get the first level ranking. A consideration important for experts estimating HEPs is the difficulty of getting an expert to make a large number of judgments reliably with only a vague understanding of the relationship of the judgments to the desired outputs.

Ludke et al. (1977), notwithstanding the controversy over the ranking procedure, seem to show that experts will and can provide the inputs necessary for Smith's procedure. It should be noted, however, that the respondents in that study preferred other procedures to the ranking one, which in some instances may be important.

Two additional papers bear mention. We discuss them separately because, unlike the research discussed above, they do not use experimental data to justify their recommendations. Instead, they are based on common sense and a good deal of applied experience in the assessment of probabilities for decision making.

The first is a paper by Spetzler and Stael von Holstein (1975) that reviews the probability assessment procedures used in probability encoding by the decision analysis group at Stanford Research Institute. These procedures have been updated and modified based on several years of application. Those interested in relevant applications of the procedures are referred to Brown, Kahr, and Peterson (1974), Howard, Matheson, and North (1972), and Spetzler and Zamora (1971).

The procedure arrived at consists of five parts. Briefly, they are as follows:

1. Motivating - The analyst attempts to get an understanding of the assessor's understanding of probabilities. Rules of probability are discussed and possible motivational biases explored.

2. Structuring Event - The desired uncertain quantity is defined. Conditionalities are explored. The expert may think of the event probability as conditional on other events. If so, the conditionalities are explicitly incorporated into the model to avoid the difficulties of dealing with combinations of uncertain quantities. The quality of the event definition is explored. For example, it is not meaningful to ask for "the probability distribution over the price of wheat in 1975,"

because of poor specification of the event. However, "the closing price of 10,000 bushels of durum wheat on June 30, 1975, at the Chicago Commodity Exchange" is a well-defined quantity.

3. Conditioning - The major purpose of this phase of the encoding process is to find out how the assessor makes estimates. Possible heuristic approaches are looked for in the subjective thought process, and their impact on the optimality of the response is explored with the assessor.

4. Encoding - A direct and an indirect procedure are used to elicit judgments. The direct procedure is simply to ask for either probabilities or odds or both. The indirect procedure uses a lottery and a probability wheel--a yellow and blue disk with the proportion of yellow constituting a variable P in a gamble. The assessor is asked to compare two gambles with equal amounts to win and lose. In one gamble the assessor wins if the event of interest (in this case, the human error) occurs; and, in the other case, the assessor wins if a spinner on the probability wheel lands on the yellow portion of the wheel. The assessor is asked to change the portion of yellow until he or she is indifferent between the two gambles. The indifference proportion P is then taken as the probability of the event of interest.

5. Verifying - Responses are checked for consistency and the assessor is given feedback on the implications of his or her responses. This is done in the form of further gambles, each pair of gambles representing what would be indifferences if the assessor were coherent and consistent.

One final point should be noted about this procedure. The length of the interview ranges from 30 to 90 minutes, depending on such things as importance and complexity of the uncertain quantity as well as the subject's previous experience with probability estimation. This constitutes an extreme handicap for the judgment of HEPs where expert time is both costly and hard to get and many probability estimates are needed. Of course, a relaxation of some portions of the procedure could be tried as a time saving device.

A second applied article is of particular relevance to the HEP judgment problem. Selvidge (1973), noting many applied situations where events as rare as 10^{-8} or 10^{-9} are of enough importance because of their very favorable or extremely severe consequences to warrant an attempt to quantify them, developed a three-step procedure for their elicitation. The steps are as follows:

1. Description and decomposition of the event and its setting.

   - Specify the event to the extent necessary so that the assessor is able to judge, without ambiguity, what probability is being asked for.

   - Identify causes of the event: sets of mutually exclusive initiating events and sequences of events.

   - Identify assumptions about elements of the event, particularly the populations of objects being considered and the time of exposure.

2. Express uncertainty in relative terms.

3. Numerically express the probability.

Phase 1 of Selvidge's procedure is similar to that of Spetzler and Stael von Holstein in that it attempts to acquaint the assessor with the exact event of interest. It does, however, include an addition, that of causal linkages. Several authors have noted the subjective impact of causal thinking on probabilistic judgment (see, for instance, Tversky and Kahneman, 1977), but Selvidge is the first to suggest that it explicitly be incorporated into an assessment procedure. The procedure also includes explicit consideration in the structuring process of the population of events of which the defined event is a subset, an element whose impact is often neglected in the judgment of uncertainty (Kahneman and Tversky, 1979).

Phase 2 of the procedure is another addition to usual elicitation methodologies. Rather than moving directly to the estimation of the unknown probability, Selvidge suggests that the judge be asked to note the likelihood of the event relative to several other rare events. She goes on to suggest the development of a master list of these events with known probabilities. Once the judge had made a number of these judgments, boundaries on the probability of the unknown event could be established to aid the judge. This aid could be of much help in light of the established problem subjects have with judging the extremes of the probability scale (Edward *et al.*, 1977. 1979).

The final phase in the procedure involves direct estimation of not only the probability of interest, but also the probabilities of sub or super populations of events elicited during earlier phases. The judge is also asked to determine probabilities within the causal linkages. Inconsistencies are then pointed out to the judge, who is asked to reconcile them for the final estimate.

2.3 Use of Multiple Experts in Assessing Probabilities

Using the information and judgment of multiple experts can have something positive to offer to the judgment of unknown probabilities. Not only is an equal or greater amount of information being brought to bear, but also certain statistical advantages accure (for example, reduced error variance around the estimate, and a measure of the agreement between judgments). Seaver (1976), in a review of numerous studies of group judgment, summarizes the following points regarding multiple assessors:

1. The product of multiple individuals, be it arrived at by aggregating the individual judgments or asking the individuals to come to a consensus estimate, will, on the average, be more accurate than the individual judgments primarily due to a decrease in the variance of the error around the true value. This is simply the statistical artifact that the mean of several values will always have a smaller variance around the true value than will the values from which it was derived.

2. The improved accuracy of group judgments over individual judgments appears to hold for factual judgments rather than value judgments.

3. A larger diversity of individual sources of information and opinion among group members will lead to relatively more accurate judgments from groups compared with individual judgments.

Two general approaches using multiple experts to assess probabilities have been extensively explored. The first is called the statisticized group approach. Individual estimates are made by multiple assessors and mathematically aggregated. Procedures that have been tried include everything from a simple arithmetic average of the individual estimates in the discrete case, to extremely complex procedures (for a review, see Seaver, 1976). Some attempts have been made to weight differentially the judgments of different individuals according to their expertise using either self-ratings, previous performance, or ratings by others (Seaver, 1978; Stael von Holstein, 1971b, 1972; Winkler, 1971). In all instances, the weighting procedure has had virtually no effect on the judgments. The more complex aggregation procedures generally rely on some version of Bayes' Theorem to aggregate the individual judgments into a combined group judgment (Dalkey, 1975; Morris, 1974, 1977; Winkler, 1968).

The second approach is to ask the individuals to interact as a group to come to a "group" estimate of the unknown quantity, either through reaching a consensus or through subsequently using some mathematical aggregation technique. The constraints put upon their interaction and the instructions they receive before that interaction constitute the major differences among procedures. In an extreme case, the Delphi procedure (Dalkey, 1969b) requires that the individuals not interact face-to-face at all, but, instead, they make judgments and are given feedback about what the group as a whole responded (usually represented by summary statistics), and a new set of judgments are made. If, after some number of iterations, no consensus is reached, mathematical aggregation is often used to provide the final group estimate. For most of these procedures, some convergence of estimates is expected or even required. Seaver (1976; 1978) reviews both the statisticized group and behavioral approaches to group estimation.

Results of the statisticized group procedures present a relatively consistent picture. In a large number of studies, small or no differences have been found in the quality of the group product among the aggregation procedures being compared, although often subjects preferred the simpler procedures (Gough, 1975; Rowse, Gustafson, and Ludke, 1974; Seaver, 1978; Stael von Holstein, 1972; Winkler, 1971; Winkler and Cummings, 1972). In fact, an important practical finding is that some of the more theoretically sophisticated techniques have proved difficult or impossible to apply in practice (Dalkey, 1975; Morris, 1971, 1974, 1975). Where differences have been found, they tend to support the notion that a simple weighted additive combination method is at least as good as or better than other more complex procedures (Seaver, 1976).

Much of the use of behavioral interaction has focused on two techniques: Delphi, developed by Dalkey and Helmer at the Rand Corporation in the late 60's (e.g., Dalkey, 1969a, 1969b), and the Nominal Groups Technique (NGT), developed at the University of Wisconsin (e.g., Delbecq and Van de Ven, 1971). Other procedures for group interaction or variations on the basic Delphi and NGT procedures do appear in the literature. Most studies, however, include one or both of these two techniques. In addition, in actual application, Delphi and, to a lesser degree, NGT markedly dominate. Therefore, we will not detail other procedures, but will discuss variations as appropriate.

Although found in the literature in a number of forms, the Delphi procedure is usually distinguished by three characteristics: (1) anonymity of group members; (2) iteration with controlled feedback; and (3) statistical group response (Dalkey, 1969b). Anonymity is a particularly important aspect of the Delphi procedure, as it is meant to avoid problems of dominant personalities, status incongruities, pressure for conformity, and so forth; all conditions which can reduce the quality of group assessments. More comprehensive reviews and discussion of Delphi can be found in Linstone and Turoff (1975), Pill (1971), and Sackman (1975), which also include extensive bibliographies.

NGT procedures get their name from the characteristics that the group does not interact in a normal manner, but only in a very limited or "nominal" sense. That is, the interaction allowed the group members is tightly controlled. As discussed by Delbecq, Van de Ven, and Gustafson

(1975), it is a four step procedure: (1) silent judgments by individuals in the presence of the group; (2) presentation to the group of all individual judgments without discussion; (3) group discussion of each judgment for clarification and evaluation; and (4) individual reconsideration of judgments and mathematical combination.

Delphi has received probably the widest applied acceptance of any of the behavioral techniques. The procedure is easy to use, it is comfortable for the users, it gives an answer to the problem, and probably most important, it does not require that participants be assembled in the same geographic location. Anyone who has tried to put together a meeting of high-level experts in any field understands the difficulty and cost saved by removing this requirement.

The problem with Delphi is the lack of evidence that it gives good answers. Only two studies (Dalkey, 1969a, 1969b) support Delphi as superior to even simple face-to-face discussion groups. And, this evidence is based on nonexperts, very few questions (20, of which Delphi's answers were better for 13), and only two groups of subjects. On the other hand, there is considerable evidence that Delphi results in answers that are no better or worse than other procedures (Brockhoff, 1975; Van de Ven and Delbecq, 1974).

Although neither Delphi nor NGT was developed as a probability estimation technique, they have been used for such purposes. Four recent studies have been performed on this topic. The first such study (Gustafson, Shukla, Delbecq, and Walster, 1973) compared four procedures for determining a group judgment of the likelihood ratio of two hypotheses. In addition to Delphi and NGT, a statisticized group and an interaction group were compared. In the interaction group, members freely discussed ideas, but made judgments individually which were then aggregated mathematically. Using the geometric mean of the deviation (GMD) of the group likelihood ratios from the true likelihood ratios as the measure of goodness, the NGT produced the best estimates (GMD = 78) and the Delphi groups produced the worst (GMD = 128). The statisticized (GMD = 114) and the interacting groups (GMD = 111) were about the same.

A second study by Gough (1975) claimed to confirm the Gustafson _et al_. findings. Again, the nominal groups attained the best performance, followed by the interaction groups, Delphi, and a group asked to reconsider individual estimates with no formal exchange of information. But, a subsequent analysis of the Gough data indicated the differences were not statistically significant (Fischer, 1981). Fischer, using procedures similar to those of Gustafson _et al_. (1973), except for the substitution of consensus groups for the interacting groups, found no significant differences between procedures. Both Fischer and a subsequent study by Seaver (1978) suggest that the Gustafson _et al_. results were the consequence of the particular method used to evaluate judgments.

The study by Seaver (1978) compared Delphi, nominal groups, a modified nominal group procedure, a consensus procedure, and a no interaction group. In the modified nominal group procedure, subjects each presented their assessment verbally along with any reasons underlying the assessment or information that might be useful to other group members, but were not allowed free discussion. The consensus procedure gave no instructions except that the group must arrive at a group consensus of the probability.

In addition, a number of techniques for aggregating individual judgments to form statisticized group judgments were compared. Three weighting procedures were used to aggregate the individual assessments; equal weights, weights derived from self-ratings for each judgment, and weights derived from a model proposed by DeGroot (1974). In the DeGroot model, individuals are assumed to iteratively revise their own probabilities as weighted linear combinations of the revealed probabilities of other group members. A constant matrix of weights, $\underline{W}$, with elements w_{ij}, the weight assigned by the i person to person j; and a vector of initial individual probability distributions, P, with elements P_i,....P_m for the m individuals is also assumed.

DeGroot shows that, after n iterations,

$$P^{(n)} = \underline{W}P^{(n-1)} = \underline{W}^{n}P,$$

is the vector of probabilities, which will converge if there is a vector $W^* = (w^*_1, \ldots., w^*_m)$ such that:

$$\lim_{n\to\infty} w^{n}_{ij} = w^*_j$$

where w^{n}_{ij} as an element of $\underline{w}^{n}$. The elements of W^* can be found by solving the set of linear equations $w^*\underline{W} = w^*$ subject to the constraint:

$$\sum_{j=1}^{m} w^* = 1.$$

Aggregation models tested were the linear model, the geometric mean model and the likelihood ratio model. The linear and geometric mean models are, as the names imply, simply the arithmetic procedures of a linear weighted average and a geometric weighted average of the individual probability estimates. The likelihood ratio model updated a uniform prior probability using each of the individual likelihood ratios as pieces of information in the Bayesian formulation. According to this formulation, each of the individual likelihood ratios constitute a datum (Ω_i) in the equation:

$$\Omega_n = \Omega_o \cdot$$

Ω_o is the prior odds for the hypothesis in question and Ω_n is the posterior odds after all data have been aggregated.

The quadratic scoring rule, a proper scoring rule similar to the Brier score, was used to score the group estimates stemming from each of these procedures. The formula for this scoring procedure is:

$$S_k = 2P(\theta_k) - \sum_{i=1}^{n} P(\theta_j)^2$$

where S_k is the score if event θ_k occurs, $P(\theta_k)$ is the probability assigned to event k and $P(\theta_j)$ is the probability assigned to the other j events. No significant difference was found among the assessments produced by the different behavioral procedures.

2.4 Problems and Biases in the Assessment of Subjective Probability

An examination of the literature on probability assessment and psychological measurement raises the strong possibility that first cut judgments of probabilities often do not (with notable exceptions discussed earlier) correspond to objective values. There are those for whom this external criterion approach to validation of probability judgments is meaningless, i.e., the "subjectivists," but for the purposes of this review, we have assumed that the notion of a "true" probability, external to the judge, is meaningful. Thus, we are able to look at the impacts of error, both random and systematic, in the assessment of probability.

A number of biases have been described since the original finding of "conservatism" in probability revision (Phillips and Edwards, 1966). In fact, a unified theory of judgment for probabilities and inferential judgment of all kinds is being developed (see Slovic, Fischhoff, and Lichtenstein, 1977, for a more complete review). We will not attempt to discuss each finding in detail, but will merely name and briefly describe the biases relevant to the subjective judgment of human error probabilities.

von Winterfeldt (1980) presents the following conceptualization of errors affecting subjective judgment:

1. Unspecific random error in judgments

   a. Imprecise measurement, rounding errors - The tendency to use round numbers (possibly to avoid the appearance of unjustified over precision) is an error of this type. Assessors have been found to use a limited number of values (.05, .10, ..., .95, etc.) to represent all of their judgments. Thus, events that the assessor does not feel are of the same magnitude may be forced into the same category or value.

b. *Internal fluctuation of the response generating mechanism* - This type of error can occur even when the assessor has a strong sense of the likelihood of the event's occurrence. The error occurs when the method by which the assessor turns a cognitive sense of likelihood into a judgment fluctuates in response to factors that are irrelevant to the judgment. Examples of factors that could be expected to effect judgment stability are fatigue, concentration, or stress.

c. *Labile values, shifts of emphasis* - This error reflects changes over time in the values that the assessor expresses, which may result from shifts in the information the assessor is using to make his or her judgment or shifts in the relative weight given to alternative information sets.

2. Response biases

a. *Central tendency* - The tendency to shift response in both halves toward the middle of the response scale. For example, if probability estimates are to be made on a percentage scale, assessors often shift their responses toward the 50 percent point (Johnson, 1977). Notice that such central tendency can lead to incorrectly coherent judgments. That is, if both high (.5-1.0) and low (0-.5) are shifted toward the middle of the scale they may still sum to one, and yet both judgments deviate from the true judgment toward the center.

b. *Avoidance of extremes* - Similar to central tendency bias, but here the judge is pulled away from extreme values rather than drawn toward the middle (Lichtenstein *et al*., 1978). The effect is thought to be related to scale bounding (the assessor does not want to use too extreme a response and not have room for more extreme ones later). This suggests the use of unbounded scales (odds, likelihood ratios) rather than bounded ones (probabilities, percentages).

c. *Use of the whole scale* - Another finding is that probability assessors ignore extreme anchors and use the whole scale (Stillwell *et al*., 1977). For example, when a scale is anchored by a *very* likely outcome at one end and a very unlikely one at the other, and several alternatives are presented that are all near the event of low likelihood, assessors will tend to spread their responses over the entire range of the scale. However, there is some evidence that the relative spacing appropriately reflects relative probability ratios (Goodman, 1973; Stillwell *et al*., 1977).

d. Anchoring and insufficient adjustment - The assessor anchors his or her judgment on some readily available point on a scale (e.g., the 50 percent point in judging the likelihood of an event). Then adjustment is made in the appropriate direction. However, usually that adjustment is insufficient (Tversky and Kahneman, 1974).

3. Processing biases

a. Conservatism - Conservatism refers to the finding that assessors do not revise their probabilities upon the receipt of information as much as they should according to Bayes' Theorem, the normative rule (Phillips and Edwards, 1966). Consequently, judged probabilities assessed after the receipt of information are less extreme and do not cohere with posterior probabilities calculated from likelihood ratios and prior probabilities using Bayes' Theorem.

b. Representativeness - This error results from assessors considering a sample from a population as more likely if that sample "represents" or resembles the population. For example, when asked to estimate the relative likelihood of samples generated from one of two binomial distributions, assessors tend to generate estimates by using the similarity of proportion of successes (s/n) to that of the two possible parent populations rather than the formally correct difference between successes and failures (s-f).

c. Availability - This is the tendency to overestimate the likelihood of events that are "available," i.e., those for which relevant examples are easy to recall, and underestimate those for which recall is difficult. For example, when asked to judge the frequency of death from various causes, assessors have been found to overestimate events that receive much media attention (e.g., plane crash, homicide) and underestimate those that receive little attention (e.g., emphysema, asthma) (Lichtenstein et al., 1978).

d. Neglect of base rates - Assessors tend to focus on individuating information when making categorical judgments and ignore the base rates of the categories.

e. Overconfidence - Assessors are often found to express more certainty in probabilistic answers than their "track record" (in terms of hit rate) justifies. Judgments of one and zero are much more common than the extremity of their meaning (always and never) makes logical. This finding may be related to perception of control studies where the people will greatly overstate the likelihood of success on a task when they are given some element of control versus the case where the outcome is either completely out of control (random) or in the control of others. Such a subconscious bias might occur in numerous instances where the events being considered are human actions.

f. Nonregressiveness - This bias reflects the failure to be cognizant of imperfect relationships between predictor variables and the quantity to be judged (in this case probability). A statistical fact is that prediction should regress toward the mean of predicted values in light of an imperfect relationship. For instance, when predicting the IQ of a child from that of the parents, the relationship is not perfect and, therefore, the prediction should be closer to the mean of all IQs (i.e., 100) than the mean of the two parents' IQs. Assessors will often simply use the mean of the predictors or the equivalent mean of deviation of the outcome variable rather than the properly regressive value.

4. Context effects

a. Neglect of relevant (but minor) aspects - Incoherence of judgments may be produced by inappropriate influences of context on a given judgment. A well-known example is Tversky's (1969) intransitivity of preferences, in which subjects produced fundamental intransitive preferences for gambles because they focused on different minor aspects of the gambles at different times.

b. Relevance of irrelevant aspects - Assessors may use information irrelevant to the judgment to be made. An example is the effect of magnitude of response scale on magnitude of judgment found by Stillwell *et al*. (1977). In either case, neglect of relevant aspects or the use of irrelevant aspects of the problem, errors can be produced by an inappropriate information processing strategy generated by drawing attention from relevant, or to irrelevant, contexts.

c. Isolation effect - This effect has been found in the context of gambles where assessors ignore the impact of features that are common to each of the gambles being evaluated. Although there is no evidence for or against it, this may also be the case when assessors are comparing event likelihoods.

d. Value induced bias - Wallsten (1978) discusses this bias, commonly found in medical contexts, in which the likelihood of outcomes with extremely high negative value (e.g., brain tumor, cancer) are routinely overestimated. This can be viewed as a "conservative" approach to diagnosis, always putting extra weight onto occurrences with particularly bad consequences.

Many biases are explained as attempts by the assessor to make inferences in a complex, confusing environment with only limited information storage, retrieval, and processing abilities. The assessor tries to bring his or her information to bear on the estimation problem, and in the process loses some and misuses other relevant portions. Often, in this process, the assessor uses simplification strategies that sometimes work quite well, but at other times can lead to poor and inconsistent judgments.

Cognizance of these biases and their impact on the judgment of HEPs should dictate the development of any procedure to be used in elicitation. Several of the above-discussed problems are of particular relevance to HEP judgment, but each is likely to impact the elicitation in at least a small way. One of the goals of our effort is to develop procedures that eliminate or reduce these biases in estimates of HEPs.

2.5 Training Probability Assessors

Training of assessors for HEP judgments could be the single most effective tool for improving the quality of those elicited numbers. Training can take a number of forms, depending on the time available to the assessor as well as how stringent the requirements are for the assessed numbers. The literature on the training of probability assessors is limited, but it does suggest several approaches to training, some that have been validated, others that have not.

One type of training involves feedback of the calibration of judges' responses. An early study (Adams and Adams, 1958) using this type of feedback found some improvement after feedback. Subjects were asked to determine whether pairs of words presented together were synonyms, antonyms, or unrelated. They then gave an estimate of the confidence they had in their choice on a scale defined to them in a way very similar to how we have defined calibration: "Subjects were instructed to express their confidence in terms of the percentage of responses, made at that particular level of confidence, that they expect to be correct . . . Of those responses made with confidence level p, about p% should be correct," (pp. 432-433). Only the responses .33 (there were three alternatives), .4, .5, .6, .7, .8, .9, and 1.0 were allowed. Thirteen of the fourteen subjects who were shown calibration tallies and calibration curves after each of the first four sessions showed decreases in discrepancy scores for the fifth session, while the six control subjects who did not receive this feedback showed a poorer performance. The discrepancy scores were calculated by taking the mean absolute difference of the proportion of correct responses assigned to a category and the appropriate proportion, weighted by the square root of the number of judgments in that category. Thus, if 45 percent of the items that the subject assigned to the category .8 were correct and 20 were assigned, the score for that category would be $\sqrt{20}$ (.8-.45) = 1.57. The mean decrease in discrepancy scores for the 14 experimental subjects was 48 percent (13.20-6.28) while the control subjects increased 36 percent (11.16-15.22).

Adams and Adams (1961) report that in a nonsense syllable learning task with large overconfidence after one trial, improvement occurred after 16 trials with feedback of the type discussed above. They also briefly report a "transfer of training" experiment that over five days used different experimental tasks. They were, on the first day, judgments about the proportion of blue dots in an array of blue and red dots; on the second and fourth days, judgments about the truth or falsehood of general statements; on the third day, comparisons of physical weights; and, on the fifth day, judgments (synonym, antonym, or unrelated) about pairs of words. Eight experimental subjects, given calibration feedback during the first four days, showed on the fifth day a mean absolute discrepancy score significantly lower than that of eight control (no feedback) subjects, suggesting some transfer of training. Unfortunately, they do not report their data.

Hoffman and Peterson (1972), using the Brier score, found significant differences between experimental and control groups in their performance after training. They conducted two experiments. In the first, student subjects answered /5 two-alternative questions and gave confidence judgments in the form of probabilities in each of three sessions. The questions were almanac types questions of the type "The capitol of Oregon is A) Eugene or B) Portland." The experimental group received the scoring rule feedback after each question during each of the three sessions while the control group did not. Ten of the 12 experimental subjects had better average scores for the third session than for the first, while in the control group six had better first sessions and six better third sessions.

In the second experiment, 15 military intelligence analysts served as subjects. Few of the analysts had experience with making probabilistic estimates or were familiar with scoring rules. The scoring rule used in this experiment was a variation on the Brier score in which highe scores are better. The authors do not report the type of questions used in the second experiment. They found that 12 of the 15 analysts earned higher scores in the third session and three in the first. They do not report the magnitude of the improvement.

Schaefer and Borcherding (1973) examined the effect of training on the assessment of continuous probabilities. The questions involved judging the distributions of individual characteristics of students at the university at which the subjects were also students. Distributions were assessed in two ways. In the first, the subjects assessed cumulative probability distributions using a fractile procedure. The second procedure was called "the prior equivalent sample" procedure. In this procedure, the subject is asked to provide the parameters r and n of a beta distribution. The parameter r represents the number of successes in n samples from a data generating process. The value r/n corresponds to the mean of the distribution, and the size of the parameters r and n to the level of confidence, that is, the tightness of the underlying distribution.

Subjects were run in four experimental sessions, each one week apart. During each session, 18 probability distributions were assessed by each method. At the beginning of the second, third, and fourth sessions, subjects were shown a table with estimated fractiles for method one and the corresponding fractiles for method two for each subjective probability distribution assessed. Subjects were told that ideally the true value would always fall into the same category of the distribution. They were also told to pay particular attention to their assessed distributions for which the actual value fell in the extreme categories.

Results were expressed in terms of the proportion of true values that fell in the interquartile range and the extremes of the distribution (below the .01 and above the .99 fractiles) averaged across subjects. For the perfectly calibrated assessor, 50 percent of the actual values should fall in the interquartile range and two percent (one percent above and one percent below) should fall into the two extreme categories. Table 2 shows the percentages over the four weekly sessions. Even with this substantial improvement, all distributions are still too "tight" with too few values falling in the interquartile range and too many in the extreme ranges.

Studies by Winkler (1967) and Stael von Holstein (1971a, 1972) suggest a second type of training. They found that, among individuals familiar with the subject matter, judging the likelihoods of events that had not yet occurred or for which they could not have exact information, those who were more statistically knowledgeable in addition to their substantive knowledge produced as good or better estimates than those who were only substantively knowledgeable. Winkler (1967) used three groups of subjects. A "no-stat" group (N = 15) had no statistics training beyond an introductory course. The "stat" group (N = 20) had significant statistical training including several business statistics courses and specific training in probability calculus and Bayesian statistics. The "math-stat" group included two Ph.D. candidates in statistics and a professor of statistics. Each of the subjects answered questions about the demographic characteristics of the student population at the university in which they were enrolled or were teaching. The study examined the medians of the maximum vertical discrepancies between cumulative probability distributions assessed by different methods (thus testing reliability and coherence), and found that the "math-stat" group had by far the lowest median discrepancies (.089) compared to the "stat" group (.215) and a slightly higher discrepancy for the "no-stat" group (.251).

The study by Stael von Holstein (1972) used five groups of subjects in a stock market prediction task. The groups were: (1) ten persons actively working in the stocks and bonds department of a Stockholm bank, (2) ten persons actively connected with the Swedish Stock Exchange, (3) 11 people associated with the Institute of Mathematical Statistics at the University of Stockholm (UOS), (4) 13 business administration professors at UOS, and (5) 28 business administration students at UOS. They were asked to state their probabilities for each of

TABLE 2

Percentages of Actual Values in the Interquartile Range (Correct Value = 50%)

and Extreme Ranges (Correct Value = 2%)

Session	1	2	3	4
Method 1, Interquartile Range	22.5	30.3	36.9	37.9
Method 2, Interquartile Range	15.7	38.1	37.4	48.2
Method 1, Extreme Range	38.8	14.9	16.0	11.9
Method 2, Extreme Range	49.8	21.5	15.5	5.6

five categories into which the prices of 12 stocks on the European Stock Exchange would fall two weeks after their predictions. The categories were:

1. The buying price decreases more than 3%.

2. The buying price decreases 1-3%.

3. The buying price changes at most 1%.

4. The buying price increases 1-3%.

5. The buying price increases more than 3%.

Responses were scored using the quadratic scoring rule which is a linear transformation of the Brier scoring rule and thus is also a proper scoring rule. The transformation was such that higher scores were better.

The quality of all responses was relatively poor, probably reflecting task difficulty rather than poor judgment. A subject using the strategy of simply predicting the category rates of these stocks for the previous year would score 7.2, while the average subject scored 6.69. There were small differences between the groups, with scores of 6.39, 6.75, 6.80, 6.64, and 6.74 for the bankers, stock brokers, statisticians, business administration professors, and business administration students, respectively. The important aspect seems to be that those with statistical training, i.e., the statisticians, professors, and students, seemed to do just as well as the stock brokers and somewhat better than the bankers, the latter two being considered the substantive experts.

One possible implication of this work is that experts in a substantive area will benefit from statistical information (a short explanation of the simpler rules of probability could make a large difference) when making probabilistic judgment in their own area of substantive expertise. A more extensive form of statistical training would include examples and explanation of known biases in human judgment and consistency checks that the expert could perform to ensure against them.

Another important aspect to the studies discussed above is that, in experiments where the analyst worked one-on-one with the assessor and pointed out inconsistencies, the assessor was able to generalize and improve performance on other estimation tasks (Winkler, 1971). This suggests a third type of training, training by example and counterexample. This procedure would be somewhat akin to feedback training, but would be built around specific examples where biased judgment has been shown to have a particularly strong impact. Many of the examples of Tversky and Kahneman (1974) and Kahneman and Tversky (1979) are of this type. They provide the judge with striking evidence of the biases to which judgment can be subject.

2.6 New Methods for Resolving Inconsistent Judgments

In many instances for assessing probabilities (or other types of judgmental data), an overspecified set of judgments will be available. For example, if ratio estimates of relative likelihood were made for all pairs of n events, there would be n(n-1)/2 judgments where only n-1 judgments are needed to minimally specify the rank order and interval spacing of scale values. These multiple judgments (up to n(n-1)/2) cannot be expected to be entirely consistent--some inconsistencies enter into all judgmental processes--but they are, nevertheless, desirable because of the additional information they provide. The problem then becomes one of determining their degree of consistency (high levels of inconsistency would suggest the entire judgmental process is suspect) and resolving the inconsistencies in an appropriate manner.

Saaty (1977, 1980) has developed the Analytical Hierarchy Process (AHP) to handle this problem for ratio judgments. The AHP uses the maximum eigenvalue, λ_{max}, of the matrix of ratio judgments as the scale values resolving inconsistencies. The amount of inconsistency is indexed by $(\lambda_{max}-n)/(n-1)$, and Saaty (1980, p. 51) provides a rule of thumb that if this index is less than .1, the judgments are of acceptable consistency.

Freeling (1981) suggests an alternative to the AHP which uses least squares methods to resolve inconsistencies and determine scale values, allows for differential weighting of estimates in the matrix of ratio judgments, and does not require all n(n-1)/2 ratio judgments. This latter point is important for practical purposes. We suspect that some additional judgments beyond the n-1 minimal set may be quite useful, but that certainly all n(n-1)/2 judgments are not needed.

Freeling's method can also be extended to the multiple expert situation where several judges each provide a (possibly inconsistent) set of ratio judgments. The method for reconciliation of inconsistencies across judges is again a least squares procedure. Freeling has also provided for a statistical test of the inconsistency of judgments in the matrix.

We view this procedure as very promising for situations where ratio estimates of HEPs are made. Its flexibility with respect to the number of experts and the number of judgments from each expert is desirable. The output of Freeling's method would give the ratios between the probabilities of different events; for example P(A) = 2P(B). However, to convert these values into probabilities scaled from zero to one, some further information is required. An easy way in which this might be achieved is to include one event of known probability in the assessment. For example, if we know P(A) = 0.5, then we could deduce P(B) = 0.25.

2.7 Defining and Structuring Judgments

A crucial component of the judgment of human error probability is that the expert understand the exact nature of the judgment to be made. For this reason, it is important to determine the factors that contribute to the likelihood of errors in nuclear power plant operations. These "Performance Shaping Factors" (PSFs) (Swain and Guttmann, 1980) are special circumstances, unique to the plant, operator, or specific situation that make error more or less likely than the average of the same general situation characteristics across the industry. Thus, PSFs help to more exactly identify the situation to be judged.

There are two aspects to the PSF problem. First, the more important relevant factors in any HEP situation must be identified, and second, the relationship of the PSFs to probability of human error must be determined. Mathematical psychology and psychological scaling can contribute to each of these aspects.

In some sense, the first part of the PSF problem, that of identification of the relevant PSFs for any given HEP judgment, is the problem of complete description. The more completely spelled out the PSFs are in a given situation, the better the expert will know just what event is being judged. The expert is therefore less likely to introduce personal interpretations not actually found in the situation.

Psychological scaling can make a number of contributions to the identification problem. Often PSFs are not clearly identified for a given judgment. One means of identification is through the use of expert judgment in a Multidimensional Scaling (MDS) paradigm (Kruskal and Wish, 1978) or using related techniques derived from factor analysis (Rummel, 1969). Experts are asked to provide judgments of similarity, distance in some n-dimensional space, or even simple dominance (greater or less than) for pairs of stimuli. Of course, the resolution of the solution improves as the strength of the input information increases. MDS analysis is then used to explore the judgment matrix to determine its dimensionality. In much the same way, input judgments of simple probabilities could be used in a factor analytic paradigm to examine general factors underlying the probabilistic judgments. The expert, in conjunction with the analyst, is left with the task of giving meaning to the dimensions or factors thus discovered. These dimensions or factors are the PSFs. The major difference between the two analytic approaches is the nature of the input judgment.

A second approach to the identification of PSFs is to use the social facilitation (Zajonc, 1965) aspect of group interaction. Many of the behavioral interaction techniques discussed under groups could be used as qualitative elicitation procedures for PSFs. For instance, with mild alteration to fit the qualitative nature of the output sought, the limited interaction approach associated with the nominal groups procedure lends itself well to the investigation of the PSF problem. An additional benefit accrues if the same group is used to judge the HEP for which the PSFs have been defined. That is, the problem of misunderstanding of the specific HEP to be judged is solved in that the group has defined and agreed to its exact characteristics in previous interaction.

The second part of the PSF problem, determining the relationships between the cues (PSFs) and the criterion (HEP), can benefit less from ideas of psychological scaling than from those of another area of mathematical psychology, namely behavioral decision theory. With the judgment of HEPs, we are faced with the problem of unraveling a complex pattern of interconnections relating the HEP to the PSFs as well as the PSFs to each other. It has been suggested (Swain and Guttmann, 1980) that the relationships may be multiplicative between PSFs, nonlinear between PSFs and the HEP, or relationships of a number of other kinds may exist. Certainly, we can assume that the PSFs are not independent in their contribution to the magnitude of the HEP.

How then do we determine the optimal relationship? We do not have data directly relating the parts to one another, and thus we would need to use expert judgment to arrive at the relationships. A method known as paramorphic representation of judgment (Hoffman, 1960) has been discussed as a solution to the problem. Once we have the PSFs that define the HEP situation that we want to examine, we ask the expert to provide judgments of the HEP for a large number (10+ per PSF) of combinations of those factors. Multiple regression, or possibly some biased estimation technique such as ridge regression is then used to derive a model of the judge from his or her own judgments. Thus, in an indirect manner we are able to derive the role of the individual PSFs in driving the judgments of HEP as well as to determine the interrelationship between PSFs. We are not constrained by the usual linearity and independence assumptions of linear regression if we use the experts themselves to advise us as to the models (nonmonotonic or higher order terms for instance) to be tested. These models will provide remarkably good estimates and can even replace the judges from whose judgments they were derived (e.g., Camerer, 1981; Dawes and Corrigan, 1974).

2.8 Summary

On the basis of literature reviewed in this section, the following statements regarding expert assessment of HEPs can be made:

1. Experts are generally able to provide accurate likelihood assessments both in terms of agreement with objective likelihoods (usually relative frequencies) when they exist and calibration when they do not. Assessments are improved if (a) feedback (in terms of consistency of responses, proper scoring rule, example and counterexample, or simply correct answers, when they exist) is provided, (b) the experts are well-motivated, and (c) the experts are experienced with quantitative thinking and judgments.

2. When directly estimating discrete likelihoods, estimates are more accurate and reliable when (a) made in writing rather than verbally, (b) made in odds (ratios) rather than probabilities, and (c) made on logarithmic scales.

3. Judgments are more accurate when multiple elicitation approaches are used and inconsistencies are resolved.

4. The likelihood to be elicited should be carefully and completely defined and structured, including definition of the population of which the event is a member and causal linkages with other characteristics and/or events.

5. Multiple experts with varying sources and types of information should be used, assuming, of course, that they are sufficiently knowledgeable about the question under consideration.

6. Arithmetic combinations of multiple estimates produce estimates that are as good as those produced by interacting groups, but the process is less acceptable to the experts.

7. Several biases are known for the judgments required, and should be pointed out to experts making judgments.

8. Training, particularly in probabilistic thinking, can effectively improve the experts' likelihood estimates.

9. Techniques for reconciling inconsistent ratio judgments appear promising for improving HEP estimates.

3.0 PSYCHOLOGICAL SCALING

Scaling is the process of assigning numbers to objects, events, or properties of either, in such a fashion that the numbers represent relationships among scaled entities (Coombs, Dawes, and Tversky, 1970). The purpose of scaling is to allow numbers to substitute for the objects or events in question. We may, therefore, derive further relationships by performing mathematical operations on those numbers.

Substituting numbers for psychological objects, such as preferences or attitudes, is inherently difficult. Such objects are not directly observable and thus psychologists have studied various scaling methods to enable quantitative manipulation of variables and various conditions under which such number assignment is possible. This latter enterprise is more correctly dubbed "measurement theory" (Coombs, Dawes, and Tversky, 1970; Roberts, 1979; Suppes and Zinnes, 1963).

Data rarely conform exactly to the measurement theorist's axioms. So, in addition to being able to generate numbers to represent objects, a scaling procedure often must include a means for removing judgmental "error" from the data and a means of estimating "true" scale values for the objects.

It should be noted that scaling is not just the systematic assignment of numbers to objects; the assignment must be meaningful. Cliff (1973) gives four conditions under which scaling is meaningful: (1) the scale values are indicative of consistent and numerous relationships among the data; (2) one has an underlying measurement theory that assumes knowledge of which scale value transformations still preserve the data relationships; (3) one has a detailed algorithm for transforming raw data into scale values; and (4) one can demonstrate that the obtained scales have external validity. This requires an additional data set from that which produced the scale values.

Conditions (1) and (2) above represent no problem for probability estimation and scaling. A comprehensive set of rules dictates the scale value transformations that still preserve the data relations. Conditions (3) and (4), on the other hand, raise questions for a probability scaling procedure.

Condition (3) is not a problem in the usual sense, i.e., no detailed solution procedure exists. Rather, a larger number of procedures exist that have been shown to provide different answers to the same questions. And in situations where an external criterion has been used to test the judgments, no one procedure has been shown to routinely provide the best numbers.

Still, given the wealth of assessment procedures and the quality of judgment in some judgmental situations (for example, estimates obtained from weather forecasters), we are confident that condition (3) is solvable by the appropriate choice of procedure. Solution of condition (3) dictates solution of condition (4). Thus, given that we can choose the appropriate procedure for assessment, we have fulfilled conditions for development of a scale of judgmental probabilities.

3.1 Techniques of Scaling

With the above discussion as motivation for our search of the scaling literature, we now turn to the task of detailing scaling procedures and finally determining those useful as probability assessment techniques. Psychologists have devised several scaling methods to ensure proper number assignment to objects. With respect to scaling subjective probability, two points should be noted. First, most scaling techniques produce ordinal, interval, or ratio scales, while the probability scale is absolute. Thus, general scaling procedures must be modified to produce an absolute scale. And, second, the requirements of a scaling technique must be practically feasible, i.e., they cannot ask judges for more or different information than they can provide. Failure to fulfill either of these considerations invalidates a scaling technique for use as a probability scaling procedure.

Torgerson (1958) lists several ways in which standard psychological scaling techniques differ. For our purposes, we shall rely on three basic aspects of scaling:

(1) differences in theoretical approach;

(2) differences in assessment procedures; and

(3) differences in analytical procedures.

Since the nature of this review is to highlight pragmatic concerns, points (2) and (3) will receive considerable attention, whereas information relevant to point (1) will be noted only as needed.

3.1.1 Scaling by paired comparisons. Paired comparison scaling was introduced by Thurstone (1927) as a means of scaling psychological attributes which have no physical basis. Every possible pair of *n* stimuli is presented to *n* subjects. The task of each subject is to indicate which member of the pair *dominates* the other with respect to the attribute to be scaled. If the attribute is subjective probability, for each pair we would ask, "which event is more likely?" No equality judgments are allowed, and a stimulus, generally, is not compared with itself. The relevant datum for each stimulus pair is the proportion of time that, say, stimulus *j* dominates stimulus *k*. This relationship is assumed to be symmetric so the proportion of time that *k* dominates *j* is one minus the proportion of time that *j* dominates *k*. Scale values for each stimulus are derived either according to Thurstone's law of comparative judgment (1927) or its major competitor, the Bradley-Terry-Luce model of choice (Bradley and Terry, 1952; Luce, 1959).

Torgerson (1958) provides a detailed description of these models. The law of comparative judgment assumes that each stimulus is represented by a distribution of subjective magnitude. In comparing two stimuli, a magnitude is selected randomly from each distribution, which corresponds to selecting a value randomly from the distribution of differences. The proportion of times stimulus j is judged to be greater than stimulus k is taken as an estimate of the area of the distribution of differences where the difference is positive.

The law of comparative judgment assumes the distribution of differences is a normal distribution, thus scale values are derived by taking the normal deviates associated with the proportions. As noted by Coombs (1964), this is a normal curve transformation of proportions into scale values. The Bradley-Terry-Luce model, although derived under different assumptions, is similar except that it utilizes a logistic transformation rather than a normal curve transformation.

The selection of scale-producing models reflects one's theoretical biases more than method superiority. In addition, scale values derived from both models are closely, and systematically, related (Torgerson, 1958; Yellot, 1977). Methods for controlling bias in the paired comparisons situation have been developed by Ross (1934), Torgerson (1958), and Wherry (1938).

The greatest single criticism that can be levied against the method of paired comparisons is that for large n, the number of judgments from each subject is prohibitively large. For example, with 20 stimuli, 190 pairs exist; for 50 stimuli, 1225 pairs; for 100 stimuli, 4950 pairs. Fortunately, methods for reducing the number of judgments required are well documented (Bock and Jones, 1968; David, 1963; Torgerson, 1958). These methods allow valid scale values to still be generated with little loss of information.

Torgerson (1958) suggests some relatively simple practical ways to reduce the required number of judgments. One method is to select a set of standard stimuli from the total stimuli set and compare other stimuli only against the standards. A second method, which requires a rough a priori ranking of stimuli, requires that each stimulus only be compared to others that are "close" to it in terms of the characteristic being scaled. Another method divides the stimulus set into overlapping subsets, each of which is scaled separately with the overlapping stimuli used to connect the scales. A final suggestion is to use nonoverlapping subsets of stimuli, with a set of standard stimuli, which are used to relate the subsets to one another.

Bock and Jones (1968) and David (1963) present more detailed and statistically justifiable designs based on balanced and partially balanced incomplete block designs. They also discuss similar designs for use with multiple judges where all judges do not judge the same pairs of stimuli. Such designs are important in situations where some comparisons among judges are wanted.

3.1.2 *Scaling by ranking*. Ranking techniques are formally equivalent to paired comparison techniques--all paired comparisons can be derived from a rank order of events, and vice versa (Bock and Jones, 1968; Torgerson, 1958). However, a fundamentally different judgmental process may underlie ranking. The general method of rank order (Guilford, 1954) instructs the subject to rank the stimuli in order with respect to the attribute to be scaled. If $m>1$ subjects each rank order $n>1$ objects, then there are two basic procedures that can be followed. One approach, as mentioned above, is to deduce paired comparison proportions from the rankings and proceed to treat the rankings as if they were paired comparisons.

A second basic method for handling data generated by the method of rank order deals with the problem of intransitivity in the rankings. If we anticipate intransitive rankings, we can develop rules for the consequences of intransitivity in such a way that allows an ordering of the objects. This approach is dubbed "consensus ranking" (Kemeny and Snell, 1962) or generating a "social welfare function" (Arrow, 1951; Luce and Raiffa, 1957). The general idea is to derive a group ranking which best (in some sense) summarizes the set, or profile, of individual rankings. (Note that this scaling approach is highly analogous to the general flavor of group probability assessment.) Unfortunately, it has been shown that under certain reasonable conditions, such a ranking need not exist (Arrow, 1951; Coombs, 1964). Thus, the theoretical base of such a procedure may be suspect, but this does not eliminate it as a practical scaling procedure.

3.1.3 *Scaling by sorting*. Sorting methods have been developed for situations in which the number of stimuli is large, and there is no clear choice between paired comparisons and ranking (Edwards, 1957). As such, standard sorting procedures contain elements of both ranking and paired comparisons. Basically, a sorting task consists of having subjects place either objects *or* attitudes toward objects in various ordered categories. The endpoints of these piles are labeled according to whatever attribute is being evaluated. For example, if the attribute is subjective probability, the most extreme categories might be labeled "most likely" and "most unlikely." Other piles are merely given a letter name. The subject's task, then, is to place each event in the category corresponding to its perceived likelihood.

A fundamental theoretical distinction defines the two basic sorting methods. If one assumes that the intervals between successive categories represent *equal-appearing intervals*, then one is using the method by the same name. If, on the other hand, inequalities exist in the widths of the intervals between the psychologically-scaled categories, the appropriate technique for analysis is the *method of successive intervals* (Edwards and Thurstone, 1952). Unfortunately, tests to determine whether intervals are equal do not exist. Torgerson (1958) discusses analytical models for handling sorting data.

3.1.4 Scaling by rating. With rating procedures, the stimuli are presented one at a time. The subject's task is to rate each stimulus with respect to the attribute in question. The rating may be numerical, adjectival, or graphical. In general, a rating form is a set of categories, by which a subject is required to partition a set of stimuli into mutually exclusive classes (Bock and Jones, 1968). These classes are said to represent only ordinality and contain no reference to intervals, or distance between classes. To describe the ordinal nature of rating, the label "method of successive categories" has been given to the general rating situation (Bock and Jones, 1968). If one assumes that ratings are normally distributed, then one is using Thurstone's law of categorical judgment (Torgerson, 1958). The law of categorical judgment is similar to the law of comparative judgment with the additional assumption that category boundaries are treated like additional stimuli. Rating is a relatively efficient scaling technique requiring only n judgments for n stimuli and means for estimating rather than assuming discriminal dispersions.

Rating differs from sorting primarily in the task the subject is required to perform. When sorting, subjects are given all stimuli at once, and, thus, are aware of the complete range of stimuli. When rating, only one stimulus is presented at a time, so the subject does not know what stimuli remain to be judged. Sorting clearly becomes difficult as the number of stimuli becomes large because of the complexity in trying to consider many stimuli simultaneously.

Experimental work has found that scale values are insensitive to variations in category placement, what subjects are used, or number of subjects (Bock and Jones, 1968). These and other findings indicating the predictive validity of rating scales attest to the general usefulness of the successive category approach. A practical advantage over paired comparision is that for n stimuli, only n judgments are required. Also, if one chooses to use Thurstonian methods of analysis, then one can estimate the discriminal dispersion of each stimulus rather than assume all are equal.

3.1.5 Scaling by fractionation. Fractionation methods are characterized by having subjects directly report the ratio between two subjective magnitudes. The most common method, magnitude estimation (Stevens, 1956), requires that numbers be directly assigned to objects in accordance with the subjective impressions they elicit (Jones, 1974). Experimental instructions require ratio judgments, hence the name "fractionation."

Typically, one object in the stimulus set is designated the standard and assigned an arbitrary value; this then becomes the stimulus against which all other stimuli are judged. Magnitude estimation is a quick assessment procedure which easily yields scale values.

3.2 Comparison of Scaling Techniques with Emphasis on Validity and Reliability

Very little work has been done directly comparing the various scaling techniques. However, a few studies do exist that either (1) compare the techniques on reliability (consistency) and/or validity (accuracy) or (2) examine the reliability and/or validity of single methods.

Direct comparison of scaling procedures is hindered by the fact that very few methods require the same experimental manipulations in order to generate scale values. Sometimes the stimuli differ (sensory versus attitudes), the judgments differ (dominance versus proximity), or the responses differ (pencil mark versus placing objects into piles). Still, some comparisons can and have been made.

Arora (1977), looking at the sociometric status of 13-17 year old females, found somewhat greater reliablity for scale values generated by partial ranking than for those generated by paired comparisons, although differences were not statistically significant. This was true even though there was high commonality between the two scales.

Schriesheim and Schriesheim (1978) compared scale values for expressions of frequency (e.g., "always," "sometimes") derived from magnitude estimation and from use of the law of comparative judgment on rank order judgments. Comparing these results with the results of similar earlier studies by Bass, Cascio, and O'Connor (1974) and Schriesheim and Schriesheim (1974), they found magnitude estimation to more closely achieve interval scale measurement than did ranking. The "average absolute percent scale value difference" (100 times the absolute difference between the scale values in the early study and in the later study divided by the largest scale value) was 2.59 for magnitude estimation and 12.51 for ranking. Both methods achieved a high ordinal test-retest reliability, with a Spearman rho correlation of .991 for magnitude estimation and 1.0 for ranking.

Magnitude estimation also appears to be less subject to fatigue effects than paired comparisons, especially when a large number of stimuli need to be judged. Lodge, Tanenhaus, Cross, Tursky, Foley, and Foley (1976) found that scale values derived from magnitude estimates of political opinion had high convergent validity with scale values derived from physical responses (hand grip strength and sound pressure). Correlations between scale values were greater than .95.

Apparently, magnitude estimates are reliable, relative to paired comparisons, because no recourse to theory (e.g., distributions of discriminal dispersions) is required to generate scale values. However, empirical studies of ratio judgments (Eyman and Kie, 1970; Ross and Di Lollo, 1971; Sjoberg, 1971) show context effects; responses and thus scale values depend upon the set of possible stimuli with which specific stimuli are contrasted.

Finally, it should be noted that, *in general*, no simple relationships exist among scale values generated by different methods. However, within any particular study, simple relations are usually found: often discrimination scales (e.g., from paired comparisons) are logarithmically related to magnitude scales (from ratio judgments) and scales from ratings or sorting have a quasi-log relation with these other scale types (Cliff, 1973).

3.3 Synopsis and Conclusion

This review has largely ignored theoretical distinctions among the different scaling procedures. One should always consider, however, the ease with which any scaling procedure yields consistent values, even in practical scaling situations. Wallsten and Budescu (1980) give a detailed review of how to obtain consistent subjective probability assessments.

A more practical consideration is the number and difficulty of judgments required to produce scale values. Some scaling procedures have elegant, formal methods for reducing the number of judgments, others do not. This very practical consideration is related to the previous discussion concerning reliable scale values. Typically, trials involving the same stimuli are necessary in order to promote consistent scale values. However, group scaling procedures (such as Thurstone's law of comparative judgment) circumvent this problem.

4.0 CONCLUSIONS AND RECOMMENDATIONS

A number of factors and procedures have been demonstrated to impact the quality of probabilistic judgments. In the best of all worlds, the procedure that produces the best numbers is also the procedure whose cost is the lowest in terms of time and effort invested in the elicitation. If, on the other hand, as is often the case, the factors of cost and quality are in conflict, the decision must be made based on the trade off between these factors for the specific situation. We do not expect, therefore, to propose a universally best procedure, but instead will describe factors that have been shown to contribute to higher quality judgments and discuss the practical considerations relevant to each. We will also discuss procedures for assessing probabilities suggested by scaling methods not previously used for elicitation and mapping of event probabilities, and try to judge their practical limitations.

First, and perhaps foremost, the event to be judged must be completely defined and structured. Individual latitude in definition of the event to be judged will almost necessarily add error variance and, in some cases, systematic bias to probabilistic judgment. The delineation of performance shaping factors suggested by Swain and Guttmann (1980) goes a long way in the pursuit of this goal and we have suggested aids to this process. In addition, we have suggested interactive elicitation techniques where iterative elicitation of various types concerning calibration or comparison of responses with known values, and consistency checks should provide further safeguards against event misspecification. Each of these suggestions does, however, involve an increased investment cost. Event description and scenario development is a time consuming process, as is the iterative elicitation. The procedure described in Spetzler and Stael von Holstein (1975) involves an extensive event specification phase, but the authors themselves admit that even for a very limited number of event judgments the process is often quite time consuming. This is obviously an extremely limiting factor when even moderate numbers of judgments are required. The use of judges who are familiar with nuclear power plant operation will help to minimize the time factor.

Our second recommendation is training. Indirectly, the quality performance of weather forecasters and the direct comparisons of experts with nonexperts and laboratory trained with untrained subjects show improved performance with training. We have examined a number of types of training and found each to contribute to judgment quality; of course, some types are more effective than others. An important factor with regard to training for which there are little data is the summative effect of different types. Some educated guesses can be made in this area.

Three types of training were identified: general statistics and basic probability concepts, training on the heuristics individuals use to make probabilistic judgment and the biases to which they lead, and feedback about quality of the individual's own responses. The first two of these are general and in terms of practicality can be expanded or shortened to fit the time available in the particular situation. Useful forms could range from a short tutorial to a lengthy session including sample problems, feedback, and extensive interaction between analyst and expert. The third is more problematic and will often be impossible where objective values simply do not exist for events analogous to the ones for which subjective inputs are needed. Such is likely to be the case for HEP estimation.

A third recommendation is the use of multiple experts. Several of the group procedures have the practical feature that the groups do not have to actually meet in order to provide "group" estimates. This is true for most of the "statisticized" procedures and the Delphi "interactive" procedure. A second consideration is that groups of experts with mixed expertise have been shown to produce less redundant and more independent information, and thereby better estimates, of unknown quantities. Also, there is little, if any, increase in cost using mixed expertise. Finally, probably the most important finding is that groups in general do provide better estimates of unknown quantities than do individuals, although the use of multiple individuals may sometimes be impractical.

One other aspect of the use of multiple experts warrants attention. If multiple groups are used and structured such that within groups knowledge and experience are relatively homogeneous, while between groups they are heterogeneous, estimation would be improved in two ways. From the within-group homogeneity, we could expect members to compare experiences, to facilitate each other to think more deeply about the judgment, to consider more of the relevant aspects of the question, and thus to arrive at group estimates that are more representative of the total amount of information at that group's disposal. The group estimate would still, of course, benefit from the reduced error variance in the estimate discussed earlier.

The second source of improvement would be the heterogeneity between groups. The within group homogeneity would give us stable estimates of likelihoods from specific areas of expertise. Thus, we could expect differences between groups to represent differences due to the different perspectives and information from the different types of experts. Post estimation interviews could then be utilized to determine reasons for these differences. Different points of emphasis and problems associated with information congruence, or lack thereof, among groups could be highlighted.

It is clear from the research and applied literature that multiple elicitation using markedly different procedures (usually one direct and one indirect), feedback comparing the results, and revision by the judge to correct inconsistencies is an effective procedure for improving probability estimates. Both of the more well-known and widely used procedures from applied contexts (Selvidge's procedure and that discussed in Spetzler and Stael von Holstein) have arrived at forms of this process through much thought and applied experience. However, as with several of the procedures leading to higher quality judgments, it should be noted that this can be a time consuming process.

Few new techniques for estimating probabilities judgmentally based on scaling methods were suggested by a review of the scaling literature. Most scaling techniques have been used in some form for probability estimation. Paired comparison techniques, however, appear promising and have received little use for probability estimation. These techniques should be investigated further to enhance their operability for HEP estimation.

Techniques based on ratio estimation (fractionation)--odds and likelihood ratio estimates in this context--have been shown to be more effective than direct probability estimation, although there appear to be some systematic biases in such judgments, particularly for very unlikely events. New developments such as Freeling's (1981) reconciliation procedure, may improve the use of such ratio judgments. These, and additional techniques, can be further developed into complete procedures including, where necessary, methods for transforming psychological scales into probability scales. One problem to be addressed, for example, is the number of events with known objective probability that would be needed to transform a psychological scale. If a certain relationship is assumed (e.g., linear or logarithmic) two known probabilities would be enough, but undoubtedly more would be useful to increase reliability. Given the scarcity of objective data and the difficulties involved in collecting them, such an increase in reliability may come at a large cost.

REFERENCES

Adams, J. K., and Adams, P. A. Realism of confidence judgments. *Psychological Review*, 1961, *68*, 33-45.

Adams, P. A., and Adams, J. K. Training in confidence judgments. *American Journal of Psychology*, 1958, *71*, 747-751.

Alpert, M. A., and Raiffa, H. *A progressive report on the training of probability assessors*. Harvard University, unpublished manuscript, 1969.

Arora, S. A comparative study of two methods of measuring pupils' status in the classroom. *Journal of Psychological Researches*, 1977, *21*, 143-146.

Arrow, K. *Social choice and individual values*. New York: Wiley, 1951.

Barclay, S., and Peterson, C. R. *Two methods for assessing probability distributions* (Technical Report 73-1). McLean, VA: Decisions and Designs, Inc. 1973.

Bartos, J. A. *The assessment of probability distributions for future security prices*. Indiana University, Ph.D. dissertation, 1969.

Bass, B. M., Cascio, W. R., and O'Connor, E. J. Magnitude estimation of expressions of frequency and amount. *Journal of Applied Psychology*, 1974, *59*, 313-320.

Beach, L. R. A note on the intrasubject similarity of subjective probabilities obtained by estimates and by bets. *Organizational Behavior and Human Performance*, 1974, *11*, 250-252.

Blanchard, R. E., Mitchell, M. B., and Smith, R. L. *Likelihood of accomplishment scale for a sample of man-machine activities*. Santa Monica, CA: Dunlap & Assoc., 1966.

Bock, R. D., and Jones, L. V. *The measurement and prediction of judgment and choice*. San Francisco: Holden-Day, 1968.

Bradley, R. A., and Terry, M. E. The rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika*, 1952, *39*, 324-345.

Brockoff, K. The performance of forecasting groups in computer dialogue and face-to-face discussion. In Linstone, H., and Turoff, M. (Eds.), *The Delphi Method: Techniques and applications*. Reading, MA: Addision- Wesley, 1975.

Brown, R. V., Kahr, P. S., and Peterson, C. Decision analysis for the ma..ager. New York: Holt, Rinehart, and Winston, 1974.

Camerer, C. General conditions for the success of bootstrapping models. Organizational Behavior and Human Performance, 1981, 27, 411-422.

Cliff, N. Scaling. Annual Review of Psychology, 1973, 473-505.

Coombs, C. H. A theory of data. New York: Wiley, 1964.

Coombs, C. H. Dawes, R. M., Tversky, A. Mathematica! psychology, an elementary introduction. Englewood Cliffs, NJ: Prentice-Hall, 1970.

Dalkey, N. Analyses from a group opinion study. Futures, 1969, 1, 541-551. (a)

Dalkey, N. An experimental study of group opinion: The Delphi method. Futures, 1969, 1, 408-426. (b)

Dalkey, N. Toward a theory of group estimation. In Linstone, H., and Turoff, M. (Eds.). The Delphi Method: Techniques and applications. Reading, MA: Addison-Wesley, 1975.

David, H. A. The method of paired comparisons. New York: Hafner, 1963.

Dawes, M., and Corrigan, B. Linear models in decision making. Psychological Bulletin, 1974, 81, 95-106.

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

Delbecq, A., and Van de Ven, A. A group process model for problem identification and program planning. Journal of Applied Behavioral Science, 1971, 7, 466-492.

Delbecq, A., Van de Ven, A., and Gustafson, D. Group techniques for program planning. Glenview, IL: Scott, Foresman, 1975.

DeSmet, A. A., Fryback, D.C., and Thornbury, J. R. A second look at the utility of radiographic skull examination for trauma. American Journal of Radiology, 1979, 132, 95-99.

Domas, P. A., Goodman, B. C., and Peterson, C. R. Bayes' Theorem: Response scales and feedback (Technical Report No. 037230-5-T). The University of Michigan, Engineering Psychology Laboratory, September, 1972.

DuCharme, W. M., and Donnell, M. L. Intrasubject comparison of four response modes for "subjective probability" assessment. *Organizational Behavior and Human Performance*, 1973, *10*, 108-117.

DuCharme, W. M. *A review and analysis of the phenomenon of conservatism in human inference*. Houston, TX: Rice University, unpublished manuscript, 1969.

Edwards, A. L., and Thurstone, L. L. An internal consistency check for scale values by the method of successive intervals. *Psychometrika*, 1952, *17*, 169-180.

Edwards, A. L. *Techniques of attitude scale construction*. New York: Appleton-Century-Crofts, 1957.

Edwards, W. The theory of decision making. *Psychological Bulletin*, 1954, *51*, 380-417.

Edwards, W., John, R. S., and Stillwell, W. G. *Research on the technology of inference and decision* (Technical Report 77-6). Los Angeles, CA: University of Southern California, Social Science Research Institute, 1977.

Edwards, W., John, R. S., and Stillwell, W. G. *Research on the technology of inference and decision* (Technical Report 79-1). Los Angeles, CA: University of Southern California, Social Science Research Institute, 1979.

Einhorn, H. J., and Hogarth, R. M. Behavioral decision theory: Processes of judgment and choice. *Annual Review of Psychology*, 1981, *32*, 53-88.

Eyman, R. K., and Kie, P. J. A model for partitioning judgment error in psychophysics. *Psychological Bulletin*, 1970, *74*, 35-46.

Fine, T. L., and Kaplan, M. Joint orders in comparative probability. *Annals of Probability*, 1977, *5*, 161-179.

Fischer, G. W. When oracles fail - A comparison of four procedures for aggregating subjective probability forecasts. *Organizational Behavior and Human Performance*, 1981, *28*, 96-110.

Freeling, A. N. S. *Reconciliation of inconsistent ratio judgments*. Falls Church, VA: Decision Science Consortium, Inc., 1981.

Fujii, T. *Conservatism and discriminability in probability estimation as a function of response mode*. University of Michigan, unpublished manuscript, 1967.

Goodman, B. C. *Direct estimation procedures for eliciting judgments about uncertain events* (Engineering Psychology Laboratory Technical Report 011313-5-T). University of Michigan, 1973.

Gough, R. The effect of group format on aggregate subjective probability distributions. In Wendt, D., and Vlek, C. (Eds.). *Utility, probability, and human decision-making*. Dordrecht-Holland: Reidel, 1975.

Green, P. E. Critique of L. H. Smith's article. *Management Science*, 1967, *14*, 250-252.

Guilford, J. P. *Psychometric methods*, 2nd Ed. New York: McGraw-Hill, 1954.

Gulliksen, H. *Theory of mental tests*. New York: 1950.

Gustafson, D., Shukla, R., Delbecq, A., and Walster, G. A comparative study of differences in subjective likelihood estimates made by individuals, interacting groups, Delphi groups, and nominal groups. *Organizational Behavior and Human Performance*, 1973, *9*, 280-291.

Hoffman, J., and Peterson, C. R. *A scoring rule to train probability assessors* (Engineering Psychology Laboratory Technical Report 037230-4-T). Ann Arbor, MI: University of Michigan, 1972.

Hoffman, P. J. The paramorphic representation of clinical judgment. *Psychological Bulletin*, 1960, *57*, 116-131.

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

Hunns, D. M., and Daniels, B. K. *The method of paired comparisons*, undated, SINDOC (80)90.

John, R. S., and Edwards, W. *Estimating subjective probability distributions* (Working Paper). Los Angeles: University of Southern California, Social Science Research Institute, 1977.

Johnson, E. M. *The perception of tactical intelligence indications: A replication* (Technical Paper 282). Alexandria, VA: U.S. Army Research Institute for the Behavorial and Social Sciences, 1977.

Jones, F. N. Overview of psycholophysical scaling methods. In E. C. Carterette and M. P. Friedman (Eds.) *Handbook of perception*, *Vol. 2*. New York: Academic Press, 1974.

Kabus, I. You can bank on uncertainty. *Harvard Business Review*, 1976, *54*, 95-105.

Kahneman, D., and Tversky, A. Intuitive prediction: Biases and corrective procedures. Management Science, 1979, 12, 313-327.

Kelley, C. W., and Peterson, C. R. Probability estimates and probabilistic procedures in current intelligence analysis (Report on Phase 1). Gaithersburg, MD: Federal Systems Division, IBM Corporation, 1970.

Kemeny, J. G., and Snell, J. L. Mathematical models in the social sciences. New York: Blaisdell, 1962.

Kendall, M. G. Ranks and measures. Biometrika, 1962, 49.

Krantz, D. H., Luce, R. D., Suppes, P., and Tversky, A. Foundations of measurement, Vol. 1. New York: Academic Press, 1971.

Kruskal, J. B., and Wish, M. Multidimensional scaling. Beverly Hills: Sage Publications, 1978.

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.

Linstone, H., and Turoff, M. The Delphi method: Techniques and applications. Reading, MA: Addison-Wesley, 1975.

Lodge, M., Tanenhaus, J., Cross, D., Tursky, B., Foley, M. A., and Foley, H. The calibration and cross-model validation of ratio scales of political opinion in survey research. Social Science Research, 1976, 5, 325-347.

Luce, R. D. Individual choice behavior. New York: Wiley, 1959.

Luce, R. D., and Raiffa, H. Games and decisions. New York: Wiley, 1957.

Ludke, R. L., Stauss, F. F., and Gustafson, D. H. Comparison of five methods for estimating subjective probability distributions. Organizational Behavior and Human Performance, 1977, 19, 162-179.

Lusted, L. B. A study of the efficiency of diagnostic radiologic procedures (Final report on diagnostic efficacy). Chicago, IL: Efficacy Study Committee of the American College of Radiology, 1977.

Meister, D. Comparative analysis of human reliability models (Report L0074-107). Westlake Village, California, Bunker-Ramo Electronics Systems Division, 1971.

Morris, P. Bayesian expert resolution (Ph.D. dissertation, University Microfilm No. 72-5959). Ann Arbor, MI: 1971.

Morris, P. Decision analysis expert use. Management Science, 1974, 20, 1233-1241.

Morris, P. Modeling experts (unpublished manuscript). Xerox Corporation, Palo Alto Research Center, 1975.

Morris, P. Combining expert judgments: A Bayesian approach. Management Science, 1977, 23, 679-693.

Morrison, D. G. Critique of: Ranking procedures and subjective probability distributions. Management Science, 1967, 14, 253-254.

Murphy, A. H., and Winkler, R. L. Credible interval temperature forecasting: Some experimental results. Monthly Weather Review, 1974, 102, 784-794.

Murphy, A. H., and Winkler, R. L. The use of credible intervals in temperature forecasting: Some experimental results. In H. Jungermann and G. de Zeeuw (Eds.) Decision making and change in human affairs. Dordrecht-Holland: D. Reidel Publishing Company, 1977. (a)

Murphy, A. H., and Winkler, R. L. Can weather forecasters formulate reliable probability forecasts of precipitation and temperatives? National Weather Digest, 1977, 2, 2-9. (b)

Nunnally, J. Psychometric theory. New York: McGraw Hill, 1979.

Phillips, L. D., and Edwards, W. Conservatism in a simple probability inference task. Journal of Experimental Psychology, 1966, 72, 346-354.

Pill, J. The Delphi method: Substance, context, a critique and an annotated bibliography. Socio-Economic Planning Sciences, 1971, 5, 57-71.

Rigby, L. V., and Edelman, D. A. A predictive scale of aircraft emergencies. Human Factors, 1968, 10(5), 475-482.

Roberts, F. S. Measurement theory. Encyclopedia of mathematics and its applications, Vol. 7. Addison-Wesley Publishing Co., 1979.

Ross, R. T. Optimum orders for the presentation of pairs in the method of paired comparisons. Journal of Educational Psychology, 1934, 25, 375-382.

Ross, J., and Di Lollo, V. Judgment and response in magnitude estimation. Psychological Review, 1971, 78, 515-527.

Rowse, G., Gustafson, D., and Ludke, R. Comparison of rules for aggregating subjective likelihood ratios. Organizational Behavior and Human Performance, 1974, 12, 274-285.

Rummel, R. J. Some empirical findings on nations and their behavior. World Politics, 1969, 21, 226-241.

Saaty, T. L. A scaling method for priorities in hierarchical structures. Journal of Mathematical Psychology, 1977, 15, 234-281.

Saaty, T. L. The analytic hierarchy process. New York: McGraw-Hill, 1980.

Sackman, H. Delphi critique. Lexington, MA: D. C. Heath & Co., 1975.

Schaefer, R. E., and Borcherding, K. The assessment of subjective probability distributions: A training experiment. Acta Psychologica, 1973, 37, 117-129.

Schriesheim, C. A., and Schriesheim, J. F. Development and empirical verification of new response categories to increase the validity of multiple response alternative questionnaires. Educational and Psychological Measurement, 1974, 34, 877-884.

Schriesheim, C. A., and Schriesheim, J. F. The invariance of anchor points obtained by magnitude estimation and paired-comparison treatment of complete ranks scaling procedures: An empirical comparison and implications for validity of measurement. Educational and Psychological Measurement, 1978, 38, 977-983.

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

Seaver, D. A. Assessing probability with multiple individuals: Group interaction versus mathematical aggregation (SSRI Research Report 78-3). Los Angeles: University of Southern California, Social Science Research Institute, December 1978.

Seaver, D. A., von Winterfeldt, D., and Edwards, W. Eliciting subjective probability distributions on continuous variables. Organizational Behavior and Human Performance, 1978, 21, 379-391.

Selvidge, J. E. Assigning probabilities to rare events. In D. Wendt, and C. Vlek (Eds.) Utility, probability, and human decision making. Dordrecht-Holland: Reidel, 1975.

Selvidge, J. E. Experimental comparison of different methods for assessing the extremes of probability distributions by the fractile method (Management Science Report Series). Boulder: University of Colorado, 1975.

Sjoberg, L. Three models for the analysis of subjective ratios. Scandinavian Journal of Psychology, 1971, 12, 217-240.

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

Smith, L. H. Ranking procedures and subjective probability distributions. *Management Science*, 1967, *14*, B236-B269.

Spetzler, C. S., and Stael von Holstein, C. A. S. Probability encoding in decision analysis. *Management Science*, 1975, *22*, 340-358.

Spetzler, C. S., and Zamora, R. M. *Decision analysis of a facilities investment and expansion problem*. Proceedings of the Sixth Triennial Symposium sponsored by the Engineering Economy Division of ASEE and AIIE, June 1971, published by Engineering Economist.

Stael von Holstein, C. A. S. The effect of learning on the assessment of subjective probability distributions. *Organizational Behavior and Human Performance*, 1971, *6*, 304-315. (a)

Stael von Holstein, C. A. S. An experiment in probabilistic weather forecasting. *Journal of Applied Meteorology*, 1971, *10*, 635-645. (b)

Stael von Holstein, C. A. S. Probabilistic forecasting: An experiment related to the stock market. *Organizational Behavior and Human Performance*, 1972, *8*, 139-158.

Stevens, S. S. The direct estimation of sensory magnitudes--loudness. *American Journal of Psychology*, 1956, *69*, 1-25. (c)

Stillwell, W. G., Seaver, D. A., and Edwards, W. *The effects of response scales of likelihood ratio judgments* (SSRI Research Report 77-5). Los Angeles: University of Southern California, Social Science Research Institute, August 1977.

Suppes, P., and Zinnes, J. L. Basic measurement theory. In R. D. Luce, R. R. Bush, and E. Galanter (Eds.) *Handbook of mathematical psychology*, *Vol. 1*. New York: Wiley, 1963.

Swain, A. D. Field calibrated simulation. In *Proceedings of the symposium on human performance quantification in system effectiveness*. Washington, D.C.: Naval Material Command and The National Academy of Engineering, January 1967, IV-A-1 to IV-A-21.

Swain, A. D. Development of a human error rate data bank. In J. P. Jenkins (Ed.) *Proceedings of U.S. Navy Human Reliability Workshop*, 22-23 July, 1970. Washington, D.C.: Naval Ship Systems Command, Office of Naval Research and Naval Development Center, February 1971.

Swain, A. D. *Estimating human error rates and their effects on system reliability* (SAND77-1240). Albuquerque: Sandia, March, 1978.

Swain, A. D., and Guttmann, H. E. Human reliability analysis applied to nuclear power, *Proceedings of 14th Annual Reliability and Maintainability Conference*, Inst. of Electrical and Electronic Engineers, New York, January 1975, 116-119.

Swain, A. D., and Guttmann, H. E. *Handbook of human reliability analysis with emphasis on nuclear power plant applications* (Draft Report NUREG/CR-1278). Washington, D.C.: U.S. Nuclear Regulatory Commission, October 1980.

Thurstone, L. L. A law of comparative judgment. *Psychological Review*, 1927, *34*, 273-286.

Thurstone, L. L. Rank order as a psychophysical method. *Journal of Experimental Psychology*, 1931, *14*, 187-201.

Torgerson, W. S. *Theory and methods of scaling*. New York: Wiley, 1958.

Torgerson, W. In David Sills (Ed.) *International encyclopedia of the social sciences*. McMillian, 1968.

Tversky, A. Intransitivity of preferences. *Psychological Review*, 1969, *76*, 31-48.

Tversky, A., and Kahneman, D. Judgment under uncertainty: Heuristics and biases. *Science*, 1974, *185*, 1129-1131.

Tversky, A., and Kahneman, D. Causal schemata in judgments under uncertainty. In M. Fishbein (Ed.) *Progress in social psychology*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1977.

Van de Ven, A., and Delbecq, A. The effectiveness of nominal, Delphi, and interacting group decision making processes. *Academy of Management Journal*, 1974, *17*, 605-621.

Wallsten, T. S. *Bias in evaluating diagnostic information* (Report No. 157). Chapel Hill, N.C.: L. L. Thurstone Psychometric Laboratory, 1978.

Wallsten, T. S., and Budescu, D. V. *Encoding subjective probabilities: A psychological and psychometric review* (Draft Report). Research Triangle Park, NC: U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, April 1980.

Wherry, R. J. Orders for the presentation of pairs in the method of paired comparisons. *Journal of Experimental Psychology*, 1938, *23*, 651-666.

Winkler, R. L. The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical Association*, 1967, *62*, 776-800.

Winkler, R. L. The consensus of subjective probability distributions. *Management Science*, 1968, *15*, 61-75.

Winkler, R. L. Probabilistic prediction: Some experimental results. *Journal of the American Statistical Association*, 1971, *66*, 675-685.

Winkler, R. L. The assessment of probability distributions for future security prices. In J. L. Bicksler (Ed.), *Methodology in finance-investments*. Lexington, MA: Lexington Books, Heath, 1972, 129-148.

Winkler, R. L., and Cummings, L. On the choice of a consensus distribution in Bayesian analysis. *Organizational Behavior and Human Performance*, 1972, *7*, 63-76.

von Winterfeldt, D. *Some sources of incoherent judgments in decision analysis*. Falls Church, VA: Decision Science Consortium, Inc., November, 1980.

Yellot, J. I. The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 1977, *15*, 109-144.

Zajonc, R. Social facilitation. *Science*, 1965, *149*, 269-274.

Zlotnick, J. Bayes' theorem for intelligence analysis. In J. A. Jacquez (Ed.), *Computer diagnosis and diagnostic methods*. Springfield, IL: Charles C. Thomas, 1972, 180-190.

GLOSSARY

Brier Score - A proper scoring rule used primarily to evaluate meteorologist's forecasts. It is so structured as to minimize the judges expected score for any estimate only when true subjective probability is given. The score, S, for n forecasts each with r possibilities is defined by:

$$S = \frac{1}{n} \sum_{i=1}^{n} (1-r_j)^2$$

where r_j is the probability assigned to the event (precipitation, no precipitation) that actually occurs.

Calibration - The extent to which the probabilities assigned to events are of the same magnitude as the corresponding empirical relative frequencies. Formally, an assessor is calibrated if, over the long run, for all propositions assigned a given probability, the proportion that is true is equal to the probability assigned. An assessors' calibration can be empirically evaluated by observing the probability assessments, verifying the associated propositions, and then observing the proportion that is true in each response category. This attribute has also been called "realism," "external validity," "secondary validity," and "reliability."

Delphi Method - This name covers a wide range of procedures for control of group judgment in prediction and estimation problems. These procedures have in common three features: (1) anonymity of group members, (2) iteration of judgments with controlled feedback, and (3) statistical aggregation of final individual responses to provide the group response.

Direct Estimation Methods - Procedures for estimating probabilities in which the assessor is asked to provide a quantitative estimate of the desired value. Thus, judgments of probabilities, odds or of the ratio of the likelihoods of two occurrences would all be direct estimates.

Dominance Judgment - A judgment of direction, i.e., "greater than" or "less than," between pairs of stimuli with regard to a single attribute. For probability, the judgment would be "more likely" or "less likely."

Empirical/External Validity - The extent to which a measurement corresponds with reality usually determined by correlation between the measurement and another independent measure of reality.

Fractile Method - A method for eliciting approximations of continuous probability distributions in which the assessor is asked for values of the uncertain quantity that divide the distribution into parts with specified likelihood. A common form of this procedure is to ask for the median, dividing the distribution in half, and then values that divide each of the halves in half.

Fractionation Method - A general class of scaling methods in which the subject responds to two stimuli on the basis of the perceived ratio of subjective magnitude. In one form, the subject is presented with two stimuli and instructed to report the subjective ratio between them with respect to the designated attribute. In the other form, the subject's task is to report when two stimuli stand in a prescribed ratio.

Graph Method - A method for assessing continuous probability distributions in which the assessor simply draws the distribution on lined graph paper and labels the axes.

Graphic Rating - A particular type of rating scale in which the rating continuum is arbitrarily divided into some number of equal categories (e.g., one inch segments). The subject is asked to indicate on which segment the stimulus is appropriately placed.

Heuristic - A rule of thumb for simplifying complex problems in a way that increases the probability that a solution will be found within a "reasonable" length of time. To be contrasted with an algorithm which, through exhaustive search of all possibilities, guarantees a solution if there is one.

Indirect Elicitation Method - Any of the estimation procedures in which the values of interest are derived from the assessors' responses rather than being the actual responses. Indirect procedures require certain assumptions about the judgment or the relationship between judgment and scale for probabilities to be determined.

Likelihood Ratio - The ratio between the probability of some datum conditional on one hypothesis and the probability of the same datum conditional on an alternative hypothesis. A measure in Bayesian information processing of the change in the odds of two competing hypotheses reflecting the diagnosticity of a new datum or series of data.

Magnitude Estimation - A specific fractionation method in which direct estimates of subjective attribute ratios are obtained. From a set of *n* stimuli, one stimuli is chosen as a standard (anchoring stimulus). Each of the remaining stimuli is presented with the standard, and an estimate of the ratio is obtained. The unit of measurement is specified by assigning a number to one of the stimuli arbitrarily, the scale values of the remaining stimuli being calculated directly from the ratios.

Multidimensional Scaling Procedures - A class of scaling techniques for relating similarity judgments to points in a geometrical space. MDS is used in two ways: to discover what dimensions people use when responding to a class of stimuli, and to investigate the psychological utilization of known physical dimensions.

Nominal Group Procedure - A procedure for obtaining a group judgment of an uncertain quantity in which the interaction among group members is carefully controlled. The procedure consists of four steps: (1) silent judgments by individuals in the presence of the group; (2) presentation to the group of all individual judgments without discussion; (3) group discussion of each judgment for clarification and evaluation; (4) individual reconsideration of judgments and mathematical combination.

Odds Judgment - A judgment of likelihood in which the assessor estimates the relative probability of two hypotheses (often a hypothesis and its negation) under a given state of the world. For example, the assessor might be asked for "the odds that the next person walking through the door will be below 74 inches in height." A judgment of 4 to 1 would mean that the assessor feels that the chance that the person would be below 74 inches is four times as likely as the chance that the person will be 74 inches or taller.

Paired Comparison Method - A method for obtaining estimates of stimulus scale values. Each stimulus is paired with each other stimulus. Each pair is presented to the subject, whose task is to indicate which member of the pair appears greater with respect to the attribute to be scaled. The basic data are the proportions of times each stimulus *k* is judged greater than any other stimulus *j*.

Proper Scoring Rule - A rule assigning scores to probability assessments in such a way that an assessor can maximize the expected score only by reporting a *true* subjective probability. Thus, there is no way to improve upon a score by "hedging" an assessment or making it too extreme. One of three scoring rules are usually used in assessment of probabilities: the quadratic, spherical, and logarithmic. The Brier score, used for evaluating meteorological forecasts, is a special case of the quadratic scoring rule.

Ranking Method - A scaling procedure in which the subject ranks the stimuli in order with respect to the attribute to be scaled. From these judgments the proportion of times stimulus *k* was perceived as greater than stimulus *j* is deduced. It is assumed that, in ranking the stimuli, the subject compares each stimulus with every other one.

Rating Method - A scaling procedure having the subject rate each stimulus with respect to the attribute. The stimuli are presented one at a time. The rating may be expressed on a numerical scale, adjective scale, or a graphic scale.

Regression - One form of analysis for determining the predictability of a criterion on the basis of one or more predictors that takes into account the degree of correlation between each predictor and the criterion. The method of calculation depends on the choice of loss function (i.e., how the error distances are weighted), as well as the scaling of both independent and dependent variables. The usual procedure is to use a least-squares loss function and linear scaling.

Reliability - The extent to which measurements are repeatable. The amount of random error of measurement determines the reliability. Major potential sources of random error (and types of reliability to be checked) include different persons making the measurement, different occasions, alternative forms of the same instrument, and slight variation in circumstances. Reliability is usually measured as the correlation between the same measure taken at different times.

Statisticized Group Approach - Groups formed so that a statistical procedure can be used to obtain a group judgment from individuals making their own judgments. Statisticized groups need not be face-to-face or interact in any way and, in fact, need not even be in the same geographic location. Individual judgments are simply aggregated via some preselected rule (often they are simply averaged).

Sorting Method - A scaling procedure in which the subject's task is to sort the stimuli into piles so that the first pile contains those stimuli that are most positive with respect to the attribute; the second pile, the stimuli next most positive, etc. It is only necessary that the piles be in rank order with respect to the attribute. Often the piles may be identified with adjectives which progress from extremely positive to zero or extremely negative, depending on the particular attribute.

Tertile Method - One of the fractile methods in which the assessor is asked for values of the uncertain quantity that divide the distribution into three equally likely parts.

Distribution

U.S. NRC Distribution Contractor (CDSI) (620)
7300 Pearl Street
Bethesda, MD 20014
385 copies for AN,RX
25 copies for NTIS
210 copies for Author-Selected Distribution

Dr. Lee Abramson
Applied Statistics Branch
Management Program Analysis Office
U.S. Nuclear Regulatory Commission
Washington, DC 20555

Prof. Jack A. Adams
Department of Psychology
University of Illinois at Urbana Champaign
Champaign, IL 61820

Prof. S. Keith Adams
Department of Industrial Engineering
212 Marston Hall
Iowa State University
Ames, IA 50011

American Institutes for Research
41 North Road
Bedford, MA 01730

Dr. Arthur Bachrach
Behavioral Sciences Department
U.S. Naval Medical Research Institute
8901 Wisconsin Avenue
Bethesda, MD 20014

Dr. A. D. Baddeley
Director, Applied Psychology Unit
Medical Research Council
15 Chaucer Road
Cambridge CB22EF
England

Dr. Werner Bastl
GRS
Bereich Systeme
Forschungsgelande
8046 Garching
Federal Republic of Germany

Dr. R. B. Basu
Bell Northern Research
P. O. Box 3511, Station C
Ottawa, ON
Canada

Dr. Robert P. Bateman
Senior Scientist
Human Factors Engineering Group
Systems Research Laboratories, Inc.
2800 Indian Ripple Road
Dayton, OH 45440

Dr. Lee Roy Beach
Department of Psychology (NI-25)
University of Washington
Seattle, WA 98195

Dr. David Beattie
Ontario Hydro H-14
700 University Avenue
Toronto, ON
Canada M5G 1X6

Mr. C. J. E. Beyers
Licensing Branch (Standards)
Atomic Energy Board
Private Bag X256
Pretoria 0001
Republic of South Africa

Dr. Kairin Borcherding
Sonderforschungsbereich (SFB)
24 an der Universitat Mannheim
68 Mannheim L13 15-17
West Germany

Prof. Mark Brecht
Psychology Department
University of New Mexico
Albuquerque, NM 87131

Dr. Leon Breen
Brookhaven National Laboratories
Building 197C
Upton, NY 11973

Dr. Robert Brune
Human Performance Technologies, Inc.
P. O. Box 3816
Thousand Oaks, CA 91359

Mr. Joseph O. Bunting
Division of Waste Management
Nuclear Material Safety and Safeguards Office
U.S. Nuclear Regulatory Commission
7915 Eastern Avenue
Silver Spring, MD 20555

Mme. Annick Carnino
Electricite de France
Service de la Production Thermique
71, Rue de Miromesnil
75008 Paris
France

Dr. Alphonse Chapanis
Department of Psychology
John Hopkins University
Charles and 34th Streets
Baltimore, MD 21218

Dr. Julien M. Christensen
Director, Human Factors Office
General Physics Corporation
1010 Woodman Drive #240
Dayton, OH 45432

Dr. Patricia A. Comella
Deputy Director
Health, Siting and Waste Management Division
U.S. Nuclear Regulatory Commission
Washington, DC 20555

Dr. Vincent T. Covello
Office of Scientific, Technological, and
International Affairs
National Science Foundation
1800 G. Street, NW
Washington, DC 20550

CDR Michael Curley
Operations Research Programs
Office of Naval Research
Ballston Tower #1
800 N. Quincy Street
Arlington, VA 22217

Dr. Judith A. Daly
Program Manager, Systems Sciences Office
Defense Advanced Research Projects Agency
1400 Wilson Blvd.
Arlington, VA 22209

Dr. Ed M. Dougherty, Jr.
Technology for Energy Corporation
10770 Dutchtown Road
Knoxville, TN 37922

Dr. Gerry Doyle
Building #K-1007
Mailstop Room 1058
Union Carbide Corporation
Computer Science Division
Oak Ridge, TN 37830

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

Dr. Hillel Einhorn
Center for Decision Research
University of Chicago
1101 East 58th Street
Chicago, IL 60637

Dr. David Embrey
National Centre for Systems Reliability
UKAEA
Wigshaw Lane
Culcheth
Warrington WA3 4NE
Cheshire
England

Dr. Donald Emon
Office of Safeguards and Security
Room A21300
U.S. Department of Energy
Germantown, MD 20545

Dr. Hunter Foreman
Building #K-1007
Mailstop Room 1058
Union Carbide Corporation
Computer Science Division
Oak Ridge, TN 37830

Dr. Joseph Fragola
Scientific Applications, Inc.
274 Madison Avenue
Suite 1501
New York, NY 10016

Dr. Dennis Fryback
Health Systems Engineering
University of Wisconsin
1225 Observatory Drive
Madison, WI 53706

Dr. Kenneth Gardner
Applied Psychology Unit
Admiralty Marine Technology Establishmnt
Teddington, Middlesex TW110LN
England

Dr. Robert A. Goldbeck
Ford Aerospace & Communications Corporation
Engineering Service Division
1260 Crossman Avenue MS S-33
Sunnyvale, CA 94086

Mme. Martine Griffon
DIR-ISE
Centre d'Etudes Nucleaires de Grenoble
85X F-38041 Grenoble Cedex
France

Dr. Douglas H. Harris
President
Anacapa Sciences, Inc.
P. O. Drawer Q
Santa Barbara, CA 93102

Dr. Julie Hopson
Human Factors Engineering Division
Naval Air Development Center
Warminster, PA 18974

CDR Kent Hull
Office of Naval Research
Code 410B
Ballston Tower #1
800 N. Quincy Street
Arlington, VA 22217

Mr. David M. Hunns
Research Engineer in Reliability Technology
National Centre of Systems Reliability
UKAEA
Safety & Reliability Directorate
Wigshaw Lane
Culcheth
Warrington WA3 4NE
Cheshire
England

Dr. Anand M. Joglekar
Defense Systems Division
Honeywell, Inc.
600 Second Street, NW
Hopkins, MN 55343

Dr. Edgar Johnson
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Prof. Margaret H. Jones
Institute of Safety and Systems Management
University of Southern California
Los Angeles, CA 90007

Dr. Helmut Jungermann
Institut fur Psychologie
Technische Universitat
Dovestr 1-5
D-1000 Berlin 10, West Germany

Dr. Daniel Kahneman
University of British Columbia
Department of Psychology
#154-2053 Main Mall
University Campus
Vancouver, BC V6T 1Y7
Canada

Dr. Ralph Keeney
Woodward-Clyde Consultants
3 Embarcadero Center, Suite 700
San Francisco, CA 94111

Dr. Albert P. Kenneke
Assistant Director, Technical Review
Office of Policy Evaluation
U.S. Nuclear Regulatory Commission
1717 H. Street, N.W.
Washington, DC 20555

Mr. Howard Kunreuther
International Institute of Applied
Systems Analysis
Laxenburg Castle A-2361
Laxenburg, Austria

Mr. Warren Lewis
Human Engineering Branch
Code 8231
Naval Ocean Systems Center
San Diego, CA 92152

Dr. Sarah Lichtenstein
Decision Research
1201 Oak Street
Eugene, OR 97401

Mr. Pierre M. Lienart
Institute of Nuclear Power Operations
1800 Water Place
Atlanta, GA 30339

Mr. William J. Luckas, Jr.
Brookhaven National Laboratory
Upton, NY 11973

Dr. Robert Lupinacci
Office of Safeguards and Security
Room A21300
U.S. Department of Energy
Germantown, MD 20545

LUTAB
Attn: Library
P. O. Box 52
S-161 26 Bromma
Sweden

Mr. Gerald S. Malecki
Office of Naval Research
Engineering Psychology Programs
Ballston Tower #1
800 N. Quincy St.
Arlington, VA 22217

Dr. David Meister
1111 Wilbur Avenue
San Diego, CA 92109

Dr. Michael Melich
Communications Sciences Division
Code 7500
Naval Research Laboratory
Washington, DC 20275

Mr. Morton Metersky
Naval Air Development Center
Human Factors Engineering Division
Warminster, PA 18974

Dr. Lorna A. Middendorf
1040 Berkshire
Grosse Point Park, MI 48230

Dr. George Moeller
Human Factors Engineering Branch
Submarine Medical Research Lab
Naval Submarine Base
Box 900
Groton, CT 06340

Mr. William M. Murphey
U.S. Arms Control & Disarmament Agency
State Department Building
21st and Virginia Avenue, N.W.
Room 4947
Washington, DC 20451

Commander
Naval Air Systems Command
Human Factors Programs
NAVAIR 340F
Jefferson Plaza 1
Washington, DC 20361

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

Prof. Donald A. Norman
Center for Human Information Processing
University of California at San Diego
San Diego, CA 92093

Dr. Kent Norman
Department of Psychology
University of Maryland
College Park, MD 20742

Dr. John J. O'Hare
Assistant Director
Engineering Psychology Programs
Office of Naval Research
Ballston Tower #1
800 N. Quincy Street
Arlington, VA 22217

Dr. Jessie Orlansky
Institute for Defense Analysis
400 Army-Navy Drive
Arlington, VA 22202

Mr. Reider Ostvik
SINTEF
N7034 Trondheim
NTH
Norway

Dr. Ray Parsick
Head, Safeguards Evaluation Section
International Atomic Energy Agency
Wagramerstrasse 5, P. O. Box 100
A-1400, Vienna, Austria

Dr. Lawrence M. Potash
Project Manager, Criteria & Analysis Division
Institute of Nuclear Power Operations
1820 Water Place
Atlanta, GA 30339

Dr. E. C. Poulton
MRC Applied Psychology Unit
15 Chaucer Road
Cambridge, CB2 2EF
England
United Kingdom

Mr. Ortwin Renn
KFA Julich
KUU
Postfach 1913
5170 Julich
West Germany

Mr. Bo Rydnert
LUTAB
P. O. Box 52
S-161 26 Bromma
Sweden

Dr. Kenneth E. Sanders
Division of Safeguards
Nuclear Material Safety and Safeguards Office
U.S. Nuclear Regulatory Commission
Washington, DC 20555

Dr. David Schum
7416 Timberrock Road
Falls Church, VA 22043

Mr. Jeffrey P. Schwartz
Decision Science Consortium, Inc.
Suite 421
7700 Leesburg Pike
Falls Church, VA 22043

Ms. Mary Jo Seamann
Division of Waste Management
Nuclear Material Safety and Safeguards Office
U.S. Nuclear Regulatory Commission
7915 Eastern Avenue
Silver Spring, MD 20555

Dr. David A. Seaver (50)
Vice President
Decision Science Consortium, Inc.
Suite 421
7700 Leesburg Pike
Falls Church, VA 22043

Dr. Arthur I. Siegel
Applied Psychological Services
404 E. Lancaster
Wayne, PA 19087

Dr. Kurt J. Snapper
The Maxima Corporation
7315 Wisconsin Avenue
Suite 900N
Bethesda, MD 20014

Dr. Eugene Sparks
Material Transfer SG
Division of Safeguards
Licensing Branch
U.S. Nuclear Regulatory Commission
Mailstop 881-SS
Washington, DC 20555

Dr. Michael E. Stephens
Nuclear Safety Division
OECD Nuclear Energy Agency
38, Boulevard Suchet
F-75016 Paris
France

Ms. Catherine Stewart
SSDC
EG&G Idaho, Inc.
P. O. Box 1625
Idaho Falls, ID 83415

Dr. William G. Stillwell
The Maxima Corporation
7315 Wisconsin Avenue
Suite 900N
Bethesda, MD 20014

Mr. Jean P. Stolz
Electricite de France
Service de la Production Thermique
71, Rue de Miromesnil
75008 Paris
France

Mr. Toshiaki Tobioka
Senior Engineer
Reactor Safety Code Dev. Lab.
Division of Reactor Safety Evaluation
Tokai Research Establishment
JAERI
Tokai-mura, Naka-gun
Ibaraki-ken
Japan

Dr. Martin A. Tolcott
Director, Engin. Psychol. Prog.
U.S. Office of Naval Research
Psychological Sciences Division
Ballston Tower #1
Room 711, 800 N. Quincy St.
Arlington, VA 22217

Dr. V. R. R. Uppuluri
Mathematics & Statistics Research Dept.
Building 9704-1
Oak Ridge National Laboratory
P. O. Box 4
Oak Ridge, TN 37830

Dr. Harold P. Van Cott
Chief Scientist
Biotechnology, Inc.
3027 Rosemary Lane
Falls Church, VA 22042

Dr. Stein Weissenberger
University of California
Lawrence Livermore Laboratories
Engineering Research Division
P. O. Box 808
Livermore, CA 94550

Dr. Chris Whipple
Electric Power Research Institute
3412 Hillview Avenue
Palo Alto, CA 94304

Mr. David Whitfield
Head, Ergonomics Development Unit
Psychology Department
The University of Aston in Birmingham
Gosta Green
Birmingham B4 7ET
England
United Kingdom

Dr. Robert Williges
Human Factors Laboratory
Virginia Polytechnical Institute
and State University
130 Wittemore Hall
Blacksburg, VA 24061

Mr. Jan Wirstad
Ergonomrad AB
Box 10032
S-65010 Karlstad
Sweden

Mr. John Wreathall
NUS
4 Research Place
Rockville, MD 20850

Mr. Jan Wright
Det Norske Veritas
P. O. Box 6060 Etterstad
Oslo 6, Norway

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

Prof. Takeo Yukimachi
Department of Administrative Engineering
Keio University
Hiyoshi, Yokohama
223 Japan

Internal Sandia Distribution

1223	B. J. Bell
1223	B. H. Finley
1223	D. P. Miller
1223	R. R. Prairie
1223	L. M. Weston (50)
3141	L. J. Erickson (5)
3151	W. L. Garner (3)
8214	M. A. Pound

✿U.S. GOVERNMENT PRINTING OFFICE: 1982-0-576-021/566