NUREG/CR-2743
SAND82-7054
AN,RX
Printed March 1983
CONTRACTOR REPORT
UNLIMITED RELEASE

Procedures for Using Expert Judgment to Estimate Human Error Probabilities in Nuclear Power Plant Operations

David A. Seaver, William G. Stillwell
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

TOTAL PAGES: 126

Includes all paginated and Non-Paginated Pages.

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

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

ABSTRACT

This report describes and evaluates several procedures for using expert judgment to estimate human error probabilities (HEPs) in nuclear power plant operations. These HEPs are currently needed for several purposes, particularly for probabilistic risk assessments. Data do not exist for estimating these HEPs, so expert judgment can provide these estimates in a timely manner. NUREG/CR-2255 suggested that expert judgment can provide reasonably valid and reliable HEP estimates, if used carefully and systematically. Five judgmental procedures are described here: paired comparisons, ranking and rating, direct numerical estimation, indirect numerical estimation and multiattribute utility measurement. These procedures are evaluated in terms of several criteria: quality of judgments, difficulty of data collection, empirical support, acceptability, theoretical justification, and data processing. Situational constraints such as the number of experts available, the number of HEPs to be estimated, the time available, the location of the experts, and the resources available are discussed in regard to their implications for selecting a procedure for use. Details for implementing the procedures and necessary calculations are included in appendices. These descriptions will be the basis of subsequent research testing the use of several procedures.

ACKNOWLEDGMENTS

This work was supported by the U.S. Nuclear Regulatory Commission, Office of Nuclear Regulatory Research, with Dr. Thomas G. Ryan as program manager. It is part of a larger effort directed by Sandia National Laboratories on human reliability analysis in nuclear power plant operations. The authors would like to thank Dr. Thomas G. Ryan of the NRC, Dr. Richard R. Prairie of Sandia National Laboratories, and other Sandia staff for helpful comments and suggestions. Special thanks goes to Dr. Louise M. Weston of the Statistics, Computing, and Human Factors Division, Sandia National Laboratories, who served as technical monitor. Her careful review and thoughtful suggestions have vastly improved this report.

Decision Science Consortium, Inc. would like to express its appreciation to the authors, who contributed considerable personal time and energy on this project both as employees, and, later, as consultants. Their conscientious efforts helped to make possible this contribution to this important topic. The senior author may be reached at (703) 255-2981.

1.0 INTRODUCTION

The purpose of this report is to present a set of procedures that can be used to estimate human error probabilities (HEPs) in nuclear power plant (NPP) operations using expert judgment. It is part of an ongoing research program being conducted for the U.S. Nuclear Regulatory Commission (NRC) by Sandia National Laboratories and its subcontractors. An earlier review of relevant literature in NUREG/CR-2255 (Stillwell, Seaver, and Schwartz, 1982) indicated the potential of this approach to estimating HEPs. This research program is guided by recognition of the role humans play in NPP operations, both in producing and in mitigating accidents. A major effort in the program has been the development of the Handbook of Human Reliability Analysis with Emphasis on Nuclear Power Plant Applications (Swain and Guttmann, 1980) that is currently being revised. The Handbook lays out procedures for estimating human reliability based on probability tree diagrams called human reliability analysis (HRA) event trees, and provides probability estimates for various specific types of errors and their distributions. These probability estimates, however, are generally based on extrapolations from only partially related performance measures, or on the expert judgment of the authors. Thus, at present, a major problem in human reliability analysis is the lack of high quality HEP estimates. One way to provide better estimates is to use expert judgment in a more structured and systematic way. This report describes several procedures that can be used under a variety of circumstances to produce needed HEPs. It provides the basis for a test of several of these procedures.

1.1 Need for Expert Judgment of HEPs

Clearly the best way to obtain good HEP estimates is through well-controlled and carefully executed empirical studies. Such studies, however, would be extremely costly and time consuming at actual plants, if possible at all. Empirical data collection in control room simulators is being funded by NRC, but the data are relatively costly and time consuming to collect, and the data cannot be calibrated quickly.

Yet HEP estimates are needed now. The accident at Three Mile Island (TMI) has instigated the improvement and use of probabilistic risk assessment (PRA) to quantify the risks associated with nuclear power plants. Because, in part, of the human errors involved at TMI (Kemeny, 1979; Rogovin and Frampton, 1980), human reliability has begun to receive special attention in PRAs. The methodology and models for the human reliability analysis parts of PRAs are well-understood and developed (Swain and Guttmann, 1980), but the data needed to provide accurate inputs to the analysis are sparse. Since empirical studies cannot provide these data to meet the needs of current PRAs, expert judgment is the most viable method for obtaining HEP estimates in a timely manner.

If expert judgment is to be used to estimate HEPs, it should not be used in an ad hoc and unsystematic manner. Reliable* and valid estimates are needed. Although evidence described in the following section suggests that experts can make reliable and valid judgments, it also indicates that care must be taken in how judgments are made to avoid systematic errors and unreliability. Thus, the procedures described here have been identified and developed to systematize the use of expert judgment, and to elicit judgments using procedures that, for given circumstances, will minimize biases.

Expert judgment of HEPs is needed not only for PRAs. NPP control rooms are currently being reviewed for human factors deficiencies. The procedures described here could be used to estimate the effects of the deficiencies on human errors and to help prioritize the correction of deficiencies. Expert estimation of HEPs could also serve a useful purpose in control room design. Alternative designs could be compared and evaluated with respect to the HEPs they produce.

1.2 Rationale for the Use of Expert Judgment of Likelihood

The argument that there is no other way to obtain rapidly and completely needed HEPs is not sufficient to justify the use of expert judgment. It must also be demonstrated that such judgments are of sufficient validity to be useful and not misleading. On the balance the literature suggests that expert judgment can be used effectively, if appropriate care is given to the manner in which judgments are obtained (Stillwell et al., 1982).

Research evidence is, however, somewhat divided as to the quality (in terms of both reliability and validity) of quantitative judgment. The use of heuristic estimation rules for making quantitative estimates and their resulting biases in the judgment of probabilities are well documented. For instance, Tversky and Kahneman (1974), Kahneman and Tversky (1979), Wallsten and Budescu (1980), and von Winterfeldt (1980), list and give examples of biases in probabilistic judgment and explanations of the heuristics that may be the causes. However, these examples generally come from narrowly focused laboratory experiments or ad hoc group data collection efforts whose generality is difficult to defend. On the other hand, Lusted (1977), Murphy and Winkler (1977), and others (for example, Goodman et al., 1979; Kabus, 1976), show results that suggest the opposite conclusion, namely, that humans can provide valid, orderly estimates of likelihood.

*In this report, reliability refers to the consistency of judgments. It is usually measured as the correlation between the same judgments made at different times by the same judge or the correlation among judgments made by different experts.

This mixed evidence would seem to cloud the question of whether these probabilistic estimates should be used for decision making or any other purpose. However, upon closer examination of the full breadth of these studies, we can glean some important information about when and where to apply judgmental methods and when to use the data that result from their application.

As mentioned above, many of the studies that demonstrate bias or severe random fluctuation in probabilistic estimates are highly controlled, laboratory experiments. Thus, these experiments are suspect on a broad range of realism grounds, including, for example, subject experience and motivation, reality of stimuli and situation, and strength of experimental manipulation. A subset of judgmental studies, however, have used expert subjects performing something akin to the tasks at which they are expert, and we will look to these for evidence more directly relevant to the question of whether expert generated judgmental probabilities can usefully be used as input to HEP models.

Four substantive areas have provided the bulk of applied research using real experts. The first of these areas, military science and intelligence, has produced findings that are only indirectly related to the quality of the estimates. These findings can be summarized as follows: (1) the experts prefer to respond in numerical form when expressing uncertainty, (2) miscommunication is reduced by the use of numerical probabilities rather than verbal reports for expressing uncertainty, (3) reliability of these estimates is satisfactory (average test-retest correlation of .79 found by Johnson, 1977), and (4) satisfactory use of probabilistic estimation is being made in the intelligence community to solve traditional information processing problems.

The other three substantive areas are somewhat more interesting, since the relative frequencies for the estimated events often become available with the passage of time. The judgments of the experts can therefore be compared to an external criterion to determine their accuracy. In the area of business the evidence is relatively negative, suggesting that experts in this area are poor judges of event probability. A variety of studies (for instance, Stael von Holstein, 1972, and Bartos, 1969) in this context have shown that security analysts, bankers, stock analysts, and other financial "experts" cannot even outperform random guessing strategies. These studies must, however, be viewed with the caveat that many of the financial phenomena about which the judgments are being made are essentially random with respect to the information the analysts have to work with. Therefore, poor prediction is not necessarily demonstrative of an inability of experts to make predictive judgments.

The evidence in the other two areas, weather forecasting and medicine, is quite encouraging. Among those experts that have been extensively evaluated, weather forecasters are the best judges of probabilities. When performing what are essentially the same tasks they do for their

jobs, i.e., estimating probabilities of precipitation and temperature forecasting, they are surprisingly accurate. Their calibration (defined as the degree to which the probability assigned to an event reflects the recorded relative frequency of that event) has been shown to be very good, with plots of their judgments showing very little distance between the line describing perfect calibration and lines describing meteorologists judgments. Experts in the medical field, both nurses and doctors, in studies covering a broad range of diagnostic and prediction tasks, also show a high degree of calibration in their judgmental estimates of uncertain quantities, although there is some tendency to overestimate the likelihood of events with severe consequences.

This last finding, that events with severe consequences are overestimated, points out one of several problems that must be considered when using judgmental data. We would like to point out several considerations as particularly important in the context of the judgment of HEPs. We also would like to recommend that the user of these data take pains either to utilize methods that minimize or remove the impact of these problems or take steps to make allowance for that impact. Some of the more important considerations for judgmental estimation of human error probabilities are:

- Value/probability dependence - This is a more general version of the medical problem discussed above in which the likelihood of events with extreme consequences are overestimated. This problem could be very important in the judgment of HEPs where some of the event pathways lead to extreme crisis situations.

- Small probability estimation - Many error probabilities are likely to be extremely small, for example, less than 10^{-4}. Research evidence suggests that many methods for eliciting probabilistic judgments suffer from insensitivity in the more extreme ranges and these methods should therefore be avoided in the case of HEP judgment.

- Making judgments outside one's area of expertise - Even in the case of weather forecasters, who have been shown to be very good probability estimators, high quality of estimation does not generalize outside the judges' specific area of expertise. Therefore, in the case of HEP judgment, the expert judges must be carefully chosen and not asked to make judgments outside their area.

The techniques presented in this report are designed to take these considerations into account to the extent that the judgmental basis of these phenomena are understood. But it is also important that users of the techniques, and the experts making any required judgments, be aware of these potential problems to help minimize their effects.

1.3 Scope and Use of Report

This report does not provide actual HEP estimates. Rather, it provides directions regarding how to obtain these estimates using expert judgment. It is intended to be used by people with some background in statistics, including a knowledge of psychological scaling methods.

A large number of techniques could be used for obtaining judgmental estimates of HEPs. These techniques have been reviewed (Stillwell et al., 1982), and the most promising procedures have been selected and consolidated for presentation here. The five procedures described in this report--paired comparison, ranking and rating, direct numerical estimation, indirect numerical estimation, and multiattribute utility measurement--represent a wide range of judgmental processes, underlying assumptions, and proven usefulness. This heterogeneity has helped to ensure that at least one of the procedures can be used appropriately under most circumstances.

A caveat regarding the multiattribute utility measurement procedure is required here. This procedure, although widely used for other applications (Keeney and Raiffa, 1976), has not generally been used to estimate probabilities, nor has there been much research concerning its reliability and validity as a scaling technique. Recent research by Embrey (1981a, 1981b), however, suggests it may be a viable method for judgmental estimation of HEPs. It is thus included here so that if additional research shows it to be viable, directions for its use will be available. At this time, it cannot be recommended for use without additional research support.

The following section of this report is its heart. It includes a discussion of general requirements for the use of any judgmental procedure for estimating HEPs. Following this is an overview of the five procedures. Details of their implementation, including step-by-step instructions and worked-through examples are in the Appendices. Finally, this section contains an evaluation of the procedures and specific constraints on their use. The concluding section discusses the use of the procedures, and presents some caveats regarding their use.

Thus, this report is primarily a "how to" document with strong emphasis on understandable guidance and instructions for use. Above all else, it is meant to be practical.

2.0 JUDGMENTAL ESTIMATION PROCEDURES

2.1 General Requirements

Probably the most critical requirement for the use of judgmental procedures to estimate HEPs is that the events possibly producing human errors being considered be defined carefully and completely. The more fully the events are defined, the less they will be open to individual and variable interpretation by the experts judging their likelihood.

One of the primary factors involved in the definition of events will be performance shaping factors (PSFs). These may include a wide range of factors including stress, training, environmental factors (noise, temperature, etc.), and physical layout of controls. Swain and Guttmann (1980) provide a taxonomy of PSFs that may affect human errors in NPP operations that is reproduced here as Exhibit 2-1.

The level of detail needed in the definition of events--that is, the number and specifity of PSFs included--will vary across different uses. There clearly is a tradeoff between the generality of the HEPs estimated and the specificity of the estimates. Events that are defined with great specificity regarding their PSFs will be easier to judge, requiring relatively less subjective interpretation, and thus less variability across the judgment of multiple experts. On the other hand, HEPs estimated for events that are somewhat more generally defined, although possibly being subject to more variability, are usable in analyses of a larger number of NPP operation contexts. An example illustrating this point would arise regarding whether events should be defined and HEPs estimated for a specific NPP or for a more generic set of NPPs, say all boiling water reactor (BWR) plants. Taking the level of detail in event definitions to a probably ridiculous extreme, HEPs could be estimated for a specific operator.

The context in which the HEPs are to be used will determine the level of detail to be used. For example, if the HEPs are to be used in a PRA, the structuring of the PRA will help define the events. In any case, it is always necessary to define the events as clearly and completely as possible without adding detail that would make the events too specific for the HEPs to be useful.

Closely related to this definitional problem is the need to determine bounds on HEP estimates as well as the nominal estimates themselves. These bounds will identify the possible variability in the estimates. The variability may come from two basic sources: the events as defined may include a broad range of levels on certain PSFs (e.g., operator training), and the experts judging the HEPs may have some fundamental uncertainty regarding their likelihood no matter how well-defined the events are. Because of the first cause of variability, the more generally defined the events, the wider the range of bounds for HEP estimates.

EXHIBIT 2-1. Performance Shaping Factors

EXTERNAL		STRESSORS	INTERNAL
Situational Characteristics	**Task and Equipment Characteristics**	**Psychological Stressors**	**Organismic Factors**
Architectural features	Perceptual requirements	Suddenness of onset	Previous training/experience
Quality of environment:	Motor requirements	Duration of stress	State of current practice or skill
Temperature, humidity, and air quality	(Speed, strength, precision)	Task speed	Personality and intelligence variables
Lighting	Control-display relationships	Task load	Motivation and attitudes
Noise and vibration	Anticipatory requirements	High jeopardy risk	Emotional state
Degree of general cleanliness	Interpretation	Threats (of failure, loss of job)	Stress (mental or bodily tension)
Work hours/work breaks	Decision-making	Monotonous, degrading, or meaningless work	Knowledge of required performance standards
Availability/adequacy of special equipment, tools, and supplies	Complexity (information load)	Long, uneventful vigilance periods	Physical condition
Manning parameters	Narrowness of task	Conflicts of motives about job performance	Attitudes based on influence of family and other outside persons or agencies
Organizational structure (e.g., authority, responsibility, communication channels)	Frequency and repetitiveness	Reinforcement absent or negative	Group identification
Actions by supervisors, coworkers, union representatives, and regulatory personnel	Task criticality	Sensory deprivation	
Rewards, recognition, benefits	Long- and short-term memory	Distractions (noise, glare, movement, flicker, color)	
Job and Task Instuctions	Calculational requirements	Inconsistent cueing	
Procedures required (written or not written)	Feedback (knowledge of results)	**Physiological Stressors**	
Written or oral communications	Continuity (discrete vs continuous)	Duration of stress	
Cautions and warnings	Team structure and communication	Fatigue	
Work methods	Man-machine interface factors:	Pain or discomfort	
Plant policies and shop practices	Design of prime equipment, test equipment, manufacturing equipment, job aids, tools, fixtures	Hunger or thirst	
		Temperature extremes	
		Radiation	
		G-force extremes	
		Atmospheric pressure extremes	
		Oxygen insufficiency	
		Vibration	
		Movement constriction	
		Lack of physical exercise	

HEP estimates should not be used without careful consideration of accompanying uncertainty bounds. These bounds will provide needed information. For example, wide uncertainty bounds imply a certain lack of consensus among the experts regarding the estimates of HEPs. Conclusions drawn from HEP estimates should be based on uncertainty bounds and sensitivity analyses as well as the HEP estimates themselves.

In addition, it is important before using the results of any of the procedures to check the consistency of judgments across experts. While complete agreement is clearly an unreasonable goal (in fact some of the techniques require some variability across experts), too much inconsistency, indicating lack of agreement among the experts, will suggest that results may be of questionable validity. Thus, in this report we discuss procedures for consistency checks, as well as methods for obtaining uncertainty bounds on estimates (that may also be related to consistency across experts).

Finally, in order to use any of the procedures discussed here, appropriate experts must be available to make the required judgments. Later in this section, specific requirements (e.g., the number) regarding experts for each of the techniques are discussed. Here, the focus is on general requirements including what types and mixes of experts to use. Expertise of the following types may be useful:

- human factors,
- NPP operators,
- NPP supervisory personnel,
- nuclear and system engineers (preferably with some exposure to human engineering), and
- human factors, operating, supervisory, and system engineering knowledge of non-NPP contexts that are similar, e.g., chemical processing plants or military weapons systems.

To the extent possible, the widest range of expertise should be used for the application of any of the procedures. It is also important, however, to ensure that particularly critical expertise be used. That is, it is probably not advantageous to reduce the number of human factor experts with extensive experience in NPP operations in order to add experts on human factors in military weapons systems.

2.2 Overview of Procedures

In this subsection, we describe briefly the five procedures that are considered here. Details of their implementation are provided in Appendix A. Also note that all evaluations of these procedures are reserved for the following subsections.

2.2.1 Paired comparison procedure. A detailed description of the paired comparison procedure is given in Section A.2 of Appendix A. This procedure requires judgments of the type "event i is more likely than event j." Using judgments of this type from several experts for a set of events and Thurstone's (1927) Law of Comparative Judgment, an interval scale of likelihood can be derived (Torgerson, 1958). This model assumes that each event likelihood is represented by a distribution of subjective magnitude, and that the distribution is normal. In comparing the likelihood of two events, a magnitude is selected randomly from each distribution and the event with the higher subjective magnitude is reported to be more likely. Taken across experts, the proportion of times an event i is judged to be more likely than event j is then transformed into a normal deviate. The average of all such normal deviates for a particular event in comparison with all other events is then taken as the scale value of likelihood for that event.

In the general case, all pairs of k events are judged, requiring k(k-1)/2 judgments from each judge. There are procedures that can be used to reduce the total number of paired comparison judgments any expert must make. If many experts are accessible, each expert need not make all possible comparisons. With a total of n judges and if a minimum of m judgments per pair are needed for a reliable scale, each judge must make only 100m/n percent of all possible paired comparisons. The pairs judged by the experts must be appropriately counterbalanced to minimize biases. One counter-balancing approach would be to have each judge i, i=1,...n, judge pairs (i-1)t/n+1 through (i-1)t/n+mt/n, where t is the total number of possible paired comparisons. Note that in this formula, pair t+1, is pair 1, etc. For example, with four experts (n=4), a requirement for two judgments per pair (m=2), and nine events (k=9), thus making t=k(k-1)/2=36; expert one would judge pairs one ((i-1)t/n+1=1) through 18 ((i-1)t/n+mt/n=18). Expert two would judge pairs 10 through 27; expert three, pairs 19 through 36; and expert four, pairs 28 through 9 (28 to 36 and 1 to 9).

Other procedures allow all experts to make all required paired comparisons, but reduce the number of paired comparisons needed. One method is to use a subset of the events as standards. Each event is then compared only with all standards. The standards should be selected to be spread across the range of event probabilities.

A second procedure, requiring a rough initial ordering, involves dividing the total set of events into overlapping subsets of events. Within a subset, each event is compared against all others and the usual procedure is used to derive scale values within each subset.

The overlapping events are then used to create a joint scale of all events. Each subset must include at least two events in common with another subset. The two common events can be used to transform the scale values of one subset into scale values which are on the same scale as those of the second subset. For example, for two subsets X and Y with events e and f in common, the equations

$$s_e^Y = as_e^X + b$$

and

$$s_f^Y = as_f^X + b,$$

where s_e^Y, s_e^X, s_f^Y, and s_f^X are the scale values for event e in subset Y, event e in subset X, event f in subset Y, and event f in subset X, respectively; would be solved for a and b. Then all scale values in subset X would be multiplied by a and added to b to obtain new scale values which would be on the same scale as events in Y. If subsets have more than two events in common, least-squares procedures can be used to solve for a and b.

A similar type of procedure can be used with non-overlapping subsets and a set of standard events. In effect, the standard events become the overlapping stimuli among subsets.

This scale must then be transformed into a probability scale by assuming some fixed relationship between the scale values and probabilities. Previous research (Hunns and Daniels, undated; Pontecorvo, 1966) suggests that a logarithmic relationship is most appropriate, i.e.,

$$\log p_i = as_i + b$$

where p_i is the estimated probability of event i, s_i is the scale value of event i, and a and b are constants. In order to determine a and b, the probabilities of at least two events must be known or estimated independently. If possible, more than two events should be used to increase the reliability of the calibration. In general, if two events are used, they should be near the upper and lower ends of the scale.

If two human errors with known or validly estimated probabilities cannot be included in the set of events being considered, two approaches to determining these constants are possible. One is to include two events of a different type, e.g., death from different causes, with known probabilities. This may require judgments that are difficult to make about events on which the judges are not particularly expert thus leading to questionable calibration. The other is to have experts make direct judgmental estimates of two HEPs. At this point, there is no research bearing directly on which of these procedures is more appropriate. Our recommendation is to use the latter method, with carefully selected human errors. The errors used should be those that the experts can best estimate, and those for which evidence regarding their probabilities is best, e.g., from Swain and Guttmann (1980).

Another relationship that might be appropriate rather than the logarithmic relationship is a power relationship (Stevens, 1975):

$$p_i = \exp((\log s_i - \log a)/b).$$

It, however, has not been studied in the context of relating interval scale values of likelihood to probabilities, so until further research can be performed, we recommend use of the logarithmic relationship.

In addition to the rationale underlying the Law of Comparative Judgment, and its widespread use in psychological scaling; a major justification for the use of the paired comparison procedure is the argument that people are much better able to make qualitative, relative judgments of the type required than they are to make numerical estimates. This seems to be particularly true for extreme numerical estimates such as the very small probabilities estimated for most HEPs. For example, Lichtenstein et al., (1978) found that subjects were generally able to judge correctly which of two events was more likely (if one event was at least twice as likely as the other), but were not very good at estimating numerically the relative likelihood of the two events. Thus, if the qualitative information in paired comparisons can be used to derive probability estimates without requiring invalid additional assumptions, the HEPs produced will have a sound judgmental basis.

2.2.2 Ranking or rating procedure. Ranking and rating have been included as a single procedure because, although the judgments required are somewhat different, the psychological model underlying the development of HEPs from ranking or rating judgments is the same. A complete description of this procedure is given in Section A.3 of Appendix A.

Ranking requires each expert to rank order the set of human errors according to their likelihood of occurrence. The rating procedure requires each expert to judge each event on a given scale, e.g., from one to ten. (Results are generally insensitive to exact form of the rating scale, Bock and Jones, 1968.) These techniques can be considered together as one procedure because, in effect, each rank can be considered a different rating.

The underlying psychological model for producing scales of likelihood from ranking/rating judgments is the Law of Categorical Judgment (Torgerson, 1958), which is based on psychological principals similar to those for the Law of Comparative Judgment. Again, the likelihood of each event is assumed to be represented by a normal distribution of subjective magnitude. Category boundaries, i.e., the boundaries between different rankings or different rating categories, are also assumed to produce a normal distribution of subjective magnitude. Then both events and boundaries are scaled using a procedure similar to that for paired comparisons.

As with paired comparisons, the result of applying the Law of Categorical Judgment is an interval scale of likelihood. This must then be transformed into a probability scale using the same methods as described for paired comparisons.

Ranking and rating judgments, as well as paired comparisons, are considered to be less susceptible to bias than are numerical judgments. As discussed below the choice among these procedures will usually be made on practical grounds.

2.2.3 Direct numerical estimation. The direct estimation procedure (described in Section A.4 of Appendix A) requires the experts to provide estimates of the likelihood of the human errors. The estimates may be either in probabilities or in odds, and the response mode may vary. Responses may be written numbers, marks on a scale, or any of a number of other responses (Stillwell et al., 1982). Odds responses on a logarithmicly-spaced scale appears to be the best direct estimation procedure, particularly for relatively unlikely events as many human errors are.

When HEPs are estimated using the direct estimation procedure, some means of aggregating estimates across experts is needed to produce a single probability estimate for each human error. For paired comparisons and ranking/rating the underlying model specifies how this aggregation is to be accomplished. For direct estimation, this is not the case, so any of a number of procedures can be used ranging from Delphi to simple averaging (Stillwell et al., 1982). Since this same question of how to aggregate across experts also arises for indirect numerical estimation and for multiattribute utility measurement, a detailed discussion of this problem is presented following the discussion of these two procedures.

2.2.4 Indirect numerical estimation. As described in detail in Section A.5, Appendix A, this procedure requires each expert to compare pairs of events and to make ratio judgments as to how much more likely one event is than the other. Each event must be compared with at least one other event, such that all events are linked. For example, with four events, a, b, c, and d, a would be compared with b, b with c, and c with d. Thus, with n events, n-1 judgments are needed.

In order to convert these ratios into probabilities, the probability of one event must be known or estimated independently. Procedures for obtaining this probability are the same as those discussed for paired comparisons. Other probabilities are then calculated by applying the appropriate ratios to this probability. As with direct estimation, some procedure is required to aggregate estimates across experts. This is discussed below.

2.2.5 Multiattribute utility measurement. As was noted in the introduction, this procedure is still in an initial developmental stage, and thus is not recommended for use until additional research documents its validity. It is, however, an interesting possibility, and preliminary studies by Embrey (1981a, 1981b) suggest this procedure may be useful. Several variations are possible, with the procedure described in Section

A.6 of Appendix A based on Embrey's work appearing to be the most promising.

As used by Embrey, the procedure is based on methods advocated by Edwards (1977). The first step in the procedure is to identify the PSFs that are most relevant to the events being considered. The choice of this set of PSFs is based on a consensus of experts making the judgments. Edwards (1977) argues that the number of PSFs should not be more than 11, and that seven is a reasonable number to consider. These PSFs are then weighted regarding their relevancy to producing errors for the set of events under consideration. These weights are not general in that they would apply to all sets of events. The weights are assessed judgmentally by the experts using a procedure described in the Appendix.

Then each event must be rated with regard to how much each PSF degrades or enhances the likelihood of an error. Numerical ratings on a 0 to 100 scale are used. A rating of 0 indicates that the particular PSF has a maximally degrading effect for that event, while a rating of 100 implies a maximum enhancing effect. For example, if the PSF is the amount of feedback received by an operator, a rating of 0 might indicate misleading or incorrect feedback, 50 would indicate no feedback, and 100 would be the provision of accurate, complete, and timely feedback. It is important to have descriptive anchors for these scales to maintain consistency across experts.

An index of human error likelihood is then derived using a weighted additive model:

$$S_i = \sum_{j=1}^{n} NW_j X_{ij},$$

where S_i is the likelihood scale value of event i, NW_j is the normalized weight for PSF_j, and X_{ij} is the rating of the effect of PSF_j on event i. Since the likelihood scale produced is again an interval scale, it must be transformed into probabilities using the methods discussed for paired comparisons.

Aggregation of estimates across experts can be undertaken at either of two points: as individual weights and ratings are assessed, or after probabilities are derived using each expert's individual assessments. Our recommendation regarding where aggregation should take place depends upon the procedure used to aggregate (see discussion below). If interaction among the experts during the estimation process does not occur, aggregation should take place after probabilities have been derived individually for each expert. If, however, there is an opportunity for interaction and discussion among the experts, the aggregation should occur for each weight and rating judgment.

2.2.6 Aggregating individual judgments. As noted above, for the direct numerical, indirect numerical, and multiattribute utility procedures,

some method is needed to aggregate the HEP estimates of individual experts into a single estimate. Two methods are advocated, with the choice between them depending primarily upon whether or not the experts are physically together to discuss information and interact while making judgments. (Additional considerations regarding this choice are discussed in the following subsections.)

One method is simply to combine mathematically the individual estimates. This can, of course, be used when experts have no opportunity to interact directly. We recommend this method, as opposed to a Delphi method involving feedback and multiple estimates, because research has shown that the additional effort required for the Delphi method does not increase the validity of estimates (Seaver, 1976, 1978).

The aggregation rule we suggest is:

$$P_j = \frac{\left(\prod_{i=1}^{m} P_{ij}\right)^{\frac{1}{m}}}{\left(\prod_{i=1}^{m} (1-P_{ij})\right)^{\frac{1}{m}} + \left(\prod_{i=1}^{m} P_{ij}\right)^{\frac{1}{m}}},$$

where P_j is the aggregated probability assigned to event j, P_{ij}, is the probability estimated for event j by expert i, and m is the number of experts. This rule is appropriate only for binary events, i.e., those which either occur or do not occur. Bordley (1982) provides theoretical justification for this rule, and Seaver (1978) has shown this to be a good aggregation rule with empirical results.

The second method is based on the Nominal Group Technique (NGT) (Delbecq et al., 1975) that guides and controls interaction among the experts. It is meant to allow for the exchange of ideas and information while attempting to control for extraneous influences that may affect judgments such as pressure for conformity, or domination by certain experts because of personality. The NGT, however, because it does allow some discussion among experts may not completely control these extraneous influences. Typical use of the NGT includes the following steps.

1. Each expert makes a private judgment in the presence of all experts without discussion.

2. Again without discussion, each expert's judgment is presented to all experts.

3. Judgments are discussed for clarification and evaluation under the control of a discussion leader who is responsible for preventing dominance and focusing on relevant issues.

4. Each expert then reconsiders his/her judgment without further discussion.

5. These final judgments are combined using a mathematical rule.

For direct and indirect numerical procedures, these steps would be followed for each event and the judgments would be those required by the procedure. The aggregation rule would be the rule given above.

For the multiattribute utility procedure, these steps would be followed once for judgments of weights. That is, each expert would make all weight judgments (step 1) and then proceed through steps 2-4 considering all weights at each step. The aggregation rule for weights is to take the geometric mean of the (unnormalized) individual judgments, i.e.

$$W_j = \left(\prod_{i=1}^{m} W_{ij} \right)^{\frac{1}{m}} ,$$

where W_j is the (unnormalized) weight for PSF_j, W_{ij} is the (unnormalized) weight assigned by expert i to PSF_j, and m is the number of experts.

The rating of each event on each PSF should also follow these five steps. In this case, all steps should be carried out for the rating of one event on one PSF before proceeding to the next rating. The aggregation rule used for ratings should be a simple average of the experts' ratings.

2.3 Evaluation of Procedures

The procedures described above represent a set of procedures from which one procedure may be selected for any specific application. Which procedure should be selected will depend upon what use is to be made of the HEPS, the values and preferences of the person (people) obtaining the estimates, and specific situational constraints. These latter constraints are discussed in detail in the following subsection. Here the focus is on a general evaluation of the strengths and weaknesses of the procedures.

The following criteria have been identified as important for the evaluation of the procedures:

- quality of the judgments required,
- difficulty of data collection,
- empirical support for procedure,
- acceptability to experts making judgments,
- theoretical justification, and

- data processing requirements.

Exhibit 2-2 presents a rank ordering of the five procedures on each of these criteria, based on review of the use of these procedures and our experience with them. Below, the basis of these rankings on each criterion is discussed. It also ranks the two procedures for aggregating individual estimates that are required for the direct numerical, indirect numerical, and multiattribute utility procedures.

Regarding Exhibit 2-2, several points must be made. First, it should serve only as a guide to be used in conjunction with good professional judgment, not as a decision table. (For example, one should not necessarily use the procedure with the lowest sum of rank orders.) Second, in using this exhibit as a guide, the user should consider the relative importance of the criteria. In most cases the "quality of judgment" and "empirical support" criteria may be the most important. The user should also note that because the entries are rank orders, they do not indicate how much better or worse one procedure is compared to another on the criteria. Finally, in the evaluation here, practical situational constraints are not included. These are discussed in Section 2.4 along with their implications for selecting a procedure. A user should select the best procedure possible based on the information in Exhibit 2-2 and the following discussion within the practical constraints of the situation.

2.3.1 Quality of judgments. This criterion refers to accuracy of the basic judgments required by the procedure. It is important to keep in mind here that the HEPs being estimated will generally be quite small. Were they larger, the direct numerical procedure would not be ranked so low. Its ranking here reflects that people generally have great difficulty in assigning very small probabilities. For less extreme probabilities (e.g., .1 to .9), direct estimates can be expected to be reasonably accurate.

There is a considerable body of research showing people make better relative, indirect numerical judgments (i.e., likelihood ratios) than direct estimates (e.g., Lichtenstein et al., 1978; see also Stillwell et al., 1982 for a review). People are even better at non-numerical judgments such as those required for paired comparisons, rankings, or ratings. Among these types of judgments, paired comparisons seem to produce higher quality judgments because they require a comparison between only two events rather than simultaneous consideration of more than two events (ranking) or comparison of an event with multiple categories (rating).

As noted earlier, the multiattribute utility procedure has not previously been used to estimate probabilities so little is known about the quality of these judgments. It has been used extensively on other problems, but very little attention has been given to validity studies.

EXHIBIT 2-2. Rank Order of Procedures on Each Evaluation Criterion[1]

Evaluation Criteria	Procedure					Aggregation[2]	
	Paired Comparison	Ranking/Rating	Direct Numerical	Indirect Numerical	Multiattribute Utility	Mathematical Aggregation	Nominal Group Technique
Quality of Judgment	1	2	4	3	?	1	1
Difficulty of Data Collection	4	1	2	3	5	1	2
Empirical Support	3	4	1	1	5	1	1
Acceptability to Experts	1	2	5	4	3	2	1
Theoretical Justification	1	1	4	4	3	2	1
Data Processing	5	4	1	2	3	1	1

NOTE: Care must be taken in interpreting the numbers in this table. They indicate only rank orders and should not be interpreted as absolute ratings. That is "1" means only that the particular procedure is better on that criterion than other procedures. It does not imply any judgment regarding the absolute quality of the procedure. All numbers should be interpreted only in conjunction with discussion in accompanying text.

[1]Ranks from 1 to 5 are from best to worst.

[2]Applies only for direct numerical, indirect numerical, and multiattribute utility procedures.

With respect to the two aggregation methods, research has shown that they produce very similar results with no significant differences in quality (Seaver, 1978; Stillwell et al., 1982).

2.3.2 Difficulty of data collection. This criterion refers to such factors as the time and effort required of both the experts and the people obtaining the estimates. It may involve designing the data collection process, developing appropriate response forms, training experts in the judgments required, and making the judgments.

The multiattribute utility procedure appears to be the most time consuming and difficult to implement. It requires the experts to define the PSFs to be included in the assessments. Each expert must then make m-1 weight judgments and m x n ratings of events with respect to the PSFs where m is the number of PSFs included and n is the number of events. In addition, the judgments required (weights and ratings) are complicated ones, so considerable training and/or instructions will be required to elicit these judgments.

Less complicated judgments are required for paired comparisons, ranking, or rating procedures so training time and instructions will be less. A large number of judgments, however, is required for the paired comparison procedure. Ranking/rating, direct numerical, and indirect numerical procedures each require approximately one judgment per event, although ranking may be somewhat more difficult than rating in this respect. For numerical judgments some training and/or instructions should be provided regarding biases that typically occur in such judgments (Stillwell et al., 1982).

Use of the interactive NGT is clearly a much more difficult data collection technique. All the experts must be assembled in one place, and they are then allowed to exchange information and discuss judgments which will require considerable time. It will also require the time of someone involved in obtaining the estimates to lead the group.

2.3.3 Empirical support. Ranking on this criterion is based on the extent to which the procedure has been empirically tested as a means of estimating probabilities. The multiattribute utility procedure has very little empirical support. Embrey (1981b) describes a small-scale, preliminary test of the methodology that produced some promising results in terms of the agreement between probabilities obtained with the procedure and known probabilities. These results are somewhat controversial, however, because of a low degree of interjudge consistency and because the probabilities being estimated were not extreme as HEPs are likely to be.

Although there has been considerable empirical support for the use of paired comparisons and ranking or rating as scaling techniques, they have not received much attention as procedures for estimating probabilities. Blanchard et al., (1966) and Rigby and Edelman (1968) have used

paired comparisons in human reliability analysis, and Munns and Daniels (undated) used the procedure to estimate the probability of a train disaster. Some unpublished work by Stillwell and Seaver used paired comparison information to estimate the probability of death from various causes and obtained a correlation of .75 between estimated probabilities and relative frequencies.

A particularly weak link in both procedures is the transformation of the scale values into probabilities. Although a logarithmic relationship has been suggested (Munns and Daniels, undated; Pontecorvo, 1966), empirical support is weak.

Direct and indirect numerical procedures are both supported by an extensive amount of research and practical application (see Stillwell et al., 1982). Little of this support, however, reflects estimation of very small probabilities.

For the three procedures requiring some type of aggregation, both mathematical aggregation and the NGT have considerable support (e.g., Seaver, 1976, 1978; Stillwell et al., 1982). The NGT method also has been used extensively in problem solving tasks where probability estimates or other numerical judgments are not required (e.g., Delbecq et al, 1975).

2.3.4 Acceptability to experts. The acceptability of the procedure to the experts providing judgments is important because if the resulting HEP estimates are to be used, they must have the support of the experts. This support will occur only if the experts are satisfied with the way in which the HEPs were estimated. The procedure must also be accepted by users of the HEPs, e.g., PRA specialists, and by regulators (e.g., NRC) who make decisions based on this information.

The numerical judgment procedures may be the least acceptable because experts may feel that an accurate probability or relative likelihood judgment cannot be made (even though research has shown that in many cases they can be). The direct numerical judgment procedure is particularly susceptable to this criticism because of the extremeness of the HEPs.

The multiattribute utility procedure has considerable face validity, primarily because it deals explicitly with PSFs that human factors experts will feel are very important considerations in estimating HEPs. In this procedure, however, the experts may be somewhat uncomfortable with the unusual nature of the judgments required.

Ranking, rating, and particularly paired comparisons are likely to be more acceptable because the judgments are relatively simple ones, and the experts may be expected to believe they can make such judgments relatively well. Questions of acceptability may arise, however, because the experts do not understand how these judgments are used to derive

HEPs. These questions may be partially alleviated by a careful explanation of the procedures.

With respect to aggregation methods, this criterion is very important. Experts are much more likely to accept a procedure such as the NGT in which they have a chance to interact and discuss information with colleagues, so they can consider more information and can understand reasons for differences in judgments. Social psychological research has shown that involvement in a group produces a strong feeling of responsibility for and acceptance of group products and decisions (Seaver, 1976).

2.3.5 Theoretical justification. This criterion refers to the extent to which a formal model underlies the procedure. There is no particular theoretical justification for the numerical procedures: their justification is derived empirically. Multiattribute utility measurement does have a very strong theoretical basis (Keeney and Raiffa, 1976), but not as a probability estimation procedure. In addition, there is no underlying psychological theory for the types of judgments required.

On the other hand, both paired comparisons and ranking/rating are based on well-established psychological theories of judgment (Torgerson, 1958). But these two procedures also suffer from a lack of justification as probability estimation procedures. Again the weak link theoretically is the transformation from scale values to probabilities, which must be justified empirically.

The NGT was developed on an extensive review of psychological theory regarding group processes (Delbecq et al., 1975). While the mathematical models used for aggregation (in the NGT as well as with no interaction) are based on a theoretical model, they have been simplified to be practical with empirical support for the simplification (Seaver, 1978).

2.3.6 Data processing requirements. This criterion refers to the amount of analysis required after the experts provide judgments to produce HEPs. As can be seen from Appendix A, none of the procedures require difficult processing. All processing can be performed by hand or relatively easily with a calculator. Thus, this criterion is probably less important than the others.

The paired comparison and ranking/rating procedures require the most processing. Paired comparisons require slightly more because of the necessary collation of the paired comparison judgments. Other processing required is simply averaging and translating scale values into probabilities.

The multiattribute utility procedure uses a weighted additive model to derive scale values, that are then translated into probabilities. It

also requires aggregation of estimates across experts. In the indirect numerical procedure, probabilities are calculated by simple multiplication and are aggregated across experts. The direct numerical procedure requires only aggregation across experts.

Both aggregation procedures use the same aggregation rules, except if the NGT is used with the multiattribute utility procedure. Then more aggregations must be performed.

2.4 Situational Constraints

Many factors in the specific context in which HEPs are to be obtained may influence the selection of a particular estimation procedure. These factors include:

- number of experts available,
- number of HEPs to be estimated,
- time available to produce estimates,
- type of experts available,
- physical location of experts,
- specificity of the errors considered,
- similarity of errors considered,
- order of magnitude of the errors,
- availability of independent estimates of some HEPS, and
- resources available.

The number of experts available is particularly critical, since the paired comparison and ranking/rating procedures require relatively more experts. Appendix B discusses the implications of the number of experts used for reliability of and uncertainty bounds for HEPs estimated by the paired comparison procedure. Because the underlying model is much the same, these implications also apply to the ranking/rating procedure. The results in Appendix B suggest that reasonable reliability can be obtained with as few as eight experts, or even fewer in some circumstances. However, if only smaller numbers of experts can make the judgments, paired comparison and ranking/rating procedures may not be sufficiently reliable.

As also noted in Appendix B, the number of HEPs to be estimated affects the reliability of estimates derived from paired comparisons (and ranking/rating as well). However, for paired comparisons, this must trade off against the number of judgments required which increases rapidly as the number of events considered increases.

Although there are some techniques for reducing the number of judgments required (discussed previously with a specific example in Appendix A), estimation of a large number of HEPs using paired comparisons will take either considerable time from each expert or a large number of experts. If a large number of HEPs are needed and there are limitations on time and the number of experts available, procedures other than paired comparisons may be more efficient.

The degree to which the time available to make the estimates constrains selection of a procedure obviously is dependent upon the number of experts available and the number of HEPs to be estimated. The more experts and fewer HEPs to be estimated, the less important time constraints are. To the extent that time does constrain the estimation process, direct and indirect numerical and ranking/rating procedures will require the least time.

The type of experts available, particularly with respect to understanding and being comfortable with the concepts of probability, can also affect procedure selection. Less probabilistically sophisticated experts will require considerable training for the direct and indirect numerical procedures (Stillwell et al., 1982). For such experts, the non-numerical procedures are to be preferred.

The physical location of experts will affect whether or not it is feasible to get them together to interact. For the paired comparison and ranking/rating procedures, there is no particular advantage to having the experts together. For the other procedures, however, if the NGT is used, the experts must be together physically, which may increase the time and costs of data collection and/or constrain which experts can participate.

The numerical procedures will generally require more specifically defined events because they require a very precise numerical judgment. This is also true of the multiattribute utility procedure. The paired comparison and ranking/rating procedures, however, do not require precise judgments. It is possible, for example, to judge which of two generally defined errors is more likely even though, because of their generality, the exact probability of either error could not be estimated.

The similarity of errors considered is of particular relevance for paired comparisons. If paired errors are too dissimilar, judgments may

be very difficult. For example, experts may have a difficult time comparing a judgmental error with an action error. If paired comparions are to be used, some grouping of errors by similarity may be useful.

On the other hand, errors that are defined the same way except for one specific change, e.g., a change in stress, should not be paired. The underlying assumptions of the psychological model are unlikely to be met in such a case, and thus may detrimentally affect results. For example, if the same error is considered under normal and high stress, all experts may indicate that the high stress error is more likely. This certainty does not necessarily reflect a belief that the likelihood of the error under high stress is very much higher than under normal stress. The actual probabilities could be quite close. The paired comparison procedure, however, could produce a large difference in probabilities.

In most cases, in estimating HEPs at least some can be expected to be quite small. To the extent this is true, it may lead direct numerical judgments to be somewhat biased, although the procedures described for direct numerical estimation in Appendix A are designed to minimize bias. If extreme HEPs are not expected, the direct numerical procedure becomes relatively more attractive.

All procedures except the direct numerical procedure require an independent estimate of at least one HEP. The indirect numerical procedure requires one such estimate while the other procedures require two. The more confidence that can be placed in independent estimates, the better these procedures are.

Lack of resources may constrain procedure selection through the incapability of obtaining the services of a needed number of experts for the necessary time. A very tight budget may eliminate paired comparison and ranking/rating procedures that require more experts and/or their time. If direct numerical, indirect numerical, or multiattribute utility procedures are used, resource constraints may force the use of mathematical aggregation rather than the NGT.

For the most part, the constraints discussed above tend to drive procedure selection away from paired comparisons and to some extent ranking/rating. On the other hand, without constraints, the general evaluation in the preceding subsection tends to favor these procedures. Thus, practical choice of a procedure will usually be a matter of the degree and firmness of the constraints. Very severe and firm constraints will usually lead to selection of the direct or indirect numerical procedures and use of mathematical aggregation. Less severe constraints will allow use of paired comparisons or ranking/rating, which are generally more attractive.

3.0 DISCUSSION

The spirit in which the procedures described in this report for estimating HEPs in NPP operations are offered is much the same as that underlying advocacy of the use of expert judgment to estimate HEPs: make the best use possible of current procedures and information, even though they may be flawed. These procedures certainly are not well-tested and validated procedures for estimating HEPs. They are, however, based on sound psychological theory and empirical support that indicate their potential usefulness. And as a practical matter, they can be used now to obtain needed HEPs.

The selection of which procedure to use is probably less important than the decision to use expert judgment to estimate needed HEPs. These five procedures have all been screened and developed to be reasonably effective and valid, with the exception of the multiattribute utility procedure that has not been investigated thoroughly. Our recommendation is that this procedure not be used until it is further tested. This should not be interpreted as a negative critique: the fact that it is discussed here is a positive appraisal. Rather, we simply advocate a wait and see attitude.

Among the procedures described, based on available evidence the paired comparison and ranking/rating procedures appear to have the potential to provide the best HEP estimates. The paired comparison procedure in particular has been used to a small degree in human reliability analysis (Blanchard et al., 1966; Rigby and Edelman, 1968) as well as extensively as a psychological scaling technique. It however, generally presents the most practical problems in application.

Some practical constraints can be alleviated by using the ranking/rating procedure that is based on the same underlying psychological model of judgment as the paired comparison procedure. If practical constraints are severe, either of the numerical estimation procedures can be used, although under most circumstances, the indirect numerical procedure is recommended. These procedures can be used to obtain HEPs relatively quickly, with a few experts, and with low costs. These practical savings, however, can be expected to result in some decrease in the quality of the HEP estimates obtained.

Two parts of the implementation of any of these procedures will probably have more effect on the results than will the selection of which procedure to use. It is critical that the "right" experts be selected and that the "right" events be considered. In Section 2.1, these topics were discussed, so here only a few important points are made. In selecting experts, the best available people should, of course, be used. But it is also necessary to use experts who have an open mind about the use of expert judgment to estimate HEPs, and are willing to put their exper-

tise on the line in their judgments. Experts who do not believe that judgment can be used to estimate HEPs are unlikely to give their judgments sufficient consideration because they do not think the judgments are sufficiently valid to estimate HEPs.

To some extent, the "right" events will be defined outside the context of estimating HEPs using expert judgment. For example, they may be defined by a PRA. What is important, is the clarity and completeness of the definition of the events. Good definitions including consideration of PSFs will simplify judgments and improve their reliability and validity.

Although the procedures described here can be used now to estimate HEPs, they undoubtedly can be improved through additional testing. Such testing would also increase the confidence of users in the results of analyses using judgmentally-estimated HEPs. There are several specific areas in which research is needed.

For the paired comparison, ranking/rating and multiattribute utility procedures, scale values must be transformed into probabilities. Currently, the relationship between scale values and probabilities is thought to be best-defined as logarithmic. This relationship, however, has relatively little empirical support, so additional research should investigate the extent to which it is logarithmic across events and experts.

The transformation of scale values into probabilities also requires independent estimates of at least two HEPs for paired comparison and ranking/rating and one HEP for indirect numerical estimation. If such estimates for human errors are not available, the independent estimates can be obtained either by direct numerical judgment or by including two events of a different type with known probabilities (though the latter approach would require empirical support before use). The effects of including two events of a different type in the set of human errors should be investigated for different types of events.

In addition to these specific research topics, more general research is needed investigating the reliability and validity of the procedures for probabilities generally and HEPs specifically using real events and experts. In particular such research should determine the extent to which the procedures produce different estimates. Even though there is at present some qualitative indication of differences (which of course may be wrong), quantification of differences would be helpful in establishing tradeoffs between the quality of estimates and practical considerations (time, cost, number of experts, etc.). Not only should the research quantify the differences, but it should also, to the extent possible, determine which estimates are more valid and the circumstances under which different procedures produce more valid estimates.

Thus, we conclude this report with two recommendations. (1) The procedures described herein should be used now to estimate HEPs as they are needed. (2) A program of research should be undertaken to test further and validate the use of these procedures.

4.0 REFERENCES

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

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, CA: Holden-Day, 1968.

Bordley, R.F. A multiplicative formula for aggregating probability assessments. Management Science, 1982, 28, 1137-1148.

Breiman, L. Probability. Reading, MA: Addison-Wesley, 1968.

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

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

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

Embrey, D.E. A new approach to the evaluation and quantification of human reliability in systems assessment. Proceedings of Third National Reliability Conference - Reliability 81, 1981, Birmingham, England, 5B/1/1 - 5B/1/12. a

Embrey, D.E. The use of performance shaping factors and quantified expert judgment in the evaluation of human reliability: An initial appraisal. Skelmersdale, England: Human Reliability Associates, Research Report No. HR-BNL-2, September 1981. b

Goodman, B., Saltzman, M., Edwards, W., and Krantz, D.H. Prediction of bids for two-outcome gambles in a casino setting. Organizational Behavior and Human Performance, 1979, 24(3), 382-399.

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

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

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.

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

Kemeny, J.G. Reports of the technical assessment task force (Volume 1). The President's Commission on the Accident at Three Mile Island, Washington, D.C.: Government Printing Office, 1979.

Kendall, M.G. Rank correlation methods. New York: Hafner Publishing Company, Inc., 1955.

Kendall, M.G., and Babington Smith, B. On the method of paired comparisons. Biometrika, 1940, 31, 324-345.

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.

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.

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

Pontecorvo, M.M. A method of predicting human reliability. Annals of Reliability and Maintenance, 1966, 4, 337-342.

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

Rogovin, M., and Frampton, G.I., Jr. Three Mile Island (Volume 1). Washington, D.C.: Nuclear Regulatory Commission Special Inquiry Group, NUREG/CR-1250, January 1980.

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.

Siegel, S. Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill, 1956.

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. Psychophysics: Introduction to its perceptual, neural, and social prospects. New York: John Wiley and Sons, 1975.

Stillwell, W.G., Seaver, D.A., and Schwartz, J.P. Expert estimation of human error probabilities in nuclear power plant operations: A review of probability assessment and scaling (NUREG/CR-2255, SAND81-7140). Albuquerque, NM: Sandia National Laboratories, May 1982.

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.

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

Tversky, A., and Kahneman, D. Judgments under uncertainty: Heuristics and biases. Science, 1974, 185, 1124-1131.

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

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.

APPENDIX A
IMPLEMENTATION OF ESTIMATION PROCEDURES

In this appendix, the steps for implementing the five procedures are described in detail including examples and necessary calculations. People interested in using these procedures can use this appendix as a guide. The appendix is composed of a brief discussion of preparation for implementing any of the procedures (Section A.1) which should be reviewed along with Section 2.1 of the main body of this report. Then Sections A.2, A.3, A.4, A.5, and A.6 describe the paired comparison, ranking/rating, direct numerical, indirect numerical, and multiattribute utility procedures respectively. Each of these sections consists of subsections regarding

- judgments required,
- calculations required to produce HEPs,
- procedures for determining interjudge consistency, and
- procedures for estimating uncertainty bounds.

Section A.2 on paired comparisons contains an additional subsection on procedures for determining within-judge consistency. Each procedure is introduced by a summary of steps to be followed in developing desired HEP data.

A.1 Preparation

The most important ingredient in any of the estimation procedures to be discussed in these appendices is that the events about which the judgments are to be made be carefully and accurately defined. There are many approaches to event definition but the best include the following elements

- The role of PSFs in the event should be defined.
- Events not under consideration, but which might be confused with the event to be judged should be clearly separated from the event under consideration.
- Larger sets of events to which this event belongs should be described.
- Causes of the event, e.g., sets of mutually exclusive initiating events and sequences of events, should be identified.

A.2 Paired Comparison Procedure

The steps required in this procedure are the following:

1. Obtain paired comparisons.
2. Create table representing the number of experts judging each event more likely than each other event.
3. Derive table of proportions from 2.
4. Create table of normal deviates corresponding to proportions in 3.
5. Calculate mean column values in table in 4 which are scale values.
6. Obtain independent estimates of two HEPs.
7. Transform scale values into HEPs.
8. Determine within-judge consistency.
9. Determine interjudge consistency.
10. Estimate uncertainty bounds.

A.2.1 Judgments required. Each expert is presented with several pairs of events and makes a discrete choice of which of each pair is the more likely. The judgment "equally likely" is not allowed. In addition, the probability of at least two of the events in the total set must be known, but the more that are known, the better. It is important that these two events be near the upper and lower ends of the range of probabilities being estimated.

The total number of experts necessary and the number of event pairs that must be judged by each expert should be based on the availability of experts relative to the time available from each expert. Appendix B gives some general guidelines on the number of experts necessary for a given level of precision in the scale values of individual events (expressed in terms of the largest expected variance possible) and shows that good precision can be obtained with few experts.

An important consideration for the use of paired comparison techniques is the large number of pairs there are for even moderate numbers of events. For example, for 20 events there are 190 ($\frac{20 \times 19}{2}$) pairs of events. Scaling experts have noted, however, that all of these judgments are not necessarily required to get good estimates of the scale

values, and several suggestions have been made for reducing the number of judgments required.

Probably the most appropriate for the judgment of human error probabilities is to select a limited number of events as standards. As much as possible, standards selected should be spaced out over the length of the scale. Each event is then compared with each standard, giving $mn-m(m+1)/2$ independent proportions where n is the number of events and m is the number of standards. For example, with 20 events, 5 of which are taken as standards, the required judgments would be reduced from 190 to 85 $((5)(20)-5(6)/2)$.

A.2.2 Performing the calculations. If a subset of events are used as standards, the computational procedures would be identical. The difference would be in the number of rows in the initial and subsequent tables. The number of standards with which all other events are compared determines the number of rows. Assuming that we have obtained a complete set of paired comparison judgments for the events whose probabilities we wish to obtain, we will now work through an example of the application of paired comparison scaling. Assume that we have obtained the matrix of judgments for 20 judges shown in Table A-1. Each cell entry represents the number of judges who said the event listed across the top was more likely than the event listed down the side. Thus, for example, the entry for cell 1,4 (row 1, column 4) will be 20 minus the entry for cell 4,1 (row 4, column 1). These example events are taken from Swain and Guttmann (1980), Table 20-17, for probability of a maintainer failing to check value status before maintenance; the frequencies in Table A-1 are hypothetical. The frequencies shown in Table A-1 are converted into proportions by dividing each frequency by the total number of experts (20 in this case). Table A-2 shows the matrix of proportions.

The next step is to convert the proportions in Table A-2 into units reflecting the assumption that they represent proportions of the area of the normal probability distribution. This conversion is accomplished by using tables of the area under the normal distribution, that can be found in most introductory statistics texts. For example, cell entry 4,1 shows that 9 of 20, or 45% of the judges stated that event 1 was more likely than event 4, while 11 of 20 or 55% said the opposite. By looking at the table of the normal distribution we find that a z value of -.13 leaves 45% of the area of the normal distribution to the left. This z value represents the relative distance between events 4 and 1. By transforming each of the proportions in Table A-2 into unit normal deviates as described above we have the matrix shown in Table A-3.

These normal deviates are summed and the mean calculated for each column as shown in Table A-3. This column mean is the value for the event on the newly created subjective scale. For example, the scale now looks like this:

TABLE A-1

Frequency Matrix of Paired Comparison Judgments*

Events: Maintainer fails to check valve status before maintenance when:	1	2	3	4	5	6
1. short list is used with checkoff, personal safety is affected	-	15	13	11	17	19
2. short list is used with checkoff, personal safety is not affected	5	-	9	5	12	16
3. short list is used without checkoff, valve is in field of view, personal safety affected	7	11	-	8	13	17
4. NPP procedures or long list used with checkoff, personal safety affected	9	15	12	-	16	18
5. NPP procedures or long list used without checkoff, valve within field of view, personal safety not affected	3	8	7	4	-	12
6. no list is used, valve within field of view, personal safety affected	1	4	3	2	8	-

*Each cell entry represents the number of experts who said the event listed across the top was more likely than the event listed down the side.

TABLE A-2

Matrix of Proportions*

Event:	1	2	3	4	5	6
1	-	.75	.65	.55	.85	.95
2	.25	-	.45	.25	.6	.8
3	.35	.55	-	.4	.65	.85
4	.45	.75	.6	-	.8	.9
5	.15	.4	.35	.2	-	.6
6	.05	.2	.15	.1	.4	-

*Each cell entry represents the proportion of experts who said the event listed across the top was more likely than the event listed down the side.

TABLE A-3

Values of Proportions Under Normal Curve*

Event:	1	2	3	4	5	6
1	-	.67	.39	.13	1.04	1.65
2	- .67	-	- .13	- .67	.25	.84
3	- .39	.13	-	- .25	.39	1.04
4	- .13	.67	.25	-	.84	1.28
5	-1.04	- .25	- .39	- .84	-	.25
6	-1.65	- .84	-1.04	-1.28	- .25	-
Scale Value (s) $= \frac{\Sigma x}{N} =$	- .64	.06	- .11	- .45	.38	.84

where N = the number of events (6 in this case).

*Each cell entry represents the normal deviate value (Z) corresponding to the proportion for the cell shown in Table A-2.

Event #	1	4	3	2	5	6
Scale value(s)	-.64	-.45	-.11	.06	.38	.84

This subjective scale of relative distances must now be converted into a scale of probabilities. In order to do this a pair of anchors is required that relate positions on the subjective scale to those on the probability scale. In most cases these anchors will come from a pair of events, placed for judgment among the others, for which the true probabilities are known. In some cases, however, none of the events in our scale will have known probabilities, and direct estimates of the anchors must be made using the direct numerical estimation procedure described in Section A.4. These direct estimates should be made after the paired judgments so that the events used for the anchor judgments are as near the ends of the true probability scale as possible.

Probabilities are assumed to be logarithmically related to the derived scale values:

$$\log HEP = as + b$$

where s is the mean scale-rank values judged by assessors, and a and b are constants, arrived at by simultaneous solution of the two variations of the above equation that result from the two anchors. In our example, we assumed anchor values were known for event #1 of .0004 and for event #6 of .01, and thus the following two equations would be solved:

$$\begin{aligned} \log(.0004) &= a(-.64) + b \\ \log(.01) &= a(.84) + b \\ \log(.0004) - \log(.01) &= -1.48a \\ -1.3979 &= -1.48a \\ a &= .94. \end{aligned}$$

Substituting a back into the first equation we get:

$$\log(.0004) = .94(-.64) + b$$
$$b = -2.7963$$

The formula:

$$\log HEP = .94s + (-2.7963)$$

now allows calculation of the probability for each of the scale values and the following scale is arrived at for the six events:

event #	1	4	3	2	5	6
log HEP	-3.3979	-3.2193	-2.8997	-2.7399	-2.4391	-2.0000
probability	.0004	.0006	.0013	.0018	.004	.01

Although it is not a factor in our example, there can be a question about how to handle complete agreement among judges about which event is the more likely. Under the normal distribution 100% or 0% of the distribution occurs at $\pm$ infinity and thus has no meaningful z value. On the other hand, this would seem to be an extremely diagnostic bit of data and therefore should not be disregarded. Appendix B (Section B.2) shows methods for selecting a z value for complete agreement, one of which results in Table A-4. We recommend that these values be used for z in cases of complete agreement with 10 or fewer judges. When the number of judges exceeds 10, procedures described in Appendix B (Section B.2) may be used to determine the appropriate z value.

A.2.3 Within-judge consistency. For paired comparison judgments, the experts may exhibit internal inconsistencies that will be shown as circular triads. That is, event a is judged to be more likely than event b; event b more likely than event c; and event c more likely than event a. The consistency of each expert can be determined using the "coefficient of consistency" (David, 1963; Kendall and Babington Smith, 1940). This coefficient varies from zero for a completely random (maximum number of circular triads) set of judgments to one for a completely consistent (no circular triads) set.

The coefficient of consistency, k, is calculated by first determining the number of circular triads, c, for the expert:

$$c = \frac{n}{24}(n^2-1) - \frac{T}{2} ,$$

where n is the number of events and $T = \sum_{i=1}^{n} (a_i-\bar{a})^2$ and $\bar{a} = (n-1)/2$. The values a_i are the number of times event a_i was judged to be more likely than any other event. For example, if one expert's judgments are described by the following matrix where a 1 indicates that the row event was judged more likely than the column event, the a_is are the row sums.

Event	1	2	3	4	5	6	a_i	$a_i-\bar{a}$
1	-	0	1	0	1	1	3	.5
2	1	-	0	0	1	1	3	.5
3	0	1	-	0	0	0	1	-1.5
4	1	1	1	-	1	0	4	1.5
5	0	0	1	0	-	0	1	-1.5
6	0	0	1	1	1	-	3	.5

In this example, $\bar{a}=(6-1)/2=2.5$, and $T=.5^2+.5^2+(-1.5)^2+1.5^2+(-1.5)^2+.5^2=7.5$. Then $c=\frac{6}{24}(6^2-1)-\frac{7.5}{2}=5$.

The coefficient of consistency, k, is then found by the formulas

TABLE A-4

Values of z for Complete Agreement*

Number of Experts	2	3	4	5	6	8	10
z	-	1.29	1.35	1.41	1.48	1.64	1.69

*Positive z values are used when the proportion is 1 and negative values are used when the proportion is 0.

$$k = 1 - \frac{24c}{n(n^2-1)}, \quad n \text{ odd},$$

$$k = 1 - \frac{24c}{n(n^2-4)}, \quad n \text{ even}.$$

In the above example, since n is even

$$k = 1 - \frac{24 \times 5}{6(6^2-4)} = 1 - \frac{120}{192} = .375.$$

Statistical tests for the coefficient of consistency are approximate and do not perform well for relatively small values of n, i.e., those in the range of practicality for paired comparison judgments in this context. For example, with six events, the usual test will reject the null hypothesis that the judgments are random only if there are no circular triads. Thus in using the coefficient of consistency, it should be interpreted as akin to a correlation coefficient. Thus, the value of .375 in the above example, while not showing a highly consistent expert, does not show sufficient inconsistency to suggest that the data of that expert should not be used.

A.2.4 Interjudge consistency. An appropriate statistic for measuring agreement with more than a single pair of judges is the coefficient of concordance (W), which is the ratio of the variance of the sums of ranks in the sample to the maximum possible variance of the sums of the ranks. This statistic can be intuitively thought of as much like an average rank order correlation. The advantage of this statistic is that it has an exact test of statistical significance. The calculation formula for W is:

$$W = \frac{12ss}{m^2(n^3-n)},$$

where m is the number of experts, n is the number of events, and ss is the sum of squares,

$$ss = \sum_{j=1}^{n} (R_j - \bar{R})^2,$$

with R_j being the sum of ranks for event j and $\bar{R}$ being the average sum of ranks.

In order to calculate W, each expert's paired comparisons are converted into a rank order. This rank order can be derived by a count of the number of times each event was judged to be more likely than other events. The event with the largest count is ranked first and so on.

These rank orders are put into a table with m rows, one for each expert, and n columns, one for each event. Entries in the table are the ranks assigned by each expert to each event. Thus, a table such as A-5 results where there are eight experts and five events.

For this example,

$$W = \frac{12[(13-24)^2 + (16-24)^2 + (27-24)^2 + (29-24)^2 + (35-24)^2]}{64(125-5)}$$

$$= \frac{12(121 + 64 + 9 + 25 + 121)}{64(120)}$$

$$= .53.$$

This indicates that the variance on the sums of the ranks is about 53 percent of the maximum possible variance (produced if all experts agree completely). The significance of W for seven or fewer events can be found in Appendix C. If there are more than seven events, chi square tables found in most statistics text can be used to determine significance. In this case the statistic

$$T = \frac{12ss}{mn(n+1)}$$

is distributed approximately like chi square with n-1 degrees of freedom. In this formula, ss, m, and n are again the sum of squares for rank sums, the number of experts, and the number of events. Thus, in the example case, W is significant beyond the .01 level indicating basic agreement among experts.

A.2.5 Estimating uncertainty bounds. Conservative uncertainty bounds on scale values, s, can be estimated statistically using a procedure described in Appendix B. First the variance of each z value corresponding to each proportion in Table A-2 is determined. If the number of experts is ten or fewer, these variances can be obtained from Table A-6. If there are more than ten experts, procedures described in Appendix B that involve considerable calculation must be used. In Table A-6, z represents the value assigned to cells with complete agreement (e.g., from Table A-4, or from procedures in Appendix B).

Because of the tediousness of the calculations involved for 20 experts, for purposes of demonstration here we assume that the proportions in Table A-2 were generated by 10 rather than 20 experts. Then z=1.69 for 10 experts from Table A-4. For proportions not found in Table A-6, interpolation may be used as is demonstrated below.

TABLE A-5

Rank Ordering of Events*

Experts	Events				
	1	2	3	4	5
1	1	3	2	5	4
2	2	3	1	4	5
3	3	2	4	1	5
4	2	1	5	4	3
5	1	2	4	3	5
6	1	2	3	4	5
7	2	1	3	4	5
8	1	2	5	4	3
R_j	13	16	27	29	35

(R_j = sum of ranks for event j)

$\bar{R} = 24$

*A rank of 1 indicates the event chosen as more likely than other events most often, while a rank of 5 indicates the event chosen as more likely than other events least often.

TABLE A-6

Variance Estimates of z Values for Various Proportions and Numbers of Experts*

Number of Experts	Proportion				
	.5	.6	.7	.8	.9
2	$.5z^2$	$.48z^2$	$.42z^2$	$.32z^2$	$.18z^2$
3	$.25z^2$ +.139	$.257z^2$-.019z+.129	$.27z^2$ -.06z +.105	$.266z^2$-.125z+.073	$.2z^2$ -.135z+.041
4	$.125z^2$+.228	$.144z^2$-.027z+.211	$.194z^2$-.105z+.17	$.245z^2$-.211z+.131	$.226z^2$-.255z+.097
5	$.063z^2$+.261	$.083z^2$-.025z+.239	$.143z^2$-.107z+.198	$.221z^2$-.247z+.167	$.243z^2$-.344z+.152
6	$.031z^2$+.262	$.049z^2$-.019z+.244	$.104z^2$-.093z+.207	$.193z^2$-.250z+.189	$.249z^2$-.408z+.202
7	$.016z^2$+.248	$.029z^2$-.013z+.234	$.076z^2$-.074z+.205	$.166z^2$-.236z+.200	$.249z^2$-.451z+.245
8	$.008z^2$+.227	$.017z^2$-.008z+.219	$.054z^2$-.056z+.196	$.140z^2$-.211z+.201	$.245z^2$-.474z+.273
9	$.004z^2$+.205	$.010z^2$-.005z+.202	$.039z^2$-.041z+.189	$.116z^2$-.184z+.199	$.237z^2$-.483z+.296
10	$.002z^2$+.185	$.006z^2$-.003z+.184	$.027z^2$-.03z +.177	$.096z^2$-.155z+.191	$.227z^2$-.482z+.310

*Z represents the normal deviate value assigned to cells with complete agreement.

Then for the proportions in Table A-2, and the formulas for ten experts in Table A-6, the matrix of variances in Table A-7 is derived. For example, the proportion in cell 1,2 of Table A-2 is .75. From Table A-6, the variance for .7 is $.027(1.69)^2 - .03(1.69) + .177 = .2034$, and the variance for .8 is $.096(1.69)^2 - .155(1.69) + .191 = .2032$. Since .75 is halfway between .7 and .8, linear interpolation arrives at a variance that is halfway between these two estimates, .2033. This value, rounded to .203, is entered into the appropriate cell in Table A-7. This is also the estimate used for cell 2,1 since the variance for p is the same as the variance for 1-p. Other cells are filled in appropriately using the same procedure. The variance of the estimate of the scale value for each event is then calculated by summing each column and dividing by n(n-1) (in this case, n=6), as shown in Table A-7. Error bounds on the scale values--95 percent bounds--are then computed as $s \pm 2s.e.$, where s is the scale value (from Table A-3) and s.e. is the standard error of the estimate which is the square root of the variance of the estimate.

These uncertainty bounds on scale values can then be transformed into bounds on probability estimates as described in Appendix B (Section B.3). The uncertainty bounds for log HEP are $\pm$ 2as.e. where a is the constant from the equation,

$$\log HEP = as + b,$$

used to transform scale values into probabilities. In this example, a was determined to be equal to .94. Thus, the bounds on log HEP for the six events are $\pm$.323, $\pm$.342, $\pm$.337, $\pm$.331, $\pm$.337, and $\pm$.312 respectively. Tables of antilogarithms are then used to obtain the bounds on the HEPs shown in the last column of Table A-8.

Another approach to estimating uncertainty bounds is based on judgmental rather than statistical estimations. Judgmental procedures including paired comparisons, ranking/rating, direct numerical estimation, and indirect numerical estimation can all be used. The advantages and disadvantages of these various judgmental procedures to estimate uncertainty bounds are the same as those for those procedures used to estimate HEPs. Any of these four judgmental bounding procedures can be used with any of the five HEP estimation procedures. Here for the sake of consistency of presentation, we discuss the paired comparison bounding procedure. Each of the other judgmental bounding procedures is discussed in the section of this appendix describing the similar HEP estimation approach. In some applications, both a statistical and a judgmental bounding approach may be useful.

In the paired comparison bounding procedures, events are paired and presented to the experts just as described above for the paired comparison HEP estimation procedure. The experts now, however, are responding to the question:

TABLE A-7

Matrix of Variances

Event:	1	2	3	4	5	6
1	-	.203	.200	.194	.174	.114
2	.203	-	.194	.203	.196	.203
3	.200	.194	-	.196	.200	.174
4	.194	.203	.196	-	.203	.144
5	.174	.196	.200	.203	-	.196
6	.114	.203	.174	.144	.196	-
Σ	.885	.999	.964	.94	.969	.831
Variance= $\Sigma/n(n-1)$	.030	.033	.032	.031	.032	.028
Standard error = $\sqrt{\text{Variance}}$	.172	.182	.179	.176	.179	.166
s= Scale Value	-.64	.06	-.11	-.45	.38	.84
Error bounds on s (lower)	-.984	-.304	-.468	-.802	.022	.508
Error bounds on s (upper)	-.296	.424	.248	-.098	.738	1.172

TABLE A-8

Statistical Uncertainty Bounds on HEP Estimates

Event	HEP	log HEP	Bounds on log HEP	Bounds on HEP
1	.0004	-3.3979	-3.7419, -3.0539	.0002, .0009
2	.0018	-2.7399	-3.0819, -2.3979	.0008, .0040
3	.0013	-2.8997	-3.2367, -2.5627	.0006, .0027
4	.0006	-3.2193	-3.5503, -2.8883	.0003, .0013
5	.0040	-2.4391	-2.7761, -2.1021	.0017, .0079
6	.0100	-2.0000	-2.3320, -1.6680	.0047, .0215

> For which error are you more uncertain about how likely it is? Note that we are interested in your uncertainty about an estimate of likelihood rather than the estimate itself.

(If the paried comparison bounding procedure is used in conjunction with the paired comparison HEP estimation procedure, each pair of events needs to be presented only once, with two responses required.)

These paired comparisons of uncertainty are then scaled using the techniques described in Tables A-1 through A-3 and the accompanying text. This produces scale values representing the degree of uncertainty regarding the HEP of each event.

By properly phrasing a subsequent question (see below), we can assume that bounds are 95 percent bounds measured by $\pm x$ where x is an order of magnitude measure. That is, the interval (log HEP-x, log HEP+x) contains 95 percent of the possible HEP estimates. We also assume that

$$x = aT + b$$

where x is the uncertainty bound, T is the scale value derived from paired comparison scaling, and a and b are constants to be estimated as follows.

To estimate a and b, two independent estimates of bounds are needed. Assume, for example, for the six events in the example (Tables A-1 to A-3) we have obtained paired comparison scale bounding values, T, of:

Event:	1	2	3	4	5	6
T:	-.74	-.12	.02	1.22	-.46	.32

Further assume that the experts have provided an independent estimate of $x = 1.4$ for event 1 and $x = .5$ for event 4, by answering questions such as:

> For this event, a range of how many orders of magnitude would be required for you to feel 95 percent certain that the range contained the true HEP?

An alternative question that might be easier for the experts to answer would provide direct estimates of the upper and lower bounds:

> For this event, what are the upper and lower bounds on the HEP that make you 95 percent certain the true HEP falls between these bounds.

This question will provide the necessary estimate of x.

Then, the values of a and b are found by solving the simultaneous equations:

$$1.4 = a(-.74) + b$$

and

$$.5 = a(1.22) + b.$$

First, subtract the second equation from the first:

$$\begin{aligned} 1.4 &= a(-.74) + b \\ -\ .5 &= a(1.22) + b \\ \hline .9 &= a(-1.96) \\ .9/-1.96 &= a \\ -.459 &= a \end{aligned}$$

Substituting this value of a into the first equations, we find

$$1.4 = (-.459)(-.74) + b$$

$$1.4 = .34 + b$$

$$b = 1.4 - .34 = 1.06.$$

Using

$$x = -.459T + 1.06,$$

the following bounds (in orders of magnitude) are found:

Event:	1	2	3	4	5	6
T:	-.74	-.12	.02	1.22	-.46	.32
Bounds:	$\pm$1.4	$\pm$1.115	$\pm$1.051	$\pm$.5	$\pm$1.27	$\pm$.913

Since these bounds are on log HEPs, bounds on actual probabilities are found by taking first log HEP $\pm x$ and then taking antilogarithms as shown in Table A-9.

A.3 Ranking and Rating Procedures

The steps required in this procedure are the following:

1. Obtain ranking or rating judgments.

2. Create table showing the number of experts who placed each event in the various rating categories or rankings.

3. Derive cumulative frequency matrix showing the number of times each event was assigned a certain ranking/rating or lower.

TABLE A-9

Judgmental Uncertainty Bounds from
Paired Comparison Judgments

Event	log HEP	HEP	Bounds on log HEP	Bounds on HEP
1	-3.3979	.0004	-4.7979, -1.9979	.000016, .01
2	-2.7399	.0018	-3.8549, -1.6249	.00014, .024
3	-2.8997	.0013	-3.9507, -1.8487	.00011, .014
4	-3.2193	.0006	-3.7193, -2.7193	.00019, .0019
5	-2.4391	.004	-3.7091, -1.1691	.0002, .068
6	-2.0000	.01	-2.9130, -1.0870	.0012, .082

4. Transform matrix from 3 into a matrix of proportions.

5. Create table of normal deviates for proportions in table in 4.

6. Calculate row, column, and grand means for table in 5.

7. Scale values are the grand mean minus the row mean for each event.

8. Obtain independent estimate of two HEPs.

9. Transform scale values into HEPs.

10. Determine interjudge consistency.

11. Estimate uncertainty bounds.

A.3.1 Judgments required. Ranking and rating procedures are discussed together because the analytic techniques are the same regardless of the judgmental source of the raw data. Ranking procedures require that the expert rank order events by their relative likelihood. The rating procedure requires that the judge rate each event with respect to likelihood. The rating may be expressed on a numerical scale (e.g., 1 to 10), an adjectival scale, or a graphic scale. The graphic scale is arbitrarily divided into as many categories as desired by the analyst.

A.3.2 Performing the calculations. Both the ranking and rating procedures result in the same raw data form, i.e., a matrix of the frequencies with which each event is ranked or rated into each category. An example of such a matrix of HEPs is shown in Table A-10 with ratings from 10 experts. Adjectives describing the likelihood of an event are used as category labels, arranged in order of increasing likelihood. Note that these labels are not symmetric (e.g., from very unlikely to very likely) because the HEPs are all likely to be small. Alternatively, these categories could be ranks. The next step is to transform this matrix into one in which the cell entries are the frequency of times that the event was in that category or any of the categories below it (Table A-11). At this point, the last category, (e.g., category five in this example) is simply the total number of judgments made and is not used further. The entries in Table A-11 are converted into proportions and the result is Table A-12.

The next step is to assume that the proportions in Table A-12 represent proportions of the area under the normal distribution. The cell entries are then converted into unit normal deviates by using a table of values of the normal distribution found in most standard statistics texts. Each proportion from the matrix is found as F(z) (or F(x)) in the normal distribution table. The corresponding normal deviate, usually labeled as z (or x) is found and entered into the new matrix (Table A-13).

TABLE A-10

Category Frequency Judgments*

Events:	Categories				
	Almost Never	Extremely Unlikely	Very Unlikely	Unlikely	Somewhat Likely
1. Select wrong control in a group of identical controls identified by labels only	1	2	4	2	1
2. Select wrong control from a functionally grouped set of controls	4	2	2	1	1
3. Select wrong control from a panel with clearly drawn mimic lines	5	3	2	0	0
4. Turn control in wrong direction under normal operating conditions (violation of a strong populational bias)	0	1	1	2	6
5. Improperly mate a connector	1	1	2	4	2

*Each cell entry represents the number of experts who placed that event into that category.

TABLE A-11

Cumulative Frequency Matrix*

Event	Category 1	2	3	4	5
1	1	3	7	9	10
2	4	6	8	9	10
3	5	8	10	10	10
4	0	1	2	4	10
5	1	2	4	8	10

*Each cell entry represents the number of experts who placed that event into that category or any of the categories below it.

TABLE A-12

Cumulative Proportion Matrix*

Event	Category 1	2	3	4
1	.1	.3	.7	.9
2	.4	.6	.8	.9
3	.5	.8	1.0	1.0
4	0	.1	.2	.4
5	.1	.2	.4	.8

*Each cell entry represents the proportion of experts who placed that event into that category or any of the categories below it.

TABLE A-13

Matrix of Unit Normal Deviates [a]

Event	Category 1	2	3	4	Row Means
1	-1.28	-.52	.52	1.28	0
2	- .25	.25	.84	1.28	.53
3	0	.84	1.69 [b]	1.69 [b]	1.06
4	-1.69 [b]	-1.28	- .84	- .25	-1.02
5	-1.28	- .84	- .25	.84	- .38
Column Means	- .90	- .31	.39	.97	Grand Mean = .04

[a] Each cell entry represents the normal deviate value (z) corresponding to the cumulative proportion for that cell shown in Table A-12.

[b] These z values for cumulative proportions of 1 or 0 were taken from Table A-4.

As in the case of paired comparisons, there is a problem in how to handle the situation where the cell entry is unity. (This happens when all of the judgments in a cell fall at or below the upper boundary for that cell.) This information should not be discarded, but instead some value of z should be substituted that reflects the extremity of this situation. Appendix B (Section B.2) shows methods for choosing a z value for proportions of unity, one of which results in Table A-4. Using the value of $\pm$1.69 from Table A-4 for z with 10 experts and the other normal deviate values from a table of the normal distribution, the matrix in Table A-13 is obtained.

The next step is to calculate row and column means and the grand mean as shown in Table A-13. Values for the category boundaries are the column means. To get scale values, the row means for each event are subtracted from the grand mean. This procedure results in the following subjective scale for the 5 events:

Event #	3	2	1	5	4
Subjective Scale Value	-1.02	-.49	.04	.42	1.06

It still remains to convert these scale values into probabilities. The reader is referred to Section A.2.2 for rescaling procedures that are appropriate for both paired comparisons and data resulting from ranking or rating.

A.3.3 Interjudge consistency. For the case where rank order data are provided by the judge, the reader is referred to Section A.2.4 for methods of analyzing the level of agreement between judges. Rating data can be handled in much the same way as rank order data except that some provision must be made for ties in the ratings. The only difference from Section A.2.4 in terms of the method of calculating the coefficient of concordance is that the tied events are each given the average of the ranks they would have been assigned had no ties occurred.

The effect of tied ranks is to depress the value of W. If the proportion of ties is small, the effect is negligible. If the proportion of ties is large, a correction should be introduced that will slightly increase the value of W.

For example, suppose the data in Table A-10 had been generated by the ratings shown in Table A-14. From these ratings, the rank order of events for each expert can be derived as shown in Table A-15. Events with the same rating are considered to be tied in rank. Table A-15 also shows the sum of the ranks (R_j) for each event.

The coefficient of concordance, W, is then calculated as

$$W = \frac{12ss}{m^2(n^3-n) - m\sum_{m} T},$$

TABLE A-14

Ratings by Ten Experts of Five Events*

Expert	Events				
	1	2	3	4	5
1	1	2	1	3	3
2	2	1	1	5	2
3	2	1	2	2	4
4	3	2	1	5	3
5	3	1	2	4	1
6	3	3	3	5	4
7	3	4	1	4	4
8	4	1	2	5	5
9	4	5	1	5	4
10	5	3	3	5	5

*Each cell entry corresponds to the rating an expert gave to an event. As shown in Table A-10, a rating of 1 corresponds to "almost never" and a rating of 5 corresponds to "somewhat likely."

TABLE A-15

Rank Ordering of Events[a]

Expert	Events					
	1	2	3	4	5	T^b
1	1.5	3	1.5	4.5	4.5	1.0
2	3.5	1.5	1.5	5	3.5	1.0
3	3	1	3	3	5	2.0
4	3.5	2	1	5	3.5	.5
5	4	1.5	3	5	1.5	.5
6	2	2	2	5	4	2.0
7	2	4	1	4	4	2.0
8	3	1	2	4.5	4.5	.5
9	2.5	4.5	1	4.5	2.5	1.0
10	4	1.5	1.5	4	4	2.5
R_j	29.0	22.0	17.5	44.5	37.0	ΣT=13.0

(Sum of ranks for event j)

$\Sigma_j R_j = 150$

$\overline{R}_j = 30$

[a]Cell entries represent the rank ordering of the events for each expert.

[b]$T = \Sigma \frac{(t^3-t)}{12}$, where t=the number of events tied at a given rank and the summation (Σ) is across all groups of ties within the rank ordering for that expert.

where ss is the sum of squares ($\Sigma(R_j-\bar{R})^2$), m is the number of experts, n is the number of events, and $m\sum_{m}^{j}T$ is the correction for ties that is calculated as follows.

For the rank order of each expert,

$$T = \frac{\Sigma(t^3-t)}{12},$$

where t is the number of events tied at a given rank and the summation, Σ, is across all groups of ties within the rank order of that expert. For example, for expert number 1, there are two events tied at 1.5 and two events tied at 4.5. Thus, for expert 1,

$$T = \frac{(2^3-2) + (2^3-2)}{12} = \frac{(8-2) + (8-2)}{12} = 1.$$

The formula, $\sum_{m}T$, is then the sum of the T values across all experts' rank orders.

In the example in Table A-15, the corrected coefficient of concordance is then

$$W = \frac{12[(29-30)^2 + (22-30)^2 + (17.5-30)^2 + (44.5-30)^2 + (37-30)^2]}{100(125-5) - 10(13)}$$

$$= \frac{12(480.5)}{12000-130} = .486.$$

From tables for the significance of the coefficient of concordance in Appendix C, this value is significant beyond the .01 level assuring that the experts are essentially in agreement and that their "pooled" estimate is reasonable.

A.3.4 Estimating uncertainty bounds. Statistical estimates of uncertainty bounds cannot be made directly from the ranking/rating data unless the number of experts used is quite large, e.g., more than 25. Since it is unlikely that such a large number of experts would be practical, uncertainty bounds for a smaller number of experts can be computed by transforming the ranking/rating data into paired comparison data.

For example, the rating data from Table A-14 can be transformed into paired comparison data in the following way. For the ten experts, count the number of experts who rated event 2 as more likely than event 1. Ties are included as .5 in the count. Experts 1, 7, and 9 rated event 2 as more likely, and expert 6 rated them equally giving a count of 3.5. This number is entered into the cell for the event 2 column and event 1 row as in Table A-16. Also note that the entry for column 1 and row 2

TABLE A-16

Paired Comparison Data Derived from Ranking/Rating Data to Estimate Uncertainty Bounds*

Event	1	2	3	4	5
1	-	3.5	1.5	9.0	7.0
2	6.5	-	4.5	9.0	8.0
3	8.5	5.5	-	9.5	9.0
4	1.0	1.0	0.5	-	3.0
5	3.0	2.0	1.0	7.0	-

*Each cell entry represents the number of experts who said an event listed across the top was more likely than an event listed down the side. Since the data were derived from rating data, an expert who gave 2 events the same rating is counted as .5 for that cell.

is 10 - 3.5 = 6.5 where 10 is the number of experts. Using this method, Table A-16 can be completely filled. This table is then equivalent to Table A-1 and the same procedure for computing uncertainty bounds can be used that is described in Section A.2.5.

Any of the four judgmental procedures for estimating uncertainty bounds can also be used. The ranking/rating procedure is described here: The others are described in Sections A.2.5, A.4.4, and A.5.4.

Ranking or rating judgments can be elicited by asking questions such as the following:

> Rank order these events according to how certain you are about the likelihood of the event. Use 1 to represent the event for which you are most certain about the estimate and n (number of events) to represent the event for which you are most uncertain. Note that we are interested in your uncertainty about an estimate of likelihood rather than the estimate itself.

or

> Rate the events on the given scale according to how certain you are about the likelihood of the event. Note that we are interested in your uncertainty about an estimate of likelihood rather than the estimate itself.

(If the ranking/rating bounding procedure is used in conjunction with the ranking/rating HEP estimation procedure, the events need to be presented to the experts only once with two sets of rankings or ratings to be used.)

These ranking/rating data are then used to derive scale values representing the degree of uncertainty regarding the HEP of each event. The methods used to derive these scale values are the same as those used with Table A-10 through A-13 and described in the accompanying text. Once the scale values are obtained, they can be transformed into bounds on HEPs using the procedures described in Section A.2.5 for bounding scale values derived from paired comparisons.

A.4 Direct Numerical Estimation

The steps required in this procedure are the following:

1. Obtain odds judgments.

2. Aggregate individual judgments to provide a single odds judgment.

3. Transform odds estimates into HEPs.

4. Determine interjudge consistency.

5. Estimate uncertainty bounds.

A.4.1 Judgments required. Previous research (e.g., Stael von Holstein, 1972) suggests little improvement is gained in direct numerical estimation by using more than about six experts. Direct estimation procedures require that the judge provide numerical estimates of the likelihood of the error. The best procedure is to provide the expert with a logarithmically spaced scale of odds and ask him or her to mark the scale at the point that represents the odds (p/1-p)) that the event occurs. It is important that this odds scale be of sufficient detail that the sensitivity of the expert to differences in event odds be displayed. Figure A-1 shows several examples of response sheets with odds scales that might be used to assess HEPs.

Another consideration in scale design is that the scale values reflect the estimated range of the odds of the events about which the expert will make judgments. This consideration is particularly important for HEPs that are in the extreme ranges of the odds scale. Take, for example, the events listed below:

	Event	HEP*	Odds
1.	Failure to respond to annunciated legend light (one of one)	.0001	1:9999
2.	Incorrect reading of the message	.001	1:999
3.	Failure to resume attention to a legend light within 1 minute after an interruption	.001	1:999
4.	Failure to respond to a legend light if more than 1 minute elapses after an interruption	.95	19:1

In the case where the expert is to make a judgment about events 1, 2, or 3, the scale in Figure A-1 (a) would be appropriate since it provides sufficient divisions in the scale for the approximate magnitude of the odds, and the appropriate end of the scale is represented (all odds are

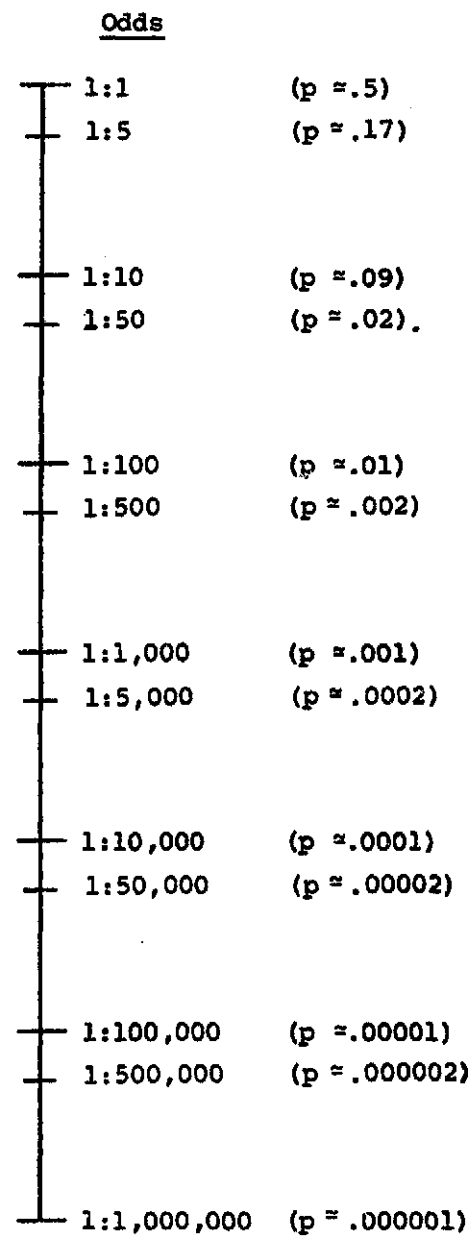
*From Swain and Guttmann (1980), p. 20-9.

Figure A-1(a)

Odds Response Scale

Mark the scale at the point that represents the odds against an average operator failing to respond to an annunciated legend light (one of one)

Mark the scale at the point that represents your judgment of the odds against an operator failing to respond to an annunciated legend light (one of one)



less than 1:1). For event 4 alone, where the odds are greater than 1:1, the scale should look like that shown in Figure A-1 (b), with a more restricted range and the scale representing odds greater than 1:1. Finally, for the case where several events covering the full range of odds are being considered, or where the approximate range of the event probability is unknown, a scale like Figure A-1 (c) should be used, where the scale covers odds both greater than and less than 1:1.

A.4.2 Performing the calculations. The only calculations necessary to produce HEPs by this procedure are the aggregation of individual expert judgments into single assessments for each event. The HEP for event j, HEP_j, is derived from the formula

$$HEP_j = \frac{\left(\prod_{i=1}^{m} Odds_{ij}\right)^{\frac{1}{m}}}{1 + \left(\prod_{i=1}^{m} Odds_{ij}\right)^{\frac{1}{m}}}$$

where $Odds_{ij}$ is the odds estimated by expert i for event j, and m is the number of experts.

To demonstrate this calculation, suppose six experts provided the odds judgments shown on Table A-17(a). These odds judgments must first be turned into decimal odds values as in Table A-17(b). Then to perform the aggregation by hand, the simplest procedure is to take logarithms of the values in Table A-17(b) creating a table such as A-17(c). The average logarithm value for each event is then computed as the sum of the numbers in the column divided by m (equal six in this example), the number of experts. Then antilogarithms are taken to provide an aggregated odds estimate for each event. These odds are converted into HEPs through the formula

$$HEP = \frac{Odds}{1+Odds}.$$

A.4.3 Interjudge consistency. The ANalysis Of VAriance (ANOVA) statistical paradigm provides a method for examining consistency across experts of judgments of probability. The method allows for variation to be separated into that accounted for by the differences among event probabilities, relative to the variation due to differences among judges. The statistic produced is called the intra-class correlation coefficient. It represents a measure of the average correlation between each pair of experts making estimates.

The analysis proceeds in two basic parts. The first is to conduct an ANOVA for repeated measures (because each expert makes a judgment of each of the events and has thus been measured repeatedly). An ANOVA is

Figure A-1(b)

Odds Response Scale for Likelihood Ratio Judgments

Mark the scale at the point that represents your judgment of the odds in favor of an operator failing to respond to a legend light if more than 1 minute elapses after an interruption

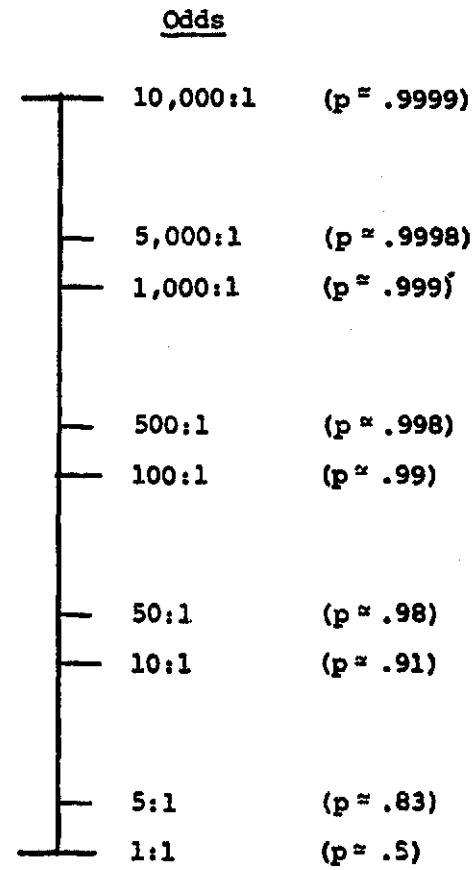


Figure A-1(c)

For each human error mark the scale at a point that represents your judgment of the odds in favor of that error occurring. Remember that, for example, odds of 1:100 mean that the error is one hundred times less likely to occur than not to occur

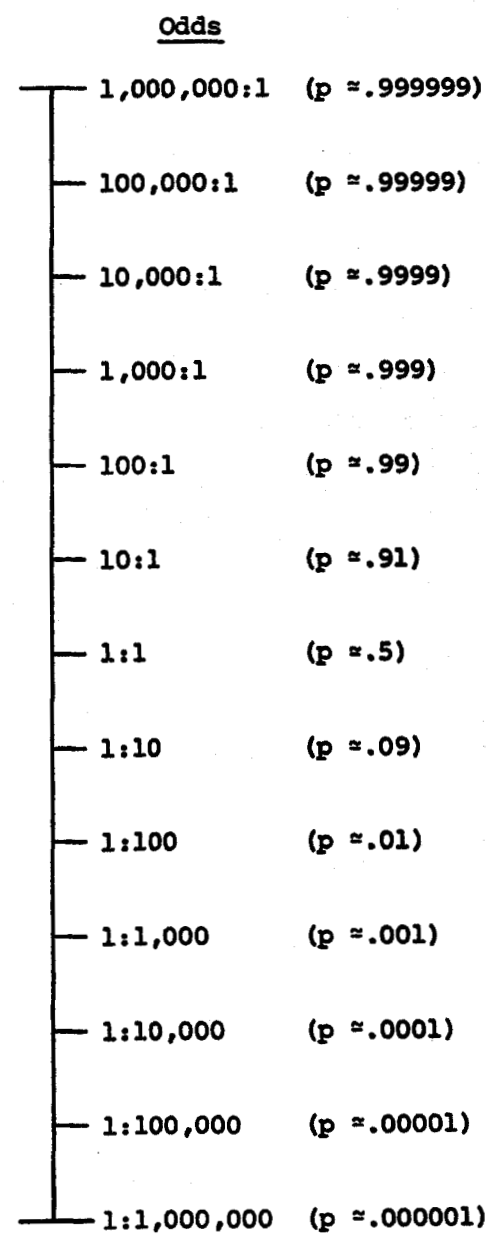


TABLE A-17

Aggregation of Direct Numerical Estimation of HEPs

a. Odds

	Event				
Expert	1	2	3	4	5
1	1:400	1:150	1:10	1:100	1:14
2	1:1000	1:100	1:9	1:75	1:9
3	1:2500	1:200	1:75	1:30	1:9
4	1:1250	1:1500	1:14	1:9	1:100
5	1:300	1:125	1:100	1:100	1:75
6	1:200	1:10	1:110	1:75	1:9

b. Decimal Values ($\times 10^{-3}$)

	Event				
Expert	1	2	3	4	5
1	2.5	6.67	100.0	10.0	71.4
2	1.0	10.0	111.1	13.3	111.1
3	.4	5.0	13.3	33.3	111.1
4	.8	.667	71.4	111.1	10.0
5	3.33	8.0	10.0	10.0	13.3
6	5.0	100.0	9.09	13.3	111.1

TABLE A-17 (Continued)

Aggregation of Direct Numerical Estimation of HEPs

c. log odds

	Event				
Expert	1	2	3	4	5
1	-2.6021	-2.1759	-1.0000	-2.0000	-1.1463
2	-3.0000	-2.0000	-0.9543	-1.8761	-0.9543
3	-3.3979	-2.3010	-1.8761	-1.4776	-0.9543
4	-3.0969	-3.1759	-1.1463	-0.9543	-2.0000
5	-2.4776	-2.0969	-2.0000	-2.0000	-1.8761
6	-2.3010	-1.0000	-2.0414	-1.8761	-0.9543
Sum/6	-2.8126	-2.1250	-1.5030	-1.6974	-1.3142
antilog (Odds)	.0015	.0075	.0314	.0201	.0485
HEP (=Odds/(1+Odds))	.0015	.0074	.0304	.0197	.0463

designed to divide the variation in the data (in this case the individual judgments) into portions due to the factor of interest (differences in probability among events) and that which is due to error (differences among judges in their estimates of the event probabilities).

The second part of the analysis is to use the estimates of the variance of events and judges to form the intraclass correlation coefficient. The coefficient represents a proportion, the proportion of the total variation in the data not produced by differences across experts in judgments. Thus, when this number is 1.0, it means that all of the variation is among events, and that the experts are perfectly consistent with each other. A value of 0 implies that none of the variation is produced by differences among events, i.e., all is because of interjudge inconsistencies.

An example should make these concepts more clear. Assume that each of 7 judges have made judgments about the probability of each of 5 events. Data that might result from these judgments are shown in Table A-18. The first step is to do a logarithmic transformation of each of these values. This is necessary to fulfill the assumption of the ANOVA model that the values be approximately normally distributed. The result is the table shown in A-19. These numbers will now constitute the input values for the ANOVA. The calculations required for an ANOVA are somewhat lengthy, and since there are a number of canned statistical computer packages that perform this type of analysis, we recommend that one of them be used (see, for example, the statistical package for the social sciences). In the example, however, the complete set of calculations is shown to demonstrate the procedure.

After the transformation, the next step in the calculations is to compute column (E_i) and row sums (J_j), and the grand sum (GT). These are shown in Table A-19. Next, we calculate the following intermediate quantities:

$$[T] = \frac{(GT)^2}{mn} = \frac{(-64.3078)^2}{7(5)} = \frac{4135.49}{35} = 118.16$$

$$[E] = \frac{\sum_{i=1}^{n}(E_i)^2}{m} = \frac{(-14.4894)^2+(-20.5879)^2+(-14.6655)^2+(-10.1426)^2+(-4.4224)^2}{7}$$

$$= \frac{971.31}{7} = 138.8$$

$$[J] = \frac{\sum_{j=1}^{m}(J_j)^2}{n} = \frac{(-10.0)^2+(-8.2006)^2+(-8.0635)^2+(-10.4437)^2+(-8.9058)^2+(-9.5406)^2+(9.1536)^2}{5}$$

TABLE A-18

Probability Estimates for 5 Events

Expert	Event 1. Failure to carry out a plant policy when there is no check on a person	2. Failure to initiate a checking function	3. Failure to use a value restoration list	4. Failure to use written calibration procedures	5. Failure to use written maintenance procedures when available
1	.020	.0001	.001	.1	.5
2	.005	.003	.02	.06	.35
3	.009	.008	.1	.02	.06
4	.015	.0004	.003	.01	.2
5	.004	.0023	.006	.15	.15
6	.01	.0009	.01	.008	.4
7	.006	.0013	.006	.05	.3

TABLE A-19

Logarithmic Transformations of HEPs for ANOVA

Judge	Event 1	2	3	4	5	J_j (Row Sum)
1	-1.6990	-4.0000	-3.0000	-1.0000	- .3010	-10.00
2	-2.3010	-2.5229	-1.6990	-1.2218	- .4559	- 8.2006
3	-2.0458	-2.0969	-1.0000	-1.6990	-1.2218	- 8.0635
4	-1.8239	-3.3979	-2.5229	-2.0000	- .6990	-10.4437
5	-2.3979	-2.6383	-2.2218	- .8239	- .8239	- 8.9058
6	-2.0000	-3.0458	-2.0000	-2.0969	- .3979	- 9.5406
7	-2.2218	-2.8861	-2.2218	-1.3010	- .5229	- 9.1536
E_i = (Column Sum)	-14.4894	-20.5879	-14.6655	-10.1426	-4.4224	GT = -64.3078 (Grand Sum)

$$= \frac{595.47}{5} = 119.1$$

and

$$[EJ] = \sum_{i=1}^{n} \sum_{j=1}^{m} (E_i J_j)^2 = (-1.6990)^2 + (-2.3010)^2 + (-2.0458)^2 + (-1.8239)^2 + \ldots + (-.3979)^2 + (-.5229)^2 = 145.9$$

In these equations, GT is the grand sum, E_i is the column sum for event i, J_j is the row sum for judge j, and m and n are the number of judges and events respe tively.

The variance estimates can now be calculated for each of the components as shown in Table A-20. By dividing each of the sums of squares of differences by the appropriate degrees of freedom (the number of levels of that factor minus 1), a quantity known as the mean squared differences is obtained. The ratios of the mean squared differences for events and judges compared to the events by judges interaction produces a ratio of variance estimates that is distributed as an F statistic. A table of the F distribution (found in any standard statistics textbook) will then give the statistical significance of each factor. In our example the effect for events was significant beyond the .001 level. The effect for judges was not significant, however, implying a high degree of consistency among experts.

The last step is to calculate the intraclass correlation coefficient from the quantities determined for the ANOVA. The formula is:

$$r = \frac{F-1}{F + (n-1)}$$

$$= \frac{20.0 - 1}{20.0 + (5-1)}$$

$$= .79,$$

where F is the F ratio for the events factor. As discussed earlier, this quantity can be interpreted as an average correlation. Thus, a value of .79 indicates a high degree of agreement among judges.

A.4.4 Estimating uncertainty bounds. Statistical estimation of uncertainty bounds for HEPs obtained from direct numerical estimation should be calculated using logarithms of the HEPs rather than the HEPs themselves. Thus, the first step is to take the logarithms of each expert's HEP estimates. The odds estimates in Table A-17(a) must be converted

TABLE A-20

ANOVA Calculations

Sources of Variation	Calculations	Sums of Squared Differences	Degrees of Freedom	Mean Square	F Ratio
Events	[E]-[T]	138.8-118.2=20.6	5-1=4	$\frac{20.6}{4}$=5.2	20.0**
Judges	[J]-[T]	119.1-118.2=.9	7-1=6	$\frac{.9}{6}$=.15	.58 n.s.
Events X Judges	[EJ]-[J]-[E]+[T]	145.9-119.1-138.8+118.2=6.2	(5-1)(7-1)=24	$\frac{6.2}{24}$=.26	
TOTAL	[EJ]-[T]	145.9-118.2=27.7	34		

**p<.001

into probabilities using the formula P = odd/(1+odds). This results in data such as in Table A-21 which shows six experts' estimates of log HEPs for five events.

The variance over experts for these log HEP estimates, V(log HEP_i), is calculated as

$$V(\log HEP_i) = \frac{m \sum_{j=1}^{m} \log HEP_{ij}^{2} - \left(\sum_{j=1}^{m} \log HEP_{ij}\right)^{2}}{m(m-1)},$$

where HEP_{ij} is the HEP estimated for event i by expert j and m is the number of experts. From the variance, the standard error of the estimate (s.e.) can be calculated by dividing by m (m=6 in this case) and taking the square root:

$$\text{s.e.} = \sqrt{\frac{V(\log HEP_i)}{m}}.$$

The approximate 95 percent uncertainty bounds for the logarithms of the HEP estimates are $\pm$2s.e. The HEPs from Table A-17(c), log HEPs, and uncertainty bounds for the example are shown in Table A-22. The bounds on HEPs are found by taking antilogarithms of the bounds on log HEPs.

In addition to this statistical procedure, uncertainty bounds can also be estimated judgmentally. Here, the use of direct numerical judgments of uncertainty bounds is described. Other judgmental methods are discussed in Sections A.2.5, A.3.4, and A.5.4.

Direct numerical estimates of uncertainty bounds can be obtained using response scales such as those shown in Figure A-1. Each expert is asked to mark on the scale the values such that he/she is 95 percent certain the HEP is between, i.e., mark the lower and upper bounds. These uncertainty bound estimates of the experts are aggregated in the same manner as are HEP estimates (see Section 4.2). That is the lower bounds estimates of individual experts are aggregated to produce a lower bound, and the upper bound estimates of the individual experts are aggregated to produce the upper bound.

A.5 Indirect Numerical Estimation

The steps required in this procedure are the following:

1. Obtain indirect numerical relative likelihood judgments.
2. Obtain independent estimate of one HEP.
3. Calculate odds from 1 and 2.

TABLE A-21

Statistical Estimation of Uncertainty Bounds
for Direct Numerical Estimation*

	Event				
Expert	1	2	3	4	5
1	-2.6031	-2.1790	-1.0414	-2.0043	-1.1761
2	-3.0004	-2.0043	-1.0000	-1.8808	-1.0000
3	-3.3981	-2.3032	-1.8808	-1.4914	-1.0000
4	-3.0973	-3.1764	-1.1761	-1.0000	-2.0043
5	-2.4786	-2.1004	-2.0043	-2.0043	-1.8808
6	-2.3032	-1.0414	-2.0453	-1.8808	-1.0000
$V(\log HEP_i)$ (Variance $\log HEP_i$)	.1748	.4658	.2517	.1567	.2215
s.e. (Standard error)	.1707	.2786	.2048	.1616	.1921
2s.e.	.3414	.5573	.4096	.3232	.3843

$$= \frac{m \sum_{j=1}^{m} \log HEP_{ij}^2 - \left(\sum_{j=1}^{m} \log HEP_{ij}\right)^2}{m(m-1)}$$

where m=the number of experts

$$= \sqrt{\frac{V \log HEP_i}{m}}$$

*Each cell entry represents the log HEP for that event by that expert.

TABLE A-22

Statistical Uncertainty Bounds on Direct Numerical Estimates of HEPs

Event	HEP	log HEP	Bounds on log HEP	Bounds on HEP
1	.0015	-2.8239	-3.1653, -2.4825	.00068, .0033
2	.0074	-2.1308	-2.6881, -1.5735	.0021, .0267
3	.0304	-1.5171	-1.9267, -1.1075	.0118, .0781
4	.0197	-1.7055	-2.0287, -1.3823	.0094, .0415
5	.0463	-1.3344	-1.7187, - .9501	.0191, .1122

4. Aggregate odds for individual experts into a single odds estimate.

5. Transform odds estimates into HEPs.

6. Determine interjudge consistency.

7. Estimate uncertainty bounds.

A.5.1 Judgments required. Indirect numerical estimation requires judgments of the relative likelihood of pairs of human errors. For example, the experts might be asked which is more likely, the failure to detect one deviant unannunciated display at a single scan for a meter with limit marks or without limit marks; and how many times more likely. As with direct numerical estimation, little is gained by using more than six experts (Stael von Holstein).

To obtain a complete set of probabilities for n events requires a minimum of n-1 such judgments and an independent estimate of the HEP of one event. Additional judgments, however, may be obtained as consistency checks. For example, event a may be judged three times as likely as event b; and b is considered twice a likely as event c. As a consistency check, a could be compared with c. A consistent judgment would be that a is six times as likely as c. If an inconsistent judgment is obtained from the expert, the inconsistency should be pointed out and the expert should reconcile his/her judgments.

The determination of which events are paired for judgments should be accomplished by first making a rough rank ordering of the events according to likelihood. Adjacent events should be paired, with any consistency checks coming from pairings with other events close in rank. This procedure will minimize biases in judgment that tend to occur as such judgments become more extreme.

In addition to the relative likelihood judgments, an independent estimate of one HEP is necessary to serve as the anchor by which to convert the relative numerical judgments into probabilities. This may be provided by including an event with known probability in the set of events considered, or by direct numerical judgment using the questions and response scales described in Section A.4.

As with direct numerical estimation, the response scale should be carefully selected. For these judgments the range of responses should be more limited than it is for direct numerical judgments because of the way in which events are paired. Only the increasing portion of the scale needs to be used since all responses will be 1:1 or greater. Also, more detailed gradations in the scale will be necessary, because some of the relative likelihoods may be quite close to 1:1. An example assessment scale is shown in Figure A-2.

Figure A-2

Assessment Scale for Likelihood Ratio Judgments

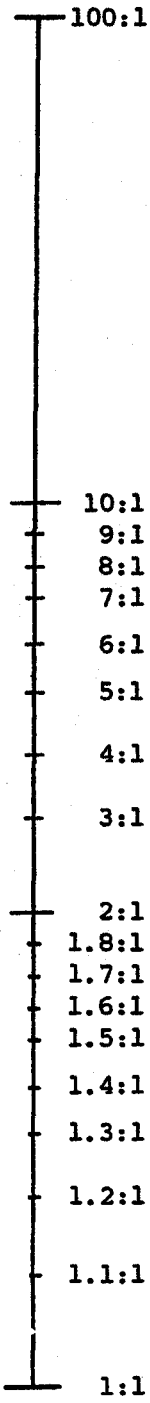
Place a mark beside the error of the following pair of human errors that is more likely:

1. In one scan, failure to detect one (of one) deviant, unannunciated display, where the display is a meter with limit mark

or

2. In one scan, failure to detect one (of one) deviant, unannunciated display, where the display is a meter without limit marks

Now mark the scale at the right at the point that represents how many times more likely the error you have chosen is than the other



If more than one expert is providing judgments, the probabilities derived from the individual judgments can be aggregated directly, or the NGT can be used to allow the experts to interact. For each judgment, it would proceed as follows.

1. With the experts together as a group, each expert, without discussion, privately judges the relatives likelihood of a pair of events.

2. Then, again without discussion, each expert presents his/her estimate to the group.

3. The judgments are discussed for clarification and evaluation, with a discussion leader controlling discussion to maintain relevance and prevent domination.

4. Each expert individually re-evaluates his/her judgment.

Once the final judgments are made, the individual estimates are aggregated using the same procedure described in Section A.4.2. If time limitations are a problem, use of the NGT will be faster if several judgments at a time are made and discussed.

A.5.2 Performing the calculations. The first calculation required is to transform the relative likelihood judgments into probabilities for each expert. Suppose one expert gave the judgments shown in Table A-23. Also suppose there was an independent estimate of the HEP of event 1=.05 or odds of 1:19 (where odds are p/(1-p)). Then the odds for event 2 are three times the odds for event 1:

$$\text{Odds}_2 = \frac{3}{1} \cdot \frac{1}{19} = \frac{3}{19} .$$

Similarly the odds for event 3 are 1/1.5 or 2/3 times the odds for event 2 or 2/19. Continuing this multiplicative process for all events leads to the odds shown in Table A-24. The HEPs are calculated as

$$\text{HEP} = \frac{\text{Odds}}{1 + \text{Odds}} .$$

After HEPs are derived for each expert, they must be aggregated across experts to obtain single HEP estimates for each event. The procedure used for aggregation is described in Section A.4.2.

A.5.3 Interjudge consistency. As in the case of direct numerical estimation, the intraclass correlation coefficient provides a measure of interjudge consistency for relative numerical estimation of likelihoods. Once HEPs have been derived from the relative likelihood judgments, the

TABLE A-23

Example of Relative Likelihood Judgments

Events*: Failure to detect one (of one) deviant unannunciated display at a single scan for display type:	Event Pair	Relative Likelihood
1. Meter with limit marks	1 x 2	1:3
2. Meter without limit marks	2 x 3	1.5:1
3. Chart recorders with limit marks	3 x 4	1:3
4. Chart recorders without limit marks	4 x 5	1:20
5. Annunciator light no longer annunciating	5 x 6	1:5
6. Legend light other than annunciator light	6 x 7	1:2
7. Indicator lamp		

*Events taken from Swain and Guttmann (1980) p. 20-12.

TABLE A-24

Odds and HEPs Derived from Relative Likelihood Judgments

Event	Odds	HEP
1	1/19	.05
2	3/19	.136
3	2/19	.095
4	6/19	.24
5	120/19	.863
6	600/19	.969
7	1200/19	.984

method described in Section A.4.3 can be used to calculate the intra-class correlation coefficient.

A.5.4 Estimating uncertainty bounds. Statistical uncertainty bounds on the HEP estimates can be derived using the same method as used for direct numerical estimation, which is described in Section A.4.4.

Judgmentally estimated uncertainty bounds can be obtained using any of the judgmental procedures described in Sections A.2.5, A.3.4, and A.4.4; or using the following procedure based on relative numerical estimates.

Each expert is given n-1 pairs of events (or more if consistency checks are wanted) and asked to judge the relative uncertainty of likelihood estimates for the two events:

> Express in terms of order of magnitude your relative uncertainty about the likelihood of occurrence of the two events. For example, if you feel that the 95 percent uncertainty bounds for event a are plus and minus one order of magnitude and the same bounds for event b are plus and minus one and a half orders of magnitude, your response should be that b is more uncertain than a and the ratio of uncertainty is 1.5 to 1.

In addition to these judgments, one independent estimate of uncertainty bounds is needed, either from an otherwise known estimate or from a direct numerical estimate using a question such as given in Section A.2.5.

Each expert's uncertainty bounds are then determined using the single independent estimate and multiplying by the appropriate relative uncertainty judgments as shown in Table A-25. For this example, it is assumed there is an independent estimate of the uncertainty bounds on event 1 of .8 orders of magnitude. To aggregate these uncertainty bounds across experts, simply take the average of the estimates for each event.

These uncertainty bounds are on log HEPs. Bounds on actual probabilities are found by first finding log HEP $\pm$ the uncertainty bound, and then taking antilogarithms of these values as was shown in Table A-9.

A.6 Multiattribute Utility Procedure

The steps required in this procedure are the following:

1. Determine relevant PSFs.

2. Rank order relative importance of PSFs.

TABLE A-25

Calculation of Uncertainty Bounds from Relative Numerical Judgments

Event Pair	Relative Uncertainty	Event	Uncertainty Bounds
		1	.8
1 x 2	2:1	2	.4
2 x 3	1:1.2	3	.48
3 x 4	3:1	4	.16
4 x 5	1:2	5	.32
5 x 6	1:1.5	6	.48
6 x 7	1:1.5	7	.72

3. Weight the PSFs.

4. Rate the effect of each PSF on each event.

5. For each expert and each event, calculate a scale value as the sum of the weighted ratings.

6. Obtain independent estimates of two events.

7. Transform scale values of individual experts into HEPs.

8. Aggregate individual HEP estimates into a single estimate.

9. Determine interjudge consistency.

10. Estimate uncertainty bounds.

A.6.1 Preliminary preparation. The simplest to use multiattribute utility procedure is one based on Edwards (1977) and described completely in Embrey (1981b). The experts must first determine which PSFs are most relevant for the set of events being considered. It is important that all consideration of PSFs be specific to the particular set of events. The initial consideration of PSFs should be a screening to reduce the number of possible PSFs to a practical number. Our recommendation is that no more than about 11 be used in subsequent scaling. This number or fewer can be obtained by having the experts rank order a set of PSFs, for example those in Exhibit 2-1 (from Swain and Guttmann, 1980). Embrey (1981b) used another set of PSFs that included:

- operator awareness of system state,
- training,
- feedback,
- clarity of responsibility,
- time,
- quality of procedures,
- complexity of decision making,
- opportunities for error correction,
- complexity of task,
- quality of supervision and checking,
- consequences of failure,
- degree of task functional isolation, and
- adverse environmental conditions.

Each PSF should be defined as well as possible. The experts should be instructed:

> For the set of events under consideration, rank these PSFs in the order of the degree to which they affect the probability of an error on these events.

This ranking can be carried out without any interaction among the experts, or the NGT can be used. In the latter case, the experts would be brought together. After explanations of the purpose of the study and descriptions of the events and PSFs to be considered, they would be asked individually to rank order the PSFs. Then each expert would present his/her rank order to the group. Following this, the experts could discuss the rank orders and the relative effects of the various PSFs. Finally, after the discussion, each expert would again rank order the PSFs. Whether or not the NGT is used, the PSFs to be used in the subsequent estimation process can be determined by averaging the ranks assigned to each PSF and using the PSFs with the highest average ranks.

A.6.2 Judgments required. At present, there is no basis, either empirical or formal, for determining the number of experts needed for this procedure. Research will be required on this subject prior to implementing the multiattribute utility procedure. The first judgments required from the experts are weights for the extent to which the PSFs affect the likelihood of error for the events specified. These weights can be elicited using the following instructions.

1. Rank order the PSFs in terms of the extent to which they determine the likelihood of an error for the events under consideration. It is important to keep in mind that we are referring to the effects of the PSFs on this specific set of events and not the general importance of the PSFs.

2. Now weight the PSFs. Arbitrarily assign a value of 10 to the PSF having the least effect on error likelihood for this specific set of events. Compare this PSF with the next higher ranked PSF. Assign a weight reflecting the ratio of the relative effect on error likelihood of the next higher ranked PSF compared with the lower ranked PSF. For example, if the higher ranked PSF has twice the effect, assign a weight of 20; or if it has one and a half times the effect, assign a weight of 15.

3. Continue comparing PSFs ranked next to each other and assigning ratio weights. For example, if the lower ranked PSF has a weight of 40 and the higher ranked PSF has three times the effect on likelihood of error, assign it a weight of 120.

4. Check for consistency among weights by comparing PSFs that are not adjacent in rank. For example, if one PSF has a weight of 150 and another a weight of 30, the former should have five times the effect.

Again this procedure can be accomplished by the experts individually, e.g., through a mail-out questionnaire, or using the NGT if experts are brought together. If the NGT is not used, all calculations should be performed independently for each expert with aggregation after probabilities have been derived.

If the NGT is used, each expert should weight the set of PSFs as described above. Then the steps of the NGT should be followed (see ranking of PSFs above), until subsequently each expert has produced a set of weights.

The next set of judgments each expert must make is an evaluation of the degree to which each PSF is "good" or "bad" as it occurs for each event. These judgments are made on a 0 to 100 scale where 0 means the PSF has the maximally degrading effect on task performance, and 100 indicates the PSF's effect is maximally enhancing. These anchors should be defined as precisely as possible.

If the NGT is being used, each evaluation estimate (for one PSF and one event) should be made using the first four steps in the NGT. It may be necessary, in order to save time, to have each expert make evaluation judgments for all events on a single PSF, and then follow the NGT for these sets of judgments. The result of this process for each expert, whether the NGT is used or not, is a set of evaluation judgments (on a 0 to 100 scale) describing the extent to which each PSF degrades or enhances each event, and a set of weights describing the relative importance of each PSF in determining the likelihood of error.

A.6.3 Performing the calculations. The exact nature of the calculations depends to some extent on whether or not the NGT is used. Calculations for when it is not used are described first.

Each expert will have made a set of judgments such as those described in Table A-26. Assume the events are the same five events described in Section A.3. The first calculation is to normalize the weights by dividing each weight (W_j) by the sum of weights. Then, for each event i, an index of likelihood, S_i, is calculated by multiplying the normalized weight, (NW_j), by the evaluation, X_{ij}, of event i on PSF_j, and summing across PSFs:

$$S_i = \sum_{j=1}^{n} NW_j X_{ij}.$$

These values for the example are shown in the last column of Table A-26.

The scale values are transformed into probabilities using the same procedure used for paired comparisons and ranking/rating procedures. Independent estimates are needed for two probabilities. With these estimates, the procedure described in Section A.2.2 is used.

TABLE A-26

Example of Multiattribute Utility Judgments[a]

Event	PSFs					Index of Likelihood (S_i)[c]
	Awareness of System State	Training	Feedback	Time	Quality of Procedures	
1	40	50	80	90	70	51.5
2	20	40	60	90	50	34.7
3	20	60	80	90	60	41.5
4	40	80	90	60	80	57.0
5	50	60	70	20	70	53.4
Weight (W_j) $\Sigma W_j = 235$	135[b]	45	30	15	10	
Normalized Weight (NW_j) = $\frac{W_j}{\Sigma W_j}$	.574	.191	.128	.064	.043	

[a] Each cell entry (X_{ij}) represents the weight given to PSF_j for event i by an expert.

[b] Each value in this row (W_j) represents the weight given to PSF_j for the set of events under consideration by an expert.

[c] The index of likelihood for event i (S_i) = $\sum_{j=1}^{n} NW_j X_{ij}$, where n refers to the number of PSFs.

For each event, each expert's HEP estimates must be aggregated to obtain a single HEP. This aggregation is accomplished using the formula

$$HEP_j = \frac{\left(\prod_{i=1}^{m} \frac{HEP_{ij}}{1-HEP_{ij}}\right)^{\frac{1}{m}}}{1+\left(\prod_{i=1}^{m} \frac{HEP_{ij}}{1-HEP_{ij}}\right)^{\frac{1}{m}}}$$

where HEP_j is the aggregated HEP estimate for event j, HEP_{ij} is the probability estimated for event j by expert i, and m is the number of experts. Suppose, for example, the HEP estimates of six experts for five events were as shown in Table A-27. To perform the aggregation by hand, the simplest procedure is to create a new table with entries $HEP_{ij}/(1-HEP_{ij})$ as shown in Table A-28(a). (Entries in Table A-28 are $HEP_{ij}/(1-HEP_{ij}) \times 10^{-3}$.) Then, logarithms of these entries are obtained as shown in Table A-28(b). These logarithms in each column (event) are summed, the sum is divided by m(6), and antilogarithms are taken resulting in:

Event:	1	2	3	4	5
Sum of logs:	-11.284	-16.136	-11.985	-10.629	-9.59
Sum/6:	- 1.881	- 2.689	- 1.998	- 1.772	-1.599
antilogarithm:	.013	.002	.010	.017	.025
HEP=antilog/ (1+antilog):	.013	.002	.010	.017	.025

These antilogarithms are equal to $\left(\prod_{i=1}^{m} \frac{HEP_{ij}}{1-HEP_{ij}}\right)^{\frac{1}{m}}$. The HEPs are then found by dividing each of these values by one plus the value.

If the NGT is used, aggregation across experts occurs at different points in the process. Individual judgments regarding weights are aggregated first using the formula

$$W_j = \left(\prod_{i=1}^{m} W_{ij}\right)^{\frac{1}{m}},$$

where W_j is the weight for PSF_j, W_{ij} is the weight assigned by expert i to PSF_j, and m is the number of experts.

For example, if six experts have assigned the weights to the five PSFs as shown in Table A-29, the calculation proceeds as follows. The product of weights for each PSF is calculated. The weight for each PSF is

TABLE A-27

Individual HEP Estimates

	Event				
Expert	1	2	3	4	5
1	.013	.001	.011	.024	.031
2	.021	.002	.003	.017	.019
3	.008	.001	.026	.026	.024
4	.005	.012	.008	.006	.018
5	.031	.001	.020	.015	.027
6	.014	.003	.007	.022	.032

TABLE A-28

Steps in Aggregation of Individual HEP Estimates

a. $HEP_{ij}/1\text{-}HEP_{ij}$ ($\times 10^{-3}$)

	Event				
Expert	1	2	3	4	5
1	13.17	1.001	11.12	24.59	31.99
2	21.45	2.004	3.009	17.29	19.37
3	8.065	1.001	26.69	26.69	24.59
4	5.025	12.15	8.065	6.036	18.33
5	31.99	1.001	20.41	15.23	27.75
6	14.20	3.009	7.049	22.49	33.06

b. $\log[(HEP_{ij}/(1\text{-}HEP_{ij})]$

	Event				
Expert	1	2	3	4	5
1	-1.880	-3.000	-1.954	-1.609	-1.495
2	-1.669	-2.698	-2.522	-1.762	-1.713
3	-2.093	-3.000	-1.574	-1.574	-1.609
4	-2.299	-1.916	-2.093	-2.219	-1.737
5	-1.495	-3.000	-1.690	-1.817	-1.557
6	-1.848	-2.522	-2.152	-1.648	-1.481

TABLE A-29

Calculation of Weights for Multiattribute Utility Procedure Using NGT*

Expert	PSFs				
	Awareness of System State	Training	Feedback	Time	Quality of Procedures
1	135	45	30	15	10
2	120	80	40	10	20
3	40	50	10	15	20
4	90	30	45	60	10
5	40	40	10	15	30
6	90	180	45	10	30
ΠW_{ij}	2.09952×10^{11}	3.888×10^{10}	2.43×10^{8}	2.025×10^{7}	3.6×10^{7}
$\log \Pi W_{ij}$	11.32212	10.58973	8.38561	7.30643	7.55630
$(\log \Pi W_{ij})/6$	1.8870	1.7650	1.3976	1.2177	1.2594
W_j	77.09	58.21	24.98	16.51	18.17
$\Sigma W_j =$	194.96				
normalized W_j $(NW_j) = \frac{W_j}{\Sigma W_j}$	.395	.299	.128	.085	.093

*Each cell entry (W_{ij}) represents the weight given to PSF_j by expert i for the set of events under consideration.

the mth root of this product which can be found by first taking the logarithm of the product; dividing by m (the number of experts), in this case, six; and taking antilogarithms. For subsequent calculations the resulting weights are normalized to sum to one. This normalization is accomplished by dividing each weight, W_j, by the sum of all weights.

Then, the evaluation judgments by each expert for each event on each PSF are aggregated by averaging across experts. For example, if the judgments of the six experts for the five events with respect to training are as shown in Table A-30, the aggregated evaluations for these events on this PSF are those shown at the bottom of Table A-30. Similar computations would be performed for each PSF.

After all aggregations are performed (weights and evaluations for each event on each PSF), a table such as that shown previously as Table A-26 and discussed with accompanying text results. Now, however, entries are already aggregated across experts rather than those of a single expert. The transformation of these resulting scale values of likelihood is again performed as discussed in Section A.2.2.

A.6.4 Interjudge consistency. The intraclass correlation coefficient is used to measure interjudge consistency in the multiattribute utility procedure as well as in the direct and indirect numerical procedues. To calculate this coefficient, HEPs of individual experts for each event must be calculated. If the NGT was not used, these probabilities have already been derived using the methods described in Section A.6.3. If the NGT was used, these individual probabilities must be derived using the same method as used when the NGT is not used. Once the individual expert HEPs are derived, the method for calculating the intraclass correlation coefficient described in Section A.4.3 can be applied.

A.6.5 Estimating uncertainty bounds. When the NGT is not used, statistical uncertainty bounds can be calculated just as they are for direct numerical estimation as described in Section 4.4. Also, any of the four judgmental methods for estimating uncertainty bounds, described in Sections A.2.5, A.3.4, A.4.4, and A.5.4, can be used.

If the NGT is used to calculate statistical uncertainty bounds, scale values for the events are computed individually for each expert, as described in Table A-26 and the accompanying text. This will provide the entries for a table such as Table A-31, for example, which shows the scale value of each event for each expert. For each event, i, the variance of scale values is calculated by the formula

$$V(S_i) = \frac{m\sum_{j=1}^{m} S_{ij}^2 - \left(\sum_{j=1}^{m} S_{ij}\right)^2}{m(m-1)},$$

TABLE A-30

Example of Aggregation of Evaluation Judgments on PSFs*

PSF: Training

Expert	Event				
	1	2	3	4	5
1	50	40	60	80	60
2	20	20	40	90	50
3	60	30	50		40
4	30	40	70	60	50
5	40	40	50	60	70
6	30	20	60	70	50
Σ	230	190	330	430	320
Σ/6	38.3	31.7	55.0	71.7	53.3

*Each cell entry represents the weight given to a PSF for an event by an expert.

TABLE A-31

Calculation of Uncertainty Bounds for Multiattribute Utility Procedure When NGT is Used*

Expert	Event				
	1	2	3	4	5
1	51.5	34.7	41.1	57.0	53.4
2	71.6	32.6	41.8	65.7	68.4
3	41.1	36.4	56.4	56.4	53.8
4	43.7	59.6	52.4	46.9	65.8
5	63.5	42.6	59.4	56.8	61.6
6	48.4	34.2	40.7	57.5	68.8
Variance(S_i) $[V(S_i)]=$ $\frac{m\sum_{j=1}^{m} S_{ij}^2 - (\sum_{j=1}^{m} S_{ij})^2}{m(m-1)}$	141.4	104.1	71.4	35.6	48.6
s.e. $= \sqrt{\frac{V(S_i)}{m}}$	4.86	4.16	3.45	2.43	2.85

where m=number of experts

*Each cell entry (S_{ij}) represents the index of likelihood for event i for expert j.

where $V(S_i)$ is the variance of the scale value for event i, S_{ij} is the scale value of event i for expert j, and m is the number of experts (six in this example). From the variance the standard error of the estimate (s.e.) can be calculated by dividing by m and taking the square root:

$$\text{s.e.} = \sqrt{\frac{V(S_i)}{m}}.$$

These standard errors for scale values are then transformed into uncertainty bounds on HEPs as described in Section A.2.5.

In addition to these statistical uncertainty bounds, any of the judgmental methods for estimating uncertainty bounds can also be used.

APPENDIX B

SUGGESTIONS REGARDING THE NUMBER OF EXPERTS AND THE HANDLING OF COMPLETE AGREEMENT IN THE PAIRED COMPARISON AND RANKING/RATING PROCEDURES

David A. Seaver and Robert C. Bromage

This appendix is intended to guide users in determining the number of experts needed for the paired comparison or rank/rating procedures. A necessary part of this guidance, and results often needed for implementing these procedures, is the development of techniques for handling complete agreement. Several practical techniques are suggested.

One of the most difficult practical problems in the use of the paired comparison procedure based on Thurstone's Law of Comparative Judgment (LCJ) or the ranking/rating procedures based on the Law of Categorical Judgment appears to be that they require a large number of judgments in order to produce reliable scale values and thus reliable probability estimates. The analysis presented in this appendix suggests that this problem is much less severe than anticipated. In fact, quite stable estimates may be produced with ten or fewer experts. The following analysis is based on the paired comparison procedure, but because of the similarity of the model underlying the ranking/rating procedure, similar results could be obtained for it.

One of the key factors in determining how many judgments are required to produce a given level of variability in the estimates is how cells corresponding to complete agreement in the data matrix, R, which has as entries r_{jk}, the proportions of times event j is judged more likely than event k, are handled in the analysis. Thurstone and others (e.g., Torgerson, 1958) suggest that such cells be treated as blanks and basically ignored in arriving at scale values. This suggestion appears to cause a loss of considerable very diagnostic data. Thus, this appendix also suggests some possible alternative ways of handling these cells and shows that use of these alternatives can significantly reduce the variability of scale values.

B.1 Number of Experts

We assume that a relevant population of experts exists who might be asked to judge the relative likelihood of various human errors. In this population, for any pair of events j and k, a proportion, r_{jk}, would indicate event j is more likely than event k($1-r_{jk} = r_{kj}$). For any particular assessment of HEPs, a sample of experts from the population is selected randomly. (This is, of course, not exactly true in practice, but at this point we have no reason to believe the experts to be actually used are significantly different from the relevant population.) In applying the method, we wish to obtain estimated scale values, s'_k, for each event k that are as close as possible to the true values s_k that

would be obtained if the entire population of experts provided judgments, and we want as little variability in these estimates as possible. Thus, for a given sample of experts, we want to determine the distribution of the variable s'_k, the estimated (observed) scale value.

Using the LCJ, for n events

$$s'_k = \frac{1}{n} \sum_{j=1}^{n} x'_{jk}$$

where $j \neq k$,

with $x'_{jk} = F^{-1}(r'_{jk})$, where r'_{jk} is the proportion of the sample who judged event j to be more likely than event k, and F is the cumulative distribution function for the normal distribution:

$$F(z) = \int_{-\infty}^{z} (1/\sqrt{2\pi})(\exp(-t^2/2))dt.$$

Assuming independence among experts (an assumption implicit in the LCJ), r'_{jk} is distributed binomially. Thus, we can obtain the distribution of $F^{-1}(r'_{jk})$, and since $s'_k = \frac{1}{n} \sum_{j=1}^{n} F^{-1}(r'_{jk})$, the variance of s'_k, $V(s'_k) = \frac{1}{n^2} \sum_{j=1}^{n} V(F^{-1}(r'_{jk}))$, assuming independence of the distributions of r'_{jk}. (This assumption is discussed below.) In practice, since $F^{-1}(r'_{jj})$ is 0, there are actually only n-1 variances in the sum. Therefore, since the multiplier is arbitrary, it makes sense to use

$$s'_k = \frac{1}{n-1} \sum_{j=1}^{n} s'_{jk}, \text{ and}$$

$$V(s'_k) = \frac{1}{(n-1)^2} \sum_{j=1}^{n} V(F^{-1}(r'_{jk}))$$

for estimates of s'_k and its variance. Table B-1 provides expected values of the mean (μ) and variance (σ^2) of estimated scale values for various values of r_{jk} (population proportion) and m (number of judges). From the fact that r'_{jk} is distributed <u>binomially</u>, this means that:

$$\mu = \sum_{x=0}^{m} F^{-1}(r'_{jk})P(x|r_{jk})$$

$$\sigma^2 = \sum_{x=0}^{m} [F^{-1}(r'_{jk})]^2 P(x|r_{jk}) - \mu^2$$

where

TABLE B-1

Mean and Variance Values for Different
Population Proportions and Number of Judges

Population Proportion

Number of Judges m		r_{jk} = .5	.6	.7	.8	.9
2	μ	0	.2z	.4z	.6z	.8z
	σ^2	$.5z^2$	$.48z^2$	$.42z^2$	$.32z^2$	$.18z^2$
3	μ	0	.152z + .062	.316z+.108	.504z +.124	.728z+.093
	σ^2	$.25z^2$+ .139	$.257z^2$-.019z+.129	$.27z^2$-.06z+.105	$.266z^2$-.125z+.073	$.2z^2$-.135z+.041
4	μ	0	.104z+.130	.232z+.227	.408z+.259	.656z+.194
	σ^2	$.125z^2$+.228	$.144z^2$-.027z+.211	$.194z^2$-.105z+.17	$.245z^2$-.211z+.131	$.226z^2$-.255z+.097
5	μ	0	.068z+.182	.166z+.323	.327z+.378	.590z+.292
	σ^2	$.063z^2$+.261	$.083z^2$-.025z+.239	$.143z^2$-.107z+.198	$.221z^2$-.247z+.167	$.243z^2$-.344z+.152
6	μ	0	.043z+.219	.117z+.396	.262z+.478	.531z+.384
	σ^2	$.031z^2$+.262	$.049z^2$-.019z+.244	$.104z^2$-.093z+.207	$.193z^2$-.250z+.189	$.249z^2$-.408z+.202
7	μ	0	.026z+.243	.082z+.450	.210z+.561	.478z+.471
	σ^2	$.016z^2$+.248	$.029z^2$-.013z+.234	$.076z^2$-.074z+.205	$.166z^2$-.236z+.200	$.249z^2$-.451z+.245
8	μ	0	.016z+.257	.058z+.486	.168z+.627	.430z+.551
	σ^2	$.008z^2$+.227	$.017z^2$-.008z+.219	$.054z^2$-.056z+.196	$.140z^2$-.211z+.201	$.245z^2$-.474z+.273
9	μ	0	.010z+.264	.04z+.510	.134z+.681	.387z+.624
	σ^2	$.004z^2$+.205	$.010z^2$-.005z+.202	$.039z^2$-.041z+.189	$.116z^2$-.184z+.199	$.237z^2$-.483z+.296
10	μ	0	.006z+.269	.028z+.528	.107z+.724	.349z+.691
	σ^2	$.002z^2$+.185	$.006z^2$-.003z+.184	$.027z^2$-.03z+.177	$.096z^2$-.155z+.191	$.227z^2$-.482z+.310
$F^{-1}(r_{jk})$		0	.26	.53	.84	1.28

$$P(x|r_{jk}) = \binom{m}{x} r_{jk}^{x}(1-r_{jk})^{m-x}$$
$$r'_{jk} = x/m,\ x=0,1,2,\ldots m$$

and $F^{-1}(r'_{jk})$ is the inverse of the cumulative normal distribution.

Example: $m = 5$, $r_{jk} = .6$

It is easiest to determine μ and σ by constructing a table. Note that $F^{-1}(0) = -z$ and $F^{-1}(1) = z$.

x	r'_{jk}	$F^{-1}(r'_{jk})$	$P(x\|r=.6)$
0	0/5	-z	$\binom{5}{0} .6^0 .4^5 = .0102$
1	1/5	-.84	$\binom{5}{1} .6^1 .4^4 = .0768$
2	2/5	-.25	$\binom{5}{2} .6^2 .4^3 = .2304$
3	3/5	.25	$\binom{5}{3} .6^3 .4^2 = .3456$
4	4/5	.84	$\binom{5}{4} .6^4 .4^1 = .2592$
5	5/5	z	$\binom{5}{5} .6^5 .4^0 = .0778$

Then:

$$\mu = -z(.0102)-.84(.0768)-.25(.2304)+.25(.3456)+.84(.2592)+z(.0778)$$
$$= .0676z+.1820.$$

$$\sigma^2 = (-z)^2(.0102)+(-.84)^2(.0768)+(-.25)^2(.2304)+(.25)^2(.3456)+(.84)^2(.2592)+(z)^2(.0778)-(.0676z + .1820)^2$$
$$= .083z^2 - .025z + .239.$$

Table B-1 provides values for μ and σ^2 for up to 10 experts. With more experts, these calculations become very tedious and are really practical only with a computer program. In this table z is the value assigned to $F^{-1}(1)$, those cells where all experts agree that one event is more likely than another. Thus, we see the relationship between how these cells are handled and the variability of estimates of scale values. This issue is discussed more fully below.

Before continuing further with the analysis, it is useful to address the issue of independence mentioned above. What this assumption states is that knowledge of event j's relationship with events other than event k

provides no extra information about r'_{jk}. For example, with four events, does knowledge that $r'_{41} = .8$ and $r'_{42} = .1$ provide any information about r'_{43}? It seems that it does not, although it does provide information about r'_{21}. However, r'_{21} is not part of the sum for s'_4, so it appears that the independence assumptions holds.

Approximate confidence intervals for s'_k can now be derived. Since for each k, all r'_{jk} are considered to be independent samples from the same distribution, the standard error of the estimate s'_k is equal to $\sqrt{V(F^{-1}(r'_{jk})/n}$, where n is the number of events. Because of the independence assumption, and the central limit theorem (e.g., Breiman, 1968, p. 186), the distribution of estimates of s_k can be assumed to be approximately normal. Since $V(F^{-1}(r'_{jk}))$ depends on the exact value of r'_{jk} and cannot be known a priori, we examine the maximum value this variance can take under various conditions. Using the central limit theorem, we conclude that at least 95 percent of the estimates of s'_k will be in the interval $(s'_k - 2\sqrt{\max V(F^{-1}(r'_{jk}))/n},\ s'_k + 2\sqrt{\max V(F^{-1}(r'_{jk}))/n})$. Obviously, the confidence will be higher than 95 percent when less than the maximum variance is achieved.

B.2 Handling Complete Agreement

This analysis has suggested some points about the choice of a value of z. Specifically, note in Table B-1 that for most reasonable values of z, the means of the $F^{-1}(r'_{jk})$ are higher than the actual value of $F^{-1}(r_{jk})$. Thus, the resulting estimates of s'_k are biased, particularly for small m. One approach to the selection of z might be to minimize the bias in these estimates. Unfortunately, there is no one value of z that will produce unbiased estimates for all x'_{jk}, for any given m. Some values of z, however, are better in this respect than others. Table B-1 indicates that as m increases the multiplier of z ($r^m-(1-r)^m$) decreases. Thus, for example with m=10 and r_{jk}=.6, the formula is .006z + .269, which is very insensitive to small changes in z. The closer r_{jk} is to 1.0 the more sensitive the estimate becomes to z, suggesting that we might want to minimize the bias for large values of r_{jk} which are sensitive to the value of z. We cannot, however, minimize the bias for $F^{-1}(1) = \infty$, so the value of r_{jk} at which we minimize bias should be the highest possible non-unity value in Table B-1. For example, with m=10, we would minimize the bias at r_{jk}=.9. Using this guide, Table B-2 results.

TABLE B-2
Values of z Minimizing Bias for Maximum r_{jk}

m	2	3	4	5	6	8	10
max r_{jk}	.50	.67	.75	.80	.83	.875	.90
z	-	1.29	1.35	1.41	1.48	1.64	1.69

This rule for selecting z produces, for example with m=8, the maximum variances shown in Table B-3.

TABLE B-3
Variances for m=8 Using Values of z to Minimize Bias

r_{jk}	.5	.6	.7	.8	.9
σ^2	.25	.25	.25	.23	.15
σ	.50	.50	.50	.48	.39

A second approach to selecting values for z is to consider the series of possible values of $F^{-1}(r'_{jk})$ for a particular m and to use a polynomial expansion to find the next value in the series. Because of symmetry, we are concerned only with values of r'_{jk} equal to or above .5. For example, for m=8 the series of possible values of r'_{jk} is .5, .625, .75, .875, and the series of $F^{-1}(r'_{jk})$ is (approximately) 0, .32, .68, 1.15. The next value in this series can be found by first taking successive differences.

0 .32 .68 1.15

.32 .36 .47

.04 .11

.07

Then, the next value is found by summing the last values in each row, i.e., z = 1.15 + .47 + .11 + .07 = 1.80. Other values for z derived by this method are given in Table B-4.

TABLE B-4
Values for z Derived Using a Polynomial Series

m	2	3	4	5	6	8	10
z	-	-	1.35	1.43	1.61	1.80	1.96

Note that these values tend to be somewhat higher than those derived using the minimizing bias guideline. The effect of this approach is to increase the maximum variances slightly. For example, with m=8 the variances are given in Table B-5.

TABLE B-5
Variances for m=8 Using Values of z from Polynomial Expansions

r_{jk}	.5	.6	.7	.8	.9
σ^2	.25	.26	.27	.27	.21
σ	.50	.51	.52	.52	.46

The maximum variance has increased from .25 to .27. Examples of other z values obtained using this method include z = 2.395 for m=20 and z = 2.85 for m=100.

An additional possibility for finding z is to use the formula:

$$F^{-1}\left[\frac{1}{2(m+1)}\right].$$

Resulting z values are:

m	2	3	4	5	6	8	10
z	-	1.15	1.28	1.39	1.47	1.59	1.70

These z values are close to those shown in Table B-2. An advantage of this method is that the z value is easy to calculate with large m.

B.3 Implications

Table B-6 shows the maximum variances and standard deviations, across populations proportions, for various numbers of experts using z values from Table B-4. From the discussion above, the standard error of the estimate of s_k is $\sigma/\sqrt{n}$, where n is the number of events; and the 95 percent confidence limits on s'_k are $s'_k \pm 2\sigma\sqrt{n}$. These values are shown in Table B-7 for various levels of m and n.

TABLE B-7

Maximum 95 Percent Confidence Limits on Scale Values for Various Number of Experts and Events

Number of Events	Number of Experts				
	4	5	6	8	10
3	±.780	±.721	±.675	±.605	±.584
4	±.675	±.624	±.585	±.524	±.506
5	±.604	±.558	±.523	±.469	±.453
6	±.551	±.510	±.478	±.428	±.413
7	±.510	±.472	±.442	±.396	±.382
8	±.477	±.441	±.414	±.370	±.358
9	±.450	±.416	±.390	±.349	±.337
10	±.427	±.395	±.370	±.331	±.320
15	±.349	±.322	±.302	±.271	±.261
20	±.302	±.279	±.262	±.234	±.226

TABLE B-6
Maximum Variances and Standard Deviations for Various Numbers of Experts

Number of Experts	4	5	6	8	10
σ^2	.456	.390	.342	.275	.256
σ	.675	.624	.585	.524	.506

These confidence limits on s'_k can be converted to limits on probabilities by recalling that scale values are transformed into probabilities by $\log p_k = as'_k + b$. With this transformation, the standard error for $\log p_k$ is a times the standard error for s'_k, so the confidence limits for $\log p_k$ are simply the confidence limits for s'_k multiplied by a. Thus, in order to have some a priori idea about uncertainty limits for HEP estimates, we need estimates of a. These can be obtained by noting that a can be estimated as the ratio of the range of log p's to the range of scale values. We know that the maximum range of scale values is 2z, which occurs if there are two events about which all experts agree; one that is more likely than all other events, and one that is less likely than all other events. Table B-4 indicates that this range may be between 2.7 and 4.0 for a reasonable number of experts (ten or less). Since these extremes are unlikely a more reasonable lower bound may be about 1.5, e.g., scale values ranging from -.75 to +.75.

The range of log p will, of course, depend upon the events being considered. Here, to provide a wide variety of possibilities, we assume it can vary between 1.0 and 6.0. Thus, the maximum ratio of ranges, a, that can be obtained is 6./1.5 = 4.0; and the minimum is 1./4. = .25. The maximum 95 percent uncertainty bounds, $\pm B$ for log p for various confidence limits on scales values (from Table B-7), that depend on the number of experts and events, can then be calculated:

$$B = aC,$$

where a is the maximum ratio of ranges defined above and C is the confidence limits on scales values (e.g., from Table B-7). These bounds, because they are on log p, are expressions of the order of magnitude of the uncertainty regarding the HEP.

This procedure can be used to help determine the number of experts required to achieve certain uncertainty bounds. For example, assume ten events are to be considered. The first step is to determine what uncertainty bounds, B, will be acceptable for the HEPs. Suppose this is

determined to be $\pm$.5 on log HEP -- plus or minus one half order of magnitude will include 95 percent of the estimates. Next, a rough estimate of the constant a -- the ratio of the range of log HEPs to the range of scale values -- must be made. A rough guess might be that the HEPs will vary from .1 to .0005, which is a range of approximately 2.25 in logarithms. Furthermore, a conservative estimate of the range of scale values is 1.5 -- suggesting that the experts will not discriminate among the events too well. Then, the estimate of a is 2.25/1.5 = 1.5. Now, using the values of B=.5 and a=1.5, B=aC is solved for C=.333.

Next Table B-7 is used to find how many experts will be required to achieve bounds of $\pm$.333 for the scale values. In Table B-7, the row for ten events is found, and going across this row, the first bounds below $\pm$.333 are found. In this case, the minimum number of experts is eight. Note that the same bounds could also be achieved with 15 events and five experts.

In making these decisions regarding the number of experts (and events) to use to achieve a certain level of uncertainty bounds, it should be noted that, for the most part, these estimates are conservative. The variances used here are maximum variances, which often will not be achieved in practice, so actual bounds may be somewhat smaller.

We note again that although the analysis described here is specific to the paired comparison procedure, because of the similarity of the underlying judgmental models, the results can also be expected to apply approximately to the ranking/rating procedure, implying that the same number of experts can be used.

APPENDIX C

TEST OF SIGNIFICANCE OF COEFFICIENT OF CONCORDANCE

Values of the coefficient of concordance, W, that are significant at the .05 level are given in Table C-1. If the value of W for the given number of experts and events is above the corresponding value in the table, there is basic agreement among the experts. The values in Table C-1 were calculated from table R in Siegel (1956).

Table C-1

Values of Coefficient of Concordance for .05 Significance Level

Number of Experts	Number of Events				
	3	4	5	6	7
3	-	-	.716	.660	.624
4	-	.619	.552	.512	.484
5	-	.501	.449	.417	.395
6	-	.421	.378	.351	.333
8	.376	.318	.287	.267	.253
9	.333	-	-	-	-
10	.300	.256	.231	.215	.204
12	.250	-	-	-	-
14	.214	-	-	-	-
15	.200	.171	.155	.145	.137
16	.187	-	-	-	-
18	.166	-	-	-	-
20	.149	.129	.117	.109	.103

Distribution

U.S. NRC Distribution Contractor (CDSI) (420)
7300 Pearl Street
Bethesda, MD 20014
395 copies for AN,RX
25 copies for NTIS
242 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

Ms. Barbara Jean Bell
Batelle Columbus Laboratories
505 King Avenue
Columbus, OH 43201

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

Dr. George J. Boggs
Telenet Technical Center
GTE Laboratories
40 Sylvan Road
Waltham, MA 02154

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

Dr. Mark Brecht
4350 West 136 Street
Hawthorne, CA 90250

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
12 Running Fox Road
Glen Arm, MD 21057

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

Dr. Gordon Clark
Dept. of Industrial and Systems Engineering
The Ohio State University
1971 Neil Avenue
Columbus, OH 43210

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

Ms. Kay Comer
Senior Reliability Engineer
General Physics Corporation

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. 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
Director
Human Reliability Associates, Ltd.
1 School House
Higher Lane, Dalton, Parbold
Lanc. WN8 7RP
England

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

Dr. George Flanagan
Engineering Physics Division
Building 6025
Oak Ridge National Laboratory
Mail Stop 09W
Oak Ridge, TN 37830

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

Mr. 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 TW11OLN
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. Paul Haas
Building 6025
Oak Ridge National Laboratory
P. O. Box X
Oak Ridge, TN 37830

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

Dr. Edward J. Kozinsky
Manager, Systems Projects
General Physics Corporation
1 Northgate Park, #201
Chattanooga, TN 37415

Harmen Kragt, M.Sc.
Univ. of Technology Eindhoven
P. O. Box 513
5600 MB Eindhoven
The Netherlands

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. Randall W. Pack
General Physics Corporation
Suite 120
1770 The Exchange
Atlanta, GA 30339

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

Dr. Thomas G. Ryan (15)
Human Engineering Section
Human Factors Branch
Division of Facility Operations
Office of Nuclear Regulatory Research
Mail Stop - Nicholson Lane
U.S. Nuclear Regulatory Commission
Washington, DC 20555

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

Mr. Franz Schneider
Ergonomics Unit, Central Safety Services
Ontario Hydro
757 MacKay Road
Pickering, Ontario L1W 3C8
Canada

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

Mr. Jeffrey P. Schwartz
4628 West Frankfurt Drive
Rockville, MD 20853

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)
Director
Planning and Evaluation Division
The Maxima Corporation
7315 Wisconsin Avenue
Suite 900N
Bethesda, MD 20014

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 (20)
Nuclear Safety Division
OECD Nuclear Energy Agency
38, Boulevard Suchet
F-75016 Paris
France

Ms. Catherine Stewart
Human Factors
PRW, Ballistic Missiles Division
513/313
Norton Air Force Base
San Bernadino, CA 92402

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

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

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

Internal Sandia Distribution

7223 B. H. Finley
7223 D. P. Miller
7223 R. R. Prairie
7223 L. M. Weston (40)
3141 L. J. Erickson (5)
3151 W. L. Garner (3)
8214 M. A. Pound

NRC FORM 335 (7-77) U.S. NUCLEAR REGULATORY COMMISSION

BIBLIOGRAPHIC DATA SHEET

1. REPORT NUMBER *(Assigned by DDC)*
NUREG/CR-2743, SAND82-7054

4. TITLE AND SUBTITLE *(Add Volume No., if appropriate)*
Procedures for Using Expert Judgment to Estimate Human Error Probabilities in Nuclear Power Plant Operations

2. *(Leave blank)*

3. RECIPIENT'S ACCESSION NO.

7. AUTHOR(S)
David A. Seaver and William G. Stillwell

5. DATE REPORT COMPLETED
MONTH: March | YEAR: 1983

9. PERFORMING ORGANIZATION NAME AND MAILING ADDRESS *(Include Zip Code)*
Decision Science Consortium, Inc.
7700 Leesburg Pike, Suite 421
Falls Church, VA 22043

DATE REPORT ISSUED
MONTH | YEAR

6. *(Leave blank)*

8. *(Leave blank)*

12. SPONSORING ORGANIZATION NAME AND MAILING ADDRESS *(Include Zip Code)*
Division of Facility Operations
Office of Nuclear Regulatory Research
U.S. Nuclear Regulatory Commission
Washington, D.C. 20555

10. PROJECT/TASK/WORK UNIT NO.

11. CONTRACT NO.
A-1188

13. TYPE OF REPORT
Technical Report - Formal

PERIOD COVERED *(Inclusive dates)*
February 1981 - March 1983

15. SUPPLEMENTARY NOTES

14. *(Leave blank)*

16. ABSTRACT *(200 words or less)* This report describes and evaluates several procedures for using expert judgment to estimate human error probabilities (HEPs) in nuclear power plant operations. These HEPs are currently needed for several purposes, particularly for probabilistic risk assessments. Data do not exist for estimating these HEPs, so expert judgment can provide these estimates in a timely manner. NUREG/CR-2255 suggested that expert judgment can provide reasonably valid and reliable HEP estimates, if used carefully and systematically. Five judgmental procedures are described here: paired comparisons, ranking and rating, direct numerical estimation, indirect numerical estimation and multiattribute utility measurement. These procedures are evaluated in terms of several criteria: quality of judgments, difficulty of data collection, empirical support, acceptability, theoretical justification, and data processing. Situational constraints such as the number of experts available, the number of HEPs to be estimated, the time available, the location of the experts, and the resources available are discussed in regard to their implications for selecting a procedure for use. Details for implementing the procedures and necessary calculations are included in appendices. These descriptions will be the basis of subsequent research testing the use of several procedures.

17. KEY WORDS AND DOCUMENT ANALYSIS
17a. DESCRIPTORS
XXXXXXXXXXXX
psychological scaling
probability assessment
expert opinion
human error probability
paired comparisons
ranking
rating
direct numerical estimation
indirect numerical estimation
multiattribute utility measurement

17b. IDENTIFIERS/OPEN-ENDED TERMS

18. AVAILABILITY STATEMENT
Unlimited

19. SECURITY CLASS *(This report)*
Unclassified

21. NO. OF PAGES
104

20. SECURITY CLASS *(This page)*
Unclassified

22. PRICE
S