NUREG/CR–3688/1 of 2
SAND84–7115
RX
Printed November 1984
CONTRACTOR REPORT

Generating Human Reliability Estimates Using Expert Judgment

Volume 1. Main Report

M. K. Comer,* D. A. Seaver,** W. G. Stillwell,** and C. D. Gaddy*

*General Physics Corporation
10650 Hickory Ridge Road
Columbia, Maryland 21044

**The Maxima Corporation
7315 Wisconsin Avenue, Suite 900N
Bethesda, Maryland 20814

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

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

Previous Reports in This Series

Expert Estimation of Human Error Probabilities in Nuclear Power Plant Operations: A Review of Probability Assessment and Scaling, NUREG/CR-2255, May 1982.

Procedures for Using Expert Judgment to Estimate Human Error Probabilities in Nuclear Power Plant Operations, NUREG/CR-2743, March 1983.

Abstract

The U.S. Nuclear Regulatory Commission is conducting a research program to determine the practicality, acceptability, and usefulness of several different methods for obtaining human reliability data and estimates that can be used in nuclear power plant probabilistic risk assessments (PRA). One method, investigated as part of this overall research program, uses expert judgment to generate human error probability (HEP) estimates and associated uncertainty bounds. The project described in this document evaluated two techniques for using expert judgment: paired comparisons and direct numerical estimation. Volume 1 of this report provides a brief overview of the background of the project, the procedures for using psychological scaling techniques to generate HEP estimates and conclusions from evaluation of the techniques. Volume 2 provides detailed procedures for using the techniques, detailed descriptions of the analyses performed to evaluate the techniques, and HEP estimates generated as part of this project.

The results of the evaluation indicate that techniques using expert judgment should be given strong consideration for use in developing HEP estimates. Judgments were shown to be consistent and to provide HEP estimates with a good degree of convergent validity. Of the two techniques tested, direct numerical estimation appears to be preferable in terms of ease of application and quality of results. The fact remains, however, that actual relative frequencies of errors are not available, so predictive validity against such a criterion has not been established. In the absence of such data, and given the practical advantages such as the time and cost of using expert judgment, this approach appears to be a feasible way to obtain needed HEP estimates for PRAs or other uses. In addition, HEP estimates for 35 tasks related to boiling water reactors (BWRs) were obtained as part of the evaluation. These HEP estimates are also included in the report.

CONTENTS

CONTENTS (continued)

CONTENTS (continued)

CONTENTS (continued)

CONTENTS (continued)

LIST OF TABLES

Page

LIST OF FIGURES

Page

ACKNOWLEDGMENTS

The work reported in this document was performed under Contract No. 37-7045 to Sandia National Laboratories (SNL). It is one of several SNL projects sponsored by the Office of Nuclear Regulatory Research, U.S. Nuclear Regulatory Commission (NRC).

The authors would like to express their appreciation to the following sponsors: Dr. Thomas G. Ryan, the NRC project monitor; Dr. Louise M. Weston, the SNL technical monitor; and Dr. Robert Easterling and Dr. Alan D. Swain of SNL for their technical reviews and assistance.

We would also like to thank Mr. Joseph N. Zerbo of General Physics Corporation for his valuable contributions related to task descriptions and the data collection effort; Mr. Stephen P. Clark for his assistance with computer programming; Dr. Julien M. Christensen, the Project Director; and the 19 General Physics instructors who participated in the data collection. Their participation and valuable insights played a major part in the project's success.

ABBREVIATIONS

ADS	automatic depressurization system
BWR	boiling water reactor
HEP	human error probability
HPCI	high pressure coolant injection
LER	Licensee Event Report
NRC	U.S. Nuclear Regulatory Commission
PRA	probabilistic risk assessment
PSF	performance shaping factor
PWR	pressurized water reactor
RCIC	reactor core isolation cooling
SLIM-MAUD	Success Likelihood Index Methodology - Multiattribute Utility Decomposition
SNL	Sandia National Laboratories

GENERATING HUMAN RELIABILITY ESTIMATES USING EXPERT JUDGMENT

1. INTRODUCTION

As more and more attention has been focused on assessing the risks associated with nuclear power plants, it has become clear that estimates of human reliability are needed as components of this assessment. The U.S. Nuclear Regulatory Commission (NRC) is supporting research to meet this need. Individual research projects that are being sponsored by the NRC in the area of human reliability data include: (1) techniques for using expert judgment, (2) nuclear power plant simulator experiments, (3) computer modeling of human performance, and (4) use of Licensee Event Report (LER) calculations. Research is being conducted in each of these areas to determine the practicality, acceptability, and usefulness of generating human reliability data that can be used in nuclear power plant probabilistic risk assessments (PRAs).

In the area of using expert judgment to generate human reliability estimates, the NRC has sponsored a multiyear program with Sandia National Laboratories (SNL) to determine whether any applicable techniques currently exist, to develop detailed procedures for using those techniques, and finally to conduct an empirical evaluation of the techniques. The results of a literature review are reported in NUREG/CR-2255, "Expert Estimation of Human Error Probabilities in Nuclear Power Plant Operations: A Review of Probability Assessment and Scaling" (Stillwell, Seaver, and Schwartz, 1982). The detailed procedures for using the techniques are contained in NUREG/CR-2743, "Procedures for Using Expert Judgment to Estimate Human Error Probabilities in Nuclear Power Plant Operations" (Seaver and Stillwell, 1983). The results of the evaluation as well as revised procedures are reported in this document. Seaver and Stillwell (1983) described procedures for using five different techniques. For reasons that are described in Section 2, this project evaluated two of those techniques: paired comparisons and direct numerical estimation. Each is described more fully in Section 3. In addition, comparisons of the results from using paired comparisons and direct numerical estimation were made with results from simulator experiments (Beare et al., 1984) and NUREG/CR-1278, "Handbook of Human Reliability Analysis with Emphasis on Nuclear Power Plant Applications" (Swain and Guttmann, 1983). The results of these comparisons are discussed in Section 4 of this volume and described in detail in Appendix B of Volume 2.

1.1 Purpose

The primary purpose of the study was to evaluate the use of expert judgment (psychological scaling techniques) to estimate probable operator errors for certain tasks. Also, the study investigated if HEP estimates could be generated using the scaling techniques and evaluated the practicality, acceptability, and usefulness of each technique. To ensure a thorough evaluation, two sets of issues were developed early in the project. The first set, Program Issues, relates to whether psychological scaling has a potential role in developing HEP estimates for PRA. Table 1 lists the six Program Issues that were addressed in the test and evaluation of the techniques. The second set, Technical Issues, relates to how to implement psychological scaling techniques if they are shown to have sufficient potential. Table 2 lists the Technical Issues. More thorough discussions of these issues and how they were evaluated are contained in Section 4 of this volume and in Appendix B of volume 2.

Table 1 Program issues

P1. Do psychological scaling techniques produce consistent judgments from which to estimate HEPs?

P2. Do psychological scaling techniques produce valid HEP estimates?

P3. Can the data collected using psychological scaling techniques be generalized?

P4. Are the HEP estimates that are generated from psychological scaling techniques suitable for use in PRAs and for entry into the Human Reliability Data Bank as described in NUREG/CR-2744 Volume 2 (Comer et al. 1983)?

P5. Can psychological scaling procedures be used by persons who are not expert in psychological scaling to generate HEP estimates?

P6. Do the experts used in the psychological scaling process have confidence in their ability to make the judgments?

Table 2 Technical issues

T1. Based on measures of consistency and comparisons with other human reliability estimates, is there any difference in the quality of estimates obtained from the two techniques?

T2. Is there any difference in the results based on the type of task that is being judged?

T3. Do education and experience have any effect on the experts' judgments?

T4. Based on the number of probability estimates and the functional relationship between the paired comparison scale and the probability scale, how should the paired comparison scale be calibrated into a probability scale?

T5. Can reasonable uncertainty bounds be estimated judgmentally?

1.2 Summary of Findings

The results of the evaluation as they relate to the issues that were listed in Tables 1 and 2 are:

- Both psychological scaling techniques more than met statistical requirements for consistency (P1).

- Both psychological scaling techniques tested were shown to have convergent validity in producing HEP estimates. It should be noted though, that predictive validity with respect to HEP estimates based on the actual relative frequency of errors could not be established because of the lack of such estimates. (This will be a difficulty in validating any procedure used to estimate HEPs.) (P2)

- The estimates generated using these techniques should be generalizable to all boiling water reactors (BWRs) (P3).

- The HEP estimates generated from these techniques are suitable for use in PRAs and for entry into the Human Reliability Data Bank (P4).

- Psychological scaling procedures can be used by persons who are not expert in psychological scaling to generate HEP estimates. Procedures described in Appendix A of Volume 2 require no special skills on the part of the person administering the data collection, only that the administrator be able to understand the instructions. However, special skills are needed to write the task statements and analyze the data. These personnel requirements are also described in Volume 2, Appendix A (P5).

- The experts used in the psychological scaling process did not have strong confidence in their ability to make the judgments. However, this lack of confidence did not affect the consistency of judgments (P6).

- Essentially no differences in measures of consistency and convergent validity were obtained from the two techniques (T1).

- There were essentially no differences in measures of consistency and convergent validity based on the type of task being judged. The techniques can be used to estimate HEPs for either Level 1 (systems) tasks or Levels 2 and 3 (components) tasks as defined in Section 4.2.2 of this volume (T2).

- Education, experience, and type of license certification did not affect the judgments of this relatively homogeneous group of experts (T3).

- The paired comparison scale should be calibrated into a probability scale using a logarithmic relationship (T4).

- Uncertainty bounds can be estimated judgmentally (T5).

A more detailed discussion of the results of the evaluation is contained in Section 4 of this volume and Appendix B of Volume 2. The actual HEP estimates that were obtained as part of the evaluation are given in Appendix C of Volume 2.

1.3 Organization of Report

This document contains four major parts: the main report; Appendix A - Instructions for the Use of Psychological Scaling Techniques; Appendix B - Evaluation Results; and Appendix C - Human Error Probability Estimates. The main report is contained in Volume 1 and the appendices are contained in Volume 2.

The main report includes an overview of psychological scaling techniques: what they are, how they are used, and how they were evaluated in this project. The intent of the main report is to provide a short, concise description for the layman.

Appendix A contains detailed procedures and step-by-step calculations for using two types of psychological scaling techniques: paired comparisons and direct numerical estimation. This appendix can be used as a stand-alone reference by anyone wishing to generate estimates with one or both techniques.

Appendix B contains a detailed description of the evaluation that was conducted for paired comparisons and direct numerical estimation. An explanation of the test methods is provided as well as a description of the results of the evaluation. Since some of the methods for evaluating the data and the techniques require some statistical knowledge, Appendix B is written primarily for those with an understanding of statistics who are interested in the details of how the evaluation was conducted.

Finally, Appendix C presents the human reliability estimates that were obtained as part of this project. This appendix is intended to be used by those who have an interest in or need for HEP estimates and the associated uncertainty bounds.

2. PSYCHOLOGICAL SCALING

Psychological scaling is the process of assigning numbers to objects, events, or their properties, in such a fashion that the numbers represent the relationships among them. In the present context, we are interested in scaling the likelihood of human error on a probability scale. Psychological scaling provides the mechanism to systematize the process of obtaining expert judgments regarding the likelihood of human error, and to transform these judgments into probabilities that meet the needs of various users such as PRA practitioners. It does so by both defining the judgments that are required and by providing a formal procedure for deriving probability estimates from these judgments.

2.1 Overview of Psychological Scaling Techniques

Numerous psychological scaling techniques have been developed based on various types of judgments and various procedures for deriving numerical scales from the judgments (Torgerson, 1958). Earlier efforts reviewed the applicability of using expert judgment to estimate probabilities and found successful applications in areas such as weather forecasting, intelligence analysis, and medical diagnosis (Stillwell, Seaver, and Schwartz, 1982). Based on this review, five techniques were selected for further consideration (Seaver and Stillwell, 1983). Four of these are briefly described here: paired comparison, ranking/rating, direct numerical estimation, and indirect numerical estimation. The fifth technique, SLIM-MAUD (Success Likelihood Index Methodology - Multiattribute Utility Decomposition), based on multiattribute utility theory, is being investigated by Brookhaven National Laboratory as part of NRC's research program and is not described here (Embrey, 1983; Embrey et al., 1984).

2.1.1 Paired Comparison Technique

Paired comparison scaling is based on judgments of the type "task a is more likely to be performed than task b." This technique assumes that each task is represented psychologically in the judge's mind by a distribution on a subjective, psychological scale. When two tasks are compared, each produces a value selected randomly from its subjective distribution, and the task with the higher value is judged to be more likely (Thurstone, 1927). By making certain assumptions about the subjective distributions of the tasks, and by having several judges make all possible paired comparisons for a set of tasks, numerical scale values can be obtained.

The scale that is derived from paired comparisons is a subjective scale and not a probability scale. Therefore, scale values for tasks must be transformed into probability estimates. This process is accomplished by including some tasks, called "anchor tasks," for which independent probability estimates are available so that the probabilities of other tasks along the subjective scale can be estimated. Details of this process are given in Appendix A of Volume 2.

2.1.2 Ranking/Rating Techniques

Although requiring different types of judgments, ranking and rating are included together because the underlying psychological model is the same and the procedures for deriving probability estimates from the judgments are similar. Ranking requires each expert to arrange the tasks under consideration according to their likelihood. Rating requires the expert to rate the likelihood of each task on a particular scale, e.g., from one to seven. Ranking and rating are similar in that each rank can be considered a different rating. These techniques are based on assumptions similar to those for paired comparison scaling involving the psychological representation of tasks by subjective distributions. Ranking and rating each produce scales that must be transformed into probability estimates using the same procedures used for paired comparisons. (See Torgerson, 1958, for additional details.)

2.1.3 Direct Numerical Estimation

The direct numerical estimation technique requires the experts to provide probability estimates for each of the tasks. The estimates of a number of experts are then combined to provide a single probability estimate for each task. Direct numerical estimation can also be used to estimate uncertainty bounds defining the probable range of HEP estimates for varying conditions, such as operator training, plant design, quality of written procedures, etc. Additional information on uncertainty bounds is discussed in Section 4.2.5 of Appendix A, Volume 2.

2.1.4 Indirect Numerical Estimation

This procedure requires experts to make ratio judgments regarding the relative likelihood of pairs of tasks, e.g., "task a is five times as likely to be performed as task b." Each task must be compared with one other task, so that all tasks are linked. For example, with four tasks, a would be compared with b, b with c, and c with d.

To convert these ratios into probabilities, an independent probability estimate for one task is needed. This known probability serves as the measure by which probabilities for the other tasks are estimated. As with the direct estimation technique, the probabilities obtained from individual expert's judgments are then combined to produce a single probability estimate for each task.

2.2 Advantages and Disadvantages of Using Psychological Scaling Techniques

Currently there are several different methods being researched for obtaining human reliability estimates for nuclear reactor operations and maintenance. These methods include:

(1) Conducting experiments using nuclear power plant training simulators to gather data on human performance.

(2) Using computer modeling techniques that have been developed to simulate human performance under a variety of conditions.

(3) Extracting information from the LER system and calculating HEPs.

(4) Generating HEP estimates using psychological scaling techniques.

In addition to these methods, HEP estimates are provided in the Handbook (Swain and Guttmann, 1983).

The advantages of using psychological scaling techniques to generate HEP estimates are primarily ones of cost and convenience. The techniques can be used in plant specific PRAs or for more generic applications. Virtually any task or set of human actions that needs to be quantified for PRA can be, using the techniques previously described in this section. Compared to the other techniques, the cost for obtaining the estimates is low. The personnel and material requirements are specified in detail in Appendix A, but basically the main cost is for the experts who make the judgments.

The chief disadvantage to using psychological scaling techniques to generate HEP estimates is that the result is only a subjective estimate and is not tied to any actual historical events. However, most of the techniques currently being researched also have this same disadvantage.

2.3 Results of Psychological Scaling

Psychological scaling has been used successfully in several applications outside the nuclear power industry. The techniques have many practical advantages over other methods for gathering data for the nuclear power industry. Despite the success of the techniques in other applications, and the potential practical advantages in this context, the results of psychological scaling must be useful. To be useful to the nuclear power industry, psychological scaling must yield acceptable HEP estimates. The techniques must also provide uncertainty bounds which represent the range over which an HEP might vary as conditions in the power plant vary.

The primary objective of the current project was to evaluate psychological scaling to determine whether it could be used to generate human reliability estimates, both HEP estimates and estimates of uncertainty bounds. The two techniques selected for testing in this study were paired comparisons and direct numerical estimation. These two techniques generally represent the extremes in terms of number of experts required and difficulty of judgments.

3. IMPLEMENTATION OF PSYCHOLOGICAL SCALING

The steps required to obtain human error probability estimates using each of two techniques, direct estimates and paired comparisons, are discussed in this section. Figure 1 provides an overview of the steps for implementing the two types of psychological scaling. Details of implementation of each of these techniques are given in Appendix A.

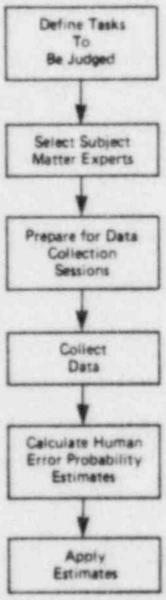

Figure 1 Steps for implementation of psychological scaling.

3.1 Define Tasks To Be Judged

Probably the most critical requirement for the use of judgmental procedures to estimate HEPs is that the tasks to be judged be defined carefully and completely. The more fully the tasks are specified, the less they will be open to variable interpretation by the experts judging their likelihood. The level of detail needed in the tasks will vary depending on the task itself and the ultimate use of the probability estimate. For example, if the task is to start a reactor feed pump at Plant X for a plant-specific PRA, the level of detail will be more specific than if the task is to restore residual heat removal cooling in a boiling water reactor for a research project. The tasks that were used in this project are presented in both Appendix B and Appendix C of Volume 2.

One consideration in defining the tasks is to specify performance shaping factors (PSFs). PSFs are those conditions which affect the performance of a specific operation in a specific situation, such as stress, situational characteristics (e.g. environment), instructions, task and equipment characteristics (e.g. plant layout), and characteristics of the individual (e.g. emotional state). (See Swain and Guttman, 1983, for a detailed discussion.) Most PSFs will be left unspecified and experts will be asked to consider "typical" conditions. Some PSFs may be specified depending on the use of the HEP estimates. In addition, when uncertainty bound estimates are obtained, they are based on possible variations in PSFs.

PRA practitioners should review the tasks but they must also be defined so that they can be easily understood by the subject matter experts. In some instances, the wording chosen by PRA practitioners or researchers is not the same as the wording used by power plant operators. For example, markings to indicate operating ranges on a meter scale may be called "limit marks" by the PRA practitioner and "meter banding" by power plant operators. Thus, tasks should include information needed by PRA practitioners worded so that they are meaningful to the experts. Examples of tasks worded for experts in this project are provided in Attachments 1 and 2 to Appendix B, Volume 2.

3.2 Select Subject Matter Experts

Experts selected must be familiar with the tasks to be judged. For example, if the tasks involve nuclear power plant operations from a control room perspective, an in-depth knowledge of plant systems, operations, and control room procedures is an essential criterion for selection of experts. If the tasks to be judged include accidents or other infrequent events, certified nuclear power plant instructors may be the best judges. Instructors have had the opportunity in nuclear power plant training simulators to witness many different operators and their reactions to simulated accident scenarios. Other types of experts considered for this project were power plant operators, human factors engineers, psychologists, and human reliability analysts. These types of experts were not chosen in favor of certified instructors because the instructors had the most appropriate background for the tasks that were to be judged.

While no exact number of experts can be specified, no fewer than six should be used for direct estimation and at least 10 to 12 should be used with paired comparisons. More experts should be used if at all practical.

3.3 Prepare and Collect Data

Preparation for data collection involves primarily preparing response booklets and instructions as described in Appendix A. Sample instructions have been developed so that data can be collected from experts by someone who is not an expert in psychological scaling. The time required for data collection will depend on the technique used. Experience in this study indicates that both HEP estimates and estimates of uncertainty bounds for 35 tasks can be collected using direct estimation in about 30 minutes per expert, while about 100 paired comparisons can be made in the same amount of time. This number of paired comparisons would provide the data needed to estimate HEPs for about 15 tasks. An additional 15 to 20 minutes will be needed for instructions.

3.4 Calculate HEP Estimates

The procedures for deriving HEP estimates from expert judgments differ for the direct numerical estimation and paired comparison scaling techniques. Each of the procedures is described briefly here and detailed, step-by-step descriptions are provided in Appendix A.

3.4.1 Direct Numerical Estimation

In direct numerical estimation, each expert provides an HEP estimate for each task. They may also be asked to estimate uncertainty bounds. HEP estimates and estimates of uncertainty bounds are then calculated by combining the estimates of individual experts. Table 3 shows an example of the type of data that is obtained and the resulting HEP and uncertainty bound estimates. Details on how individual experts' estimates are combined are provided in Appendix A.

Table 3 Sample of experts' direct estimates for Task 1

Expert	Estimate	Lower Bound	Upper Bound
1	.0010	.0005	.0500
2	.0100	.0001	.0100
3	.0030	.0002	.0200
4	.0004	.0003	.0300
5	.0050	.0004	.0400
HEP (Geometric Mean)	.0023	.0003	.0261

3.4.2 Paired Comparison Scaling

Calculation of human error probabilities based on paired comparisons is more complex than for direct estimates. A table must be constructed for each expert as shown in Table 4. For example, if Expert 1 chose Task 2 as being more likely to occur than Task 1, a "1" is placed in the table in the column for Task 2 and the row for Task 1. The "1" signifies that the task listed across the top of the table was chosen to be more likely than the task listed down the side of the table. As a second example, Expert 1 thought Task 3 was less likely than Task 1, so a "0" is placed in the table.

Once a table is prepared for each expert, the numbers for each pair of tasks or cell in the table are summed for all experts. The result is a table of the number of experts who chose a given task more likely than another task. The numbers in this table of all experts' judgments can then be translated into proportions, e.g., 15 experts out of 20 thought Task 1 was more likely than Task 2, i.e., .75. This table is then converted into a table of normal deviates using tables of normal distributions found in most statistics textbooks.

Table 4 Sample table of expert's paired comparisons

Expert 1:	Task 1	Task 2	Task 3
Task 1	-	1	0
Task 2	0	-	1
Task 3	1	0	-

The columns of this table are summed and an average value is calculated for each task based on judgments of the task relative to all other tasks. The resultant numbers are called scale values. These scale values permit ordering the tasks relative to one another.

At this point, scale values are converted into probability estimates. The tasks for which there are HEP estimates from another source, such as simulator data, can be used to determine what probabilities should be associated with each scale value. For example, assume that Task 1 has an estimated HEP of .01 based on simulator research data, and Task 3 has an estimated HEP of .001. The scale values from paired comparisons show that experts chose Task 1 as more likely than Task 2, and Task 2 as more likely than Task 3. Statistical calculations can be used to determine the HEP estimate for Task 2 because (a) independent probability estimates for Tasks 1 and 3 are known, and (b) the relationships of 1, 2, and 3 are known. A probability for each task can be calculated accordingly. The detailed steps for performing these calculations are presented in Appendix A.

3.5 Other Necessary and Useful Analyses

Prior to the use of any HEP estimates derived from the procedures described above, checks should be made of the consistency of the experts' judgments. Across-expert consistency measures the extent to which the judgments agree. If there is not a minimum level of agreement, the HEP estimates should not be used. The coefficient of concordance, described in Appendix A, provides a measure of across-expert consistency on a zero-to-one scale where 0 indicates no agreement and 1 indicates complete agreement. A statistically significant coefficient of concordance indicates that the consistency is adequate.

For the paired comparison judgments, it is also possible to measure the internal consistency of each expert's judgments. Inconsistent judgments are indicated by intransitive triads of judgments, e.g. Task 1 is judged more likely than Task 2, Task 2 more likely than Task 3, and Task 3 more likely than Task 1. If an expert is not sufficiently consistent, that expert's judgments should not be used in the derivation of HEP estimates. The coefficient of consistency (see Appendix A) provides a measure of internal consistency on a zero-to-one scale with 0 indicating the maximum possible number of intransitive triads and 1 indicating no intransitive triads. Adequate consistency is again determined by the statistical significance of the measure.

In addition to these consistency measures that should always be calculated when using psychological scaling techniques, another analysis can be performed for each technique that is not necessary but does provide useful information. This analysis is the calculation of statistical confidence limits that indicate the amount of statistical variation to be expected in the HEP estimates. This variation represents, for example, the variation that would be expected if the same experts, without remembering their previous responses, or similar groups of experts, made these same judgments many times. Procedures for calculating statistical confidence limits may be found in Appendix A of Volume 2. These limits indicate the probable range of variation for HEP estimates under typical conditions.

3.6 Application of Human Error Probability Estimates Derived From Direct Estimates or Paired Comparisons

HEP estimates derived from direct estimates or paired comparisons can be used in three application areas. First, the HEP estimates can be used to support PRAs. Second, the HEP estimates can be entered into the Human Reliability Data Bank (Comer et al., 1983) for reference by anyone interested in estimating human reliability. A final application would be any type of probabilistic study in which human error is a consideration, e.g., in design.

A specific example of this final application area would be using HEP estimates to assist in the assessment of findings from a human factors control room design review. Human factors problems in a control room could be ranked in terms of their potential contribution to human error based on their associated probabilities. This would assist the control room review team in deciding the priorities of problems for correction.

4. EVALUATION OF PSYCHOLOGICAL SCALING TECHNIQUES

As discussed in Section 1, the project sponsored by the NRC and SNL evaluated two techniques for psychological scaling. Discussions of the issues addressed, the evaluation methods used, and the results obtained are contained in this section.

4.1 Issues

Two sets of issues were developed early in the project as a means of ensuring that all essential aspects of the paired comparison and direct numerical estimation psychological scaling techniques were adequately tested. The two were program issues and technical issues, as were shown in Tables 1 and 2. Each of these issues was categorized as to whether it provided information on practicality, acceptability, or usefulness. These characteristics were used to evaluate the psychological scaling techniques and can be thought of in the following terms:

- Is psychological scaling practical to implement in terms of cost and procedural issues?
- Will the industry accept the techniques as a viable means of acquiring estimates?
- Will government and industry use psychological scaling techniques as part of the PRA process?

Table 5 lists the issues that were considered during the project; identifies the categories of practicality, acceptability, and usefulness for each; and describes the method and type of analysis that were used to address each. One or more of the following methods were chosen to address each issue:

(M1) By survey.

(M2) By conducting a formal experiment.

(M3) Through the use of a demonstration.

The three types of analysis considered were descriptive, quantitative, and comparative. The descriptive type results from observation or experience. The quantitative type results in a numerical resolution of the issue. The comparative type is used to determine the similarities and differences between choices.

Section 4.2 describes the methods that were used to evaluate the techniques, and Section 4.3 presents the results. More detail on both is contained in Appendix B of Volume 2.

Table 5 Issues, methods and analysis

Issue*	Category **	Method ***	Analysis
P1 - Consistency	A	M2	Quantitative
P2 - Validity	A	M2	Quantitative, comparative
P3 - Generalizability	P	M1, M3	Descriptive, comparative
P4 - Human Reliability Data Bank	U	M1, M3	Descriptive, comparative
P5 - Used by nonexperts	P	M3	Descriptive
P6 - Experts' confidence	A	M1, M2	Descriptive, comparative
T1 - Quality of techniques	A	M2	Quantitative, comparative
T2 - Type of task	A	M2	Quantitative, comparative
T3 - Education/experience	A	M1, M2	Quantitative, comparative
T4 - Conversion of paired comparison scale	P	M1	Quantitative, comparative
T5 - Uncertainty bounds	A	M2	Quantitative, comparative

* From Tables 1 and 2

** Practically, acceptability, usefulness

*** Method for test: M1 = survey; M2 = experiment; M3 = demonstration

4.2 Evaluation

Three methods were used to evaluate the psychological scaling techniques in terms of practicality, acceptability, and usefulness. The evaluation was through a survey, experiment, or demonstration. All three methods were combined into a single test design that was fully documented and pilot tested in the early stages of the project. A summary of the test is presented in the following subsections. Complete details on each aspect of the test are contained in Appendix B.

4.2.1 Subject Matter Experts

Nineteen NRC-certified BWR instructors served as subject matter experts. Certified instructors were considered the best subject population because they have had the opportunity to witness many different operators and their reactions to simulated accident scenarios during training. The experts had an average of 11.26 years of experience as power plant instructors or operators.

4.2.2 Task Statements

As discussed in Section 3.1, the tasks chosen and the statements made about the tasks are probably the most critical considerations in the methodology of psychological scaling. Two levels of tasks were defined for use in the evaluation. The tasks corresponded to Level 1 and Levels 2 and 3 as defined by the Human Reliability Data Bank (Comer et al., 1983). Level 1 of the Human Reliability Data Bank structure combined power plant systems with human actions that represented job duties. In this project the Level 1 tasks represented BWR systems and control room operator duties. Level 2 of the data bank structure combined equipment components with human actions defined as tasks. The tasks defined for this project included those associated with control room operators and equipment operators. Level 3 corresponded to controls and displays and task elements. To enable the estimates from this project to be compared with other existing data, the task statements had to be similar. Therefore, task statements from Levels 2 and 3 were developed so that they could be compared with Handbook and simulator tasks.

4.2.3 Materials

Task statements were presented to the experts in four-part, response booklets. Data collection session instructions were prepared. These materials are summarized in this section and described in greater detail in Appendix B to Volume 2.

The first part of the response booklet contained assumptions that applied to the tasks and examples of paired comparisons. Additional pages in the first part of the booklet contained all possible pairs of tasks. Tasks were presented from either Level 1 or Levels 2 and 3. Experts who had first responded to Level 1 tasks responded to Level 2 and 3 tasks second, and vice versa. Thus, the second part of the booklet contained the tasks and associated assumptions the experts had not seen in the first part. Also, each booklet had a random ordering of tasks within each pair and of pairs within each level. These steps were taken to minimize the influence of the order of the tasks on the experts' responses.

The third part of the booklet presented assumptions that applied to the tasks and samples of the direct estimate procedure. Tasks and scales were provided as examples of the type of response expected. Then, each page in the remainder of the third part of the booklet presented a task and a scale.

The fourth and final part of the booklet contained questions about the expert's background, e.g., years of experience. Nine questions requesting respondents' opinions about the ease of using the booklet and the clarity of the task statements concluded the booklet.

Finally, data collection session instructions were prepared. Instructions were written for the data collection session administrator to read before beginning a session. Then, additional instructions to be read to the experts were developed. These instructions included a general overview of the purpose and procedures, and more specific details to be read before paired comparisons and direct estimates were made. Assumptions for Level 1 and Level 2 and 3 tasks were also written on a board in the front of the room to remind the experts.

The response booklets and instructions were pretested before actual data collection was conducted. Based on the pretest, the booklets and instructions were refined. Examples of the instructions are contained in Appendix B of Volume 2.

4.2.4 Procedure

Data collection sessions were scheduled. A data collection session administrator directed the sessions. He read instructions to the subject matter experts, ensured that the subjects did not exchange information, and handled questions about the instructions. The session administrator did not answer impromptu questions about technical details of the task statements because inconsistencies could have been introduced. Experts were asked to note any questions or assumptions about the tasks in their response booklets or during the exit interview. By design, the session administrator had no experience with psychological scaling techniques. A psychological scaling expert was available during data collection sessions as an observer. He did not participate in the data collection session.

The data collection session was divided into four periods, one for each of the four parts of the response booklet described in Section 4.2.3. Experts were asked to make paired comparison judgments before direct estimates. Paired comparisons are relative judgments, i.e., one task more likely than another, and no probability estimates are assigned to the tasks. Direct estimates involve the assignment of probabilities to tasks. Thus, paired comparisons were less likely to influence direct estimates because comparisons among all possible pairs of tasks are less easily remembered than directly estimated values.

Responses from completed booklets were transferred to coding sheets for entry into a computer. Then, data analyses as described in the following section were performed. In order to adequately address the technical issue of calibrating the paired comparison scale values into probability estimates, this study used several sources of HEP estimates for anchor tasks and both two and four anchor tasks. Thus, analyses could determine the effects of using different sources and different numbers of anchors.

4.3 Study Results

The results of this study provide a positive evaluation of both techniques used to derive HEP estimates from the judgments of experts. Although all aspects of the program issues could not be completely resolved on the basis of this single test, it has provided considerable information relevant to these issues and substantial support for use of the techniques. The test has also resulted in the development of a detailed description of the process by which these techniques are used (Appendix A), a process that has been tested and has been shown to be workable.

The study provided results from the process of actually using the techniques and also an evaluation of those results. Both types of results are important to addressing the issues. The following discussion describes these results as they relate to program and technical issues. This discussion is intended to summarize results and does not include statistical details and data. These details and data are found in Appendix B of Volume 2.

The first results of interest are, of course, the HEP estimates themselves. Figure 2 shows these estimates for the Level 1 tasks in a format that allows easy comparison of the estimates from different sources. These estimates are also given numerically in Appendix C. For this set of tasks, there were three estimates for each task: the direct estimate, and paired comparison estimates using two and four direct estimates as anchor tasks. As can be seen, there was good agreement between the direct estimates and the paired comparison estimates with four anchors. (The results with two anchors are similar.) In fact, only one task varied by as much as an order of magnitude. Most of the other tasks varied by only a difference of 1 in the first significant digit (e.g., from .03 to .04).

Seven sets of HEP estimates for Level 2 and 3 tasks were obtained in this study: one from direct estimates and six from paired comparisons. The six paired comparison estimates came from using each of three sources of anchor task HEP estimates (direct estimates, Handbook estimates, and simulator estimates) with both two and four anchor tasks. In addition, HEP estimates for these tasks were available from the Handbook and, for four tasks, from simulator studies (Beare et al., 1984). Figures 3 through 9 show several plots that allow comparison of direct estimates, paired comparison estimates and Handbook estimates. (Again, results using two anchors were similar so are not shown.)

Again, the agreement among estimates from different sources was reasonably good, although not as good as for the Level 1 tasks. There were a few specific tasks on which the differences were somewhat more substantial. Also, paired comparison estimates with simulator anchors tended to cluster together and not to discriminate among tasks to the extent other estimates did, apparently because of the limited range of the HEP estimates for simulator anchor tasks (Figures 6 and 8). These data can also be found in numeric form in Appendix C.

In addition to making judgments to estimate HEPs, the experts were asked to estimate upper and lower uncertainty bounds that represent the range over which the HEP might be expected to vary as PSFs affecting errors vary from very adverse to very good. These uncertainty bound estimates are shown in Figures 10 and 11 for Level 1 and Level 2 and 3 tasks, respectively. (Figure 10 shows descriptive information only, since there are no system level data available from either the Handbook or the simulator.) Figure 11 also shows uncertainty bounds from the Handbook. Again there is reasonable agreement with the Handbook

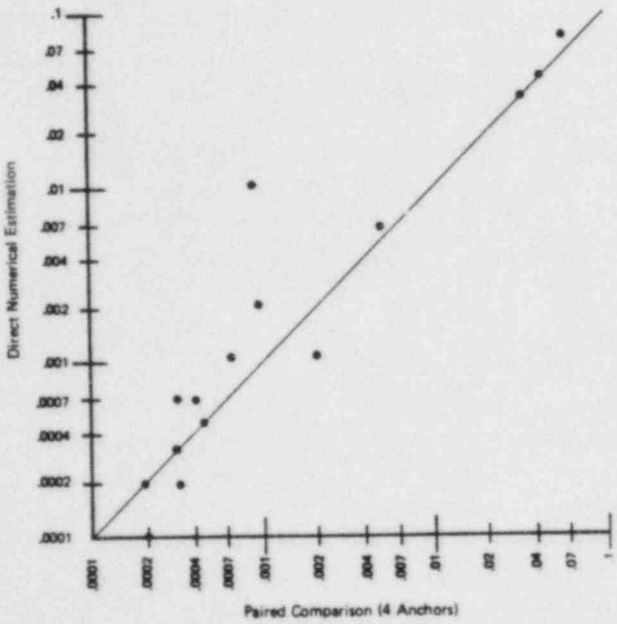

Figure 2 Level 1 direct numerical estimates and paired comparison estimates with four anchors.

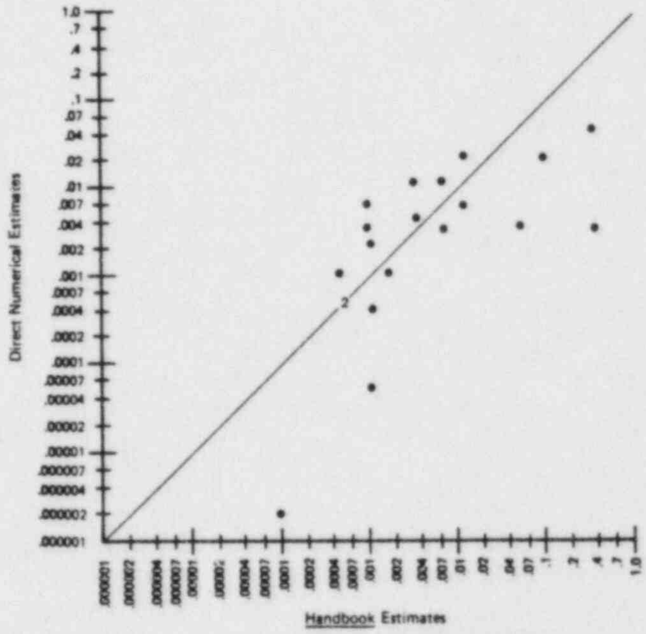

Figure 3 Level 2/3 direct numerical estimates and Handbook estimates.

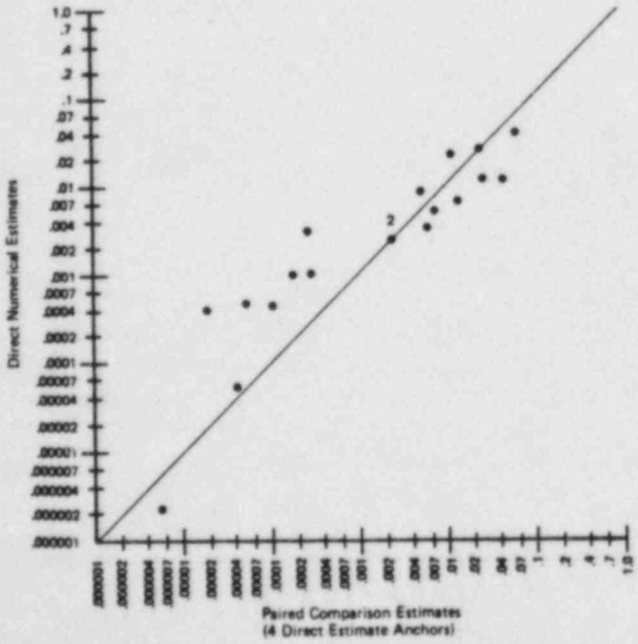

Figure 4 Level 2/3 direct numerical estimates and paired comparison estimates with four direct estimate anchors.

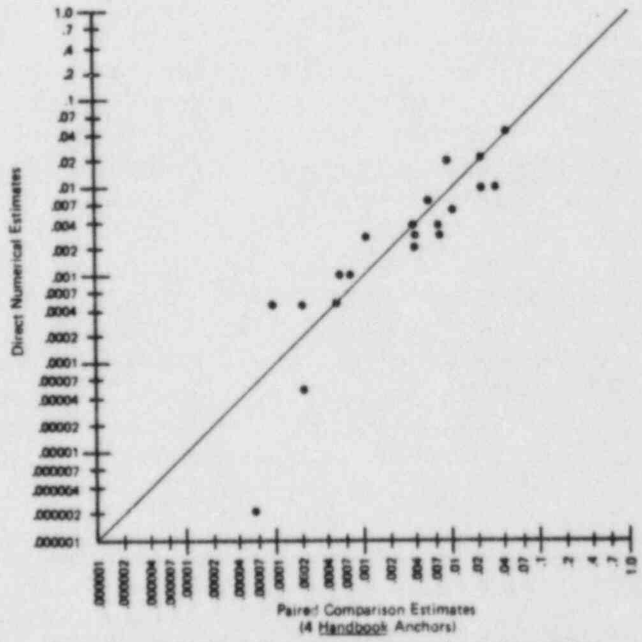

Figure 5 Level 2/3 direct numerical estimates and paired comparison estimates with four Handbook anchors.

2 = two points coincide

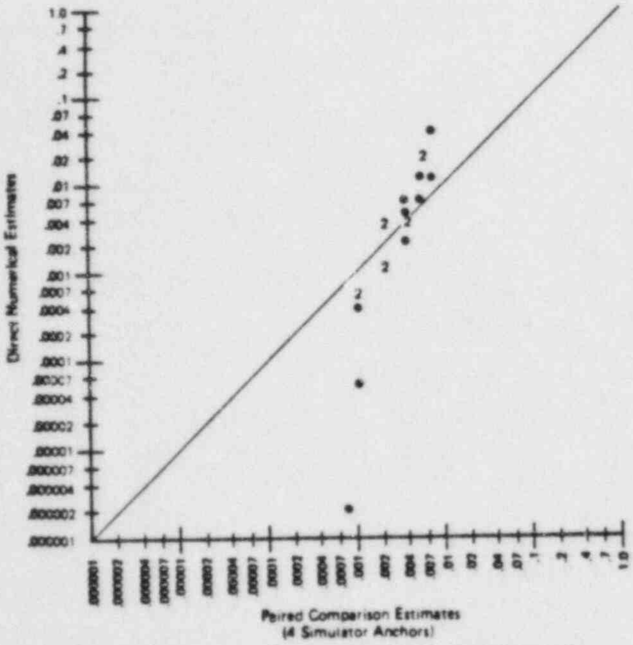

Figure 6 Level 2/3 direct numerical estimates and paired comparison estimates with four simulator anchors.

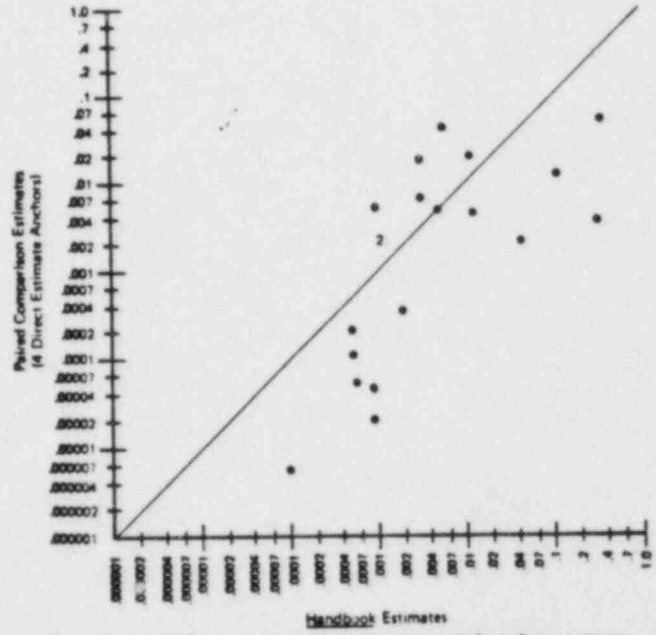

Figure 7 Level 2/3 paired comparison estimates with four direct estimate anchors and Handbook estimates.

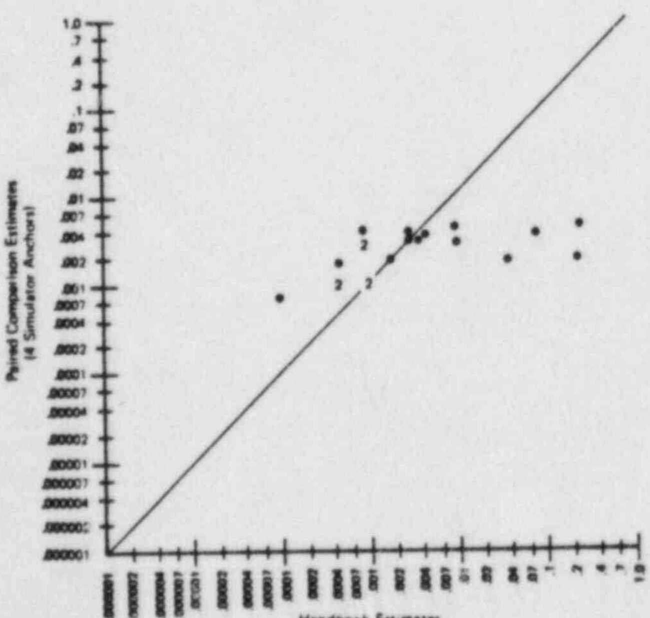

Figure 8 Level 2/3 paired comparison estimates with four simulator anchors and Handbook estimates.

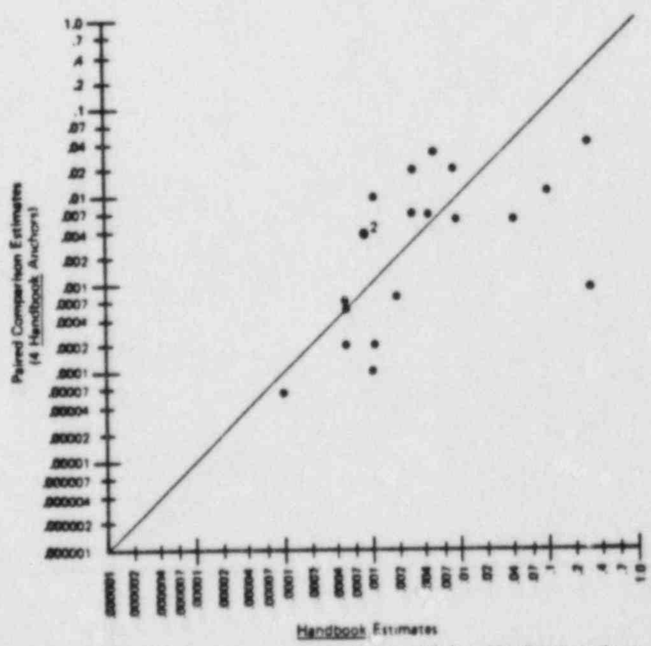

Figure 9 Level 2/3 paired comparison estimates with four Handbook anchors and Handbook estimates.

2 = two points coincide

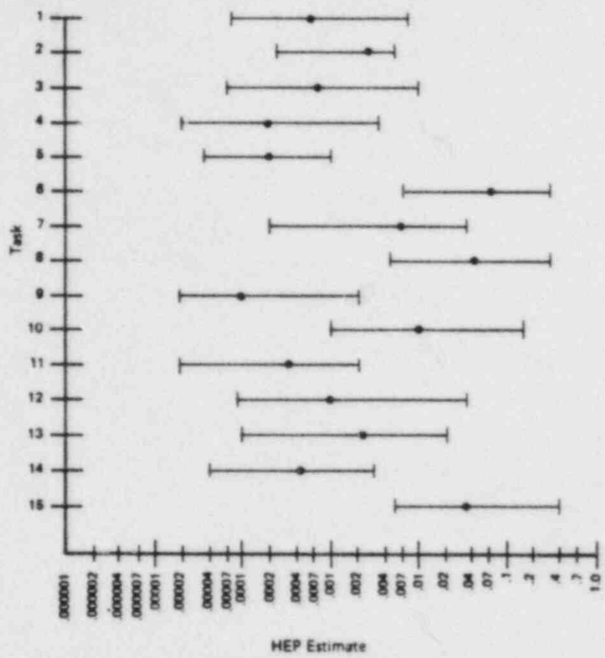

Figure 10 Direct numerical estimates and uncertainty bounds for Level 1 tasks.

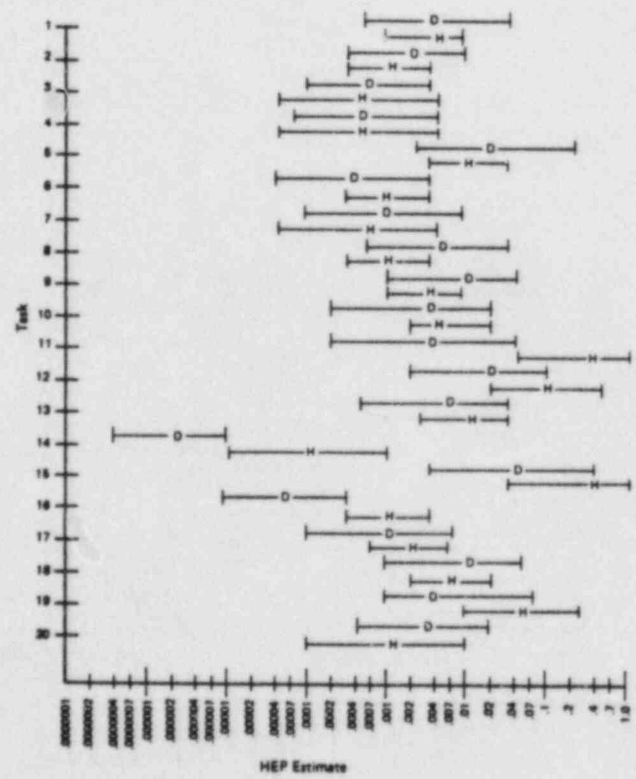

Figure 11 HEP estimates and uncertainty bounds for direct numerical estimation (D) and the <u>Handbook</u> (H) for Level 2/3 tasks.

except for a few specific tasks. The estimates of uncertainty bounds are also shown numerically in Appendix C. In most instances, the estimate from one source fell within the bounds estimated by the other source.

With these basic results in mind, the remainder of this section discusses additional analysis and study results that bear directly on the program and technical issues discussed previously.

4.3.1 Do Psychological Scaling Techniques Produce Consistent Judgments From Which to Estimate HEPs?

A comparison of expert judgments indicated that consisteny was very high and is more than adequate to use these techniques and the resulting estimates to support potential applications discussed in Section 1. Two types of consistency were measured: (1) the internal consistency of individual expert's paired comparison judgments, and (2) the across-expert consistency or agreement of judgments for both direct estimates and paired comparison judgments for the group.

The internal consistency of an expert's paired comparison judgments is a measure of the number of intransitive triads in paired comparisons. An intransitive triad is one in which task a is judged more likely than b, b more likely than c, and c more likely than a. Comparing the actual number of intransitive triads with the possible number of intransitive triads provides a measure of consistency for which 0 indicates the maximum inconsistency and 1 represents complete consistency (no intransitive triads).

The within-expert consistency was extremely high for both Level 1 and Level 2 and 3 tasks. Almost all measures were between .8 and 1. Actual measures are shown in Appendix B of Volume 2.

Across-expert consistency for both techniques was more moderate than within-expert consistency for both Level 1 and Level 2 and 3 tasks, as was to be expected with the relatively large number (19) of experts used. The coefficient of concordance used as the measure of consistency also had a range of 0 (no agreement) to 1 (complete agreement). The coefficients for paired comparisons were between .50 and .59, while those for direct estimates were near .4. Similar across-expert consistency measures obtained for the uncertainty bound estimates ranged from .34 to .40, somewhat lower than for the HEP estimates themselves, as might be expected because the bounds were more open to individual interpretation. All of these across-expert consistency measures were also highly statistically significant.

4.3.2 Do Psychological Scaling Techniques Produce Valid HEP Estimates?

The focus of analyses related to validity was on convergent validity, i.e., the degree to which different approaches to estimating HEPs produce

Table 6 Correlation coefficients for HEP estimates from various sources

Source	Correlation[1]
Level 1 Tasks	
Direct Estimates - Paired Comparisons	0.94**
Level 2 and 3 Tasks	
Direct Estimates - Paired Comparisons	0.89**
Direct Estimates - Handbook	0.68**
Paired Comparisons - Handbook	0.57*

* $p \leq .01$
** $p \leq .001$

[1] Correlations were computed using logarithms of HEP estimates.

the same estimates. Predictive validity, the extent to which the HEPs which are estimated using expert judgment are able to predict actual error rates, could not be determined because actual error rates are not known. The correlations among the different estimates were quite high as shown in Table 6.

This convergent validity was also supported by several additional analyses that are described in Section 4.1 of Appendix B. Taken together, these analyses suggest that much of the difference that did exist in estimates arises from the use of different anchors in the paired comparison estimates. In particular, using just two anchor tasks (Tasks 14 and 15) produced relatively low HEP estimates with direct estimate anchors because one task (Task 14) had a very small direct HEP estimate, and relatively high estimates with Handbook anchors because one task (Task 15) had a relatively high Handbook estimate. Using four anchors greatly reduced the effects of these extreme anchor task estimates, and therefore produced more convergence in estimates.

4.3.3 Can the Data Collected Using Psychological Scaling Techniques Be Generalized?

Generalizability was addressed primarily in the process of designing this study. Tasks, particularly Level 1 tasks, were selected and defined to be generic to all BWR plants, and therefore HEP estimates for these tasks should be appropriate with adjustments for plant-specific factors. This provides a degree of generalizability across plants, although not necessarily to pressurized water reactor (PWR) plants. Level 2 and 3 tasks, as defined, should be relatively more generalizable because they are not system-specific and they refer to operation of components, instruments, etc., that are found in most all plants.

Some data analyses also address the generalizability issue. The across-expert consistency measures and the statistical confidence limits described in Section 4.2 of Appendix B indicate the degree to which the HEP estimates obtained in this study generalize to estimates that might be obtained from other, similar experts. The moderate, though significant, results suggest that reasonably similar estimates would be obtained from other experts.

4.3.4 Are the HEP Estimates That Are Generated From Psychological Scaling Techniques Suitable for Use in PRAs and the Human Realiability Data Bank?

This issue was addressed primarily by the process by which task definitions were developed and also by the fact that consistent judgments could be obtained for these tasks.

Use of these estimates in the data bank, NUREG/CR-2744, Volume 2 (Comer et al., 1983), and in PRA is closely related because the data bank was designed to be consistent with PRA needs. This issue was addressed in part by the use of different task sets corresponding to different levels in the data bank. Tasks from all three levels were produced with Levels 2 and 3 being combined for the purpose of this study. Judgments were successfully collected and HEP estimates were derived for all tasks, indicating that these techniques can be considered for use in the data bank and therefore for PRAs. In addition, the Level 1 tasks were extensively reviewed by PRA practitioners to ensure that they were representative of tasks for which HEP estimates are needed in PRAs.

An additional important consideration was the capability to produce uncertainty bounds. Such bounds were estimated using direct numerical estimation. Both their across-expert consistency and their convergence with <u>Handbook</u> uncertainty bound estimates were reasonably good. Generally, the bounds estimated in this study were somewhat wider than the <u>Handbook</u> estimates. Since there were no actuarial data in a form suitable for comparison with the estimates from psychological scaling, estimates from the <u>Handbook</u> were used. These comparisons were made to determine convergent validity.

The HEP estimates from paired comparisons, the <u>Handbook</u>, and simulator studies were also compared with the estimated uncertainty bounds. These HEP estimates should fall between the estimates of uncertainty bounds. As shown in Appendix B, this was generally true with the exception of paired comparison estimates with two direct estimate anchors, which tended to fall below the bounds, and paired comparison estimates with two

Handbook anchors which tended to fall above the bounds. These latter results can again be attributed to the extremeness of these anchors, which was discussed above.

4.3.5 Can Psychological Scaling Procedures Be Used by Persons Who Are Not Expert in Psychological Scaling to Generate HEP Estimates?

Again, this issue was addressed in the design of the study and the process by which it was implemented rather than by data analyses. The data collection procedures, including instructions, were designed to be conducted by someone with no background in psychological scaling. These procedures, described in Appendix A, were pretested, revised, and then used in actual data collection. This data collection was conducted smoothly, with no obvious difficulties, by a nonexpert. The detailed description of procedures in Appendix A should make data collection and analysis possible without the assistance of a psychological scaling expert. In addition, the results of other analyses indicate that the experts (instructors) were able to make the necessary judgments.

The study also showed that a human reliability analyst is probably necessary to define the tasks to be judged to meet the needs of the PRA practitioner. Representative subject matter experts are also needed to evaluate whether the task statements will be understandable and meaningful to other, similar experts.

4.3.6 Do the Experts Used in the Psychological Scaling Process Have Confidence in Their Ability To Make the Judgments?

The confidence of the experts in their judgments is one indication of the reasonableness of the HEP estimates, although experience in other contexts suggest that often experts can make good probability estimates even when they are doubtful of their ability to do so. The experts in this study were systematically questioned regarding their perception of the accuracy and difficulty of the required judgments. In general, the experts were neutral about their judgments, although paired comparison judgments were considered to be somewhat more accurate than direct estimates. On a six-point scale along which 1 indicated accurate and 6 indicated inaccurate, average judgments were 2.1 for paired comparisons and 3.1 for direct estimates. The experts were also neutral regarding the difficulty of the judgments with means of 3.2 and 3.3 for paired comparisons and direct estimates, respectively. They also considered the uncertainty bound estimates to be somewhat more difficult (mean 3.9) and less accurate (mean 3.5) than direct HEP estimates or paired comparisons.

4.3.7 Is There Any Difference in the Quality of Estimates Obtained From the Two Scaling Techniques?

A primary consideration involves which of the two techniques to use. Although, as discussed in Sections 2 and 3, there are several practical considerations in answering this question, the results of this study indicate that with respect to the HEP estimates obtained, there is little difference between the two techniques. Paired comparisons have somewhat higher across-expert consistency (coefficients of .54 for Level 1 tasks, .57 for Level 2 and 3 tasks versus .39 and .42, respectively, for direct estimates), while direct estimates correlate higher with Handbook estimates (coefficient of .68 versus .40 for paired comparisons). Experts do perceive their paired comparison judgments to be somewhat more accurate. None of these differences are large nor do they appear to provide a strong basis for selecting one technique. Therefore, as is discussed more fully in the following section, selection of a technique can be based on practical considerations such as number of experts or time available.

4.3.8 Is There Any Difference in the Results Based on the Type of Task That Is Being Judged?

Because of the potential use of these techniques in PRA, the second technical issue relates to differences in results between the two task sets, the relatively complex tasks in Level 1 and the more simply defined tasks in Levels 2 and 3. Study results were not generally different for the two types of tasks. Within-expert consistency was somewhat higher for Level 1 tasks (coefficient of .89 versus .86). Also described in Section 4.3.7, across-expert consistency was slightly higher for Level 2 and 3 tasks, and the convergence among different estimates was somewhat better for Level 1 tasks. None of these differences, however, was large. In addition, and perhaps more importantly, the experts judged the tasks in both sets to be relatively easy to understand (mean of 1.7 on a six-point scale with 1 indicating easy to understand).

4.3.9 Do Education and Experience Have Any Effect on the Experts' Judgments?

An additional technical issue regarding the effects of the experts' background was also addressed. There was little variation among the experts in terms of level of education, amount of experience, or type of license or certification. Analyses indicated that these background variables were in no way related to judgments from this homogenous set of experts.

4.3.10 How Should the Paired Comparison Scale Be Calibrated Into a Probability Scale?

One of the most important technical issues, if paired comparison judgments are to be used to estimate HEPs, involves the transformation of scale values obtained from the paired comparison judgments into HEP estimates. This transformation involves an assumption with regard to how scale values are related to probabilities, and the estimation of two constant values to be used in the transformation from anchor tasks. Neither of these parts of the transformation had substantial empirical support prior to this study.

Our examination of the relationship between scale values and probabilities indicated that a logarithmic relationship was more appropriate than a linear one. The examination of the effect of anchor tasks consisted of examining the number of tasks used (two or four) and the source of the anchor task HEP estimates (direct estimates, Handbook, or simulator). Results indicate that the source of the anchor had relatively little effect if four anchor tasks were used, but had much more effect if only two anchor tasks were used. Basically, these results indicate that more than two anchor tasks should be used, if possible, to reduce the influence of any single HEP estimate used as an anchor for a set of paired comparison HEP estimates.

4.3.11 Can Reasonable Uncertainty Bounds Be Estimated Judgmentally?

The experts were able to estimate uncertainty bounds using direct numerical estimation, but these estimates were subjected to only the following limited analysis. The across-expert consistency of the uncertainty bound estimates was moderate and only slightly lower than for HEP estimates (coefficients of about .34 for lower bounds and .40 for upper bounds). Also, HEP estimates from the Handbook and simulator studies were generally between the uncertainty bound estimates.

5. CONCLUSIONS AND RECOMMENDATIONS

Based on the results of this project, several conclusions can be reached regarding the use of the techniques described in this report and tested in this study.

The conclusions are:

- Both direct numerical estimates and paired comparison judgments more than met statistical requirements for consistency.

- Convergent validity of the HEP estimates was good, particularly if the effects of using only two anchor tasks for paired comparison estimates are disregarded. It should be noted though, that predictive validity with respect to HEP estimates based on the actual relative frequency of errors could not be established because of the lack of such estimates. (This will be a difficulty in validating any procedure used to estimate HEPs.)

- The tasks and their HEP estimates should be generalizable to all BWRs. Results should also be somewhat generalizable to other, similar groups of experts. The actual extent of this latter generalizability has not been fully tested.

- Tasks can be appropriately defined and HEP estimates for them can be obtained so that the estimates can be used in PRAs and in the Human Reliability Data Bank.

- The judgments required can be obtained from experts without the use of an expert in psychological scaling. However, epxertise in human reliability, statistics, and task subject matter is needed for task selection, analysis, and judgment.

- Experts making the judgments have only a moderate degree of confidence in their judgments. (Often experts without experience in making these types of judgments will lack confidence in the judgments. Confidence will increase with experience. Lack of confidence does not imply that the judgments are not sound.)

- Only minor differences occur in the evaluations of direct numerical estimates and paired comparison estimates. One technique cannot be selected over the other on the basis of these analyses alone. In some situations, use of direct numerical estimation may be preferred to paired comparison scaling because of practical considerations such as requiring fewer experts (as few as six for direct estimation versus 10 to 12 for paired comparison) and less of the experts' time. For example, if paired comparison scaling is used to obtain uncertainty bound estimates, it will increase the amount of time required to make judgments by three (once for the HEP estimate, once for the lower bound, and once for the upper bound).

- Only minor differences in consistency and convergent validity occur in the results for the two types of tasks (Level 1 and Levels 2 and 3). Expert judgment can be used to estimate HEPs for either type of task.

- Background variables such as education, experience, and type of license/certification did not affect judgments. The extent to which this conclusion is true, beyond the specific group of instructors used as experts in this study, is not known because the group used was very homogeneous.

- For paired comparison estimates, scale values should be transformed into HEP estimates using a logarithmic relationship. Human error probability estimates for more than two tasks (e.g., four) should be used to estimate the parameters in the transformation.

- Uncertainty bounds can be estimated using direct estimates, although this study was not designed to thoroughly test the resulting estimates.

As a practical matter, this study demonstrated that either technique can be used to estimate HEPs in a timely manner. Expert judgment data can be obtained and used in a relatively cost-effective manner with tasks that are carefully defined to meet PRA needs. Psychological scaling techniques can thus be used to generate estimates without some of the difficulties of task definition or inadequate data that may affect simulator studies or field reporting. The main drawback presently in the use of expert judgment or any other procedure to estimate HEPs is the inability to establish predictive validity.

Taken together, the conclusions indicate that these techniques using expert judgment should be given strong consideration for use in developing estimates for the Human Reliability Data Bank. In addition, they can be implemented, as needed, to provide HEP estimates for PRAs.

Additional research on the use of expert judgment might be especially valuable in several areas: time-response functions, estimation and assessment of uncertainty bounds, assessment of predictive validity, and development of anchor task estimates. Time-response functions show the probability that an operator will successfully perform a task within a certain time frame, with the probability varying as the amount of time varies. The HEP estimates obtained in this study were essentially estimates for a single point in time. Time response functions provide the estimates needed for a wider range of contexts. If expert judgment can be used to obtain time-response functions, the number of overall judgments required could be reduced.

In this project, uncertainty bound estimates were obtained using expert judgment, although this study was not designed to thoroughly test the resulting estimates of bounds. Additional research could be undertaken to explore whether there are systematic biases in these estimates and to further investigate other judgmental methods for obtaining estimates of uncertainty bounds. Finally, simulator studies could provide an excellent source of anchor task HEP estimates needed for paired comparison estimates.

REFERENCES

Beare, A. N., Dorris, R. E., Bovell, C. R., Crowe, D. S., and Kozinsky, E. J., A simulator-based study of human errors in nuclear power plant control room tasks (NUREG/CR-3309, SAND 83-7095). Albuquerque, NM: Sandia National Laboratories, January 1984.

Comer, M. K., Kozinsky, E. J., Eckel, J. S., Miller D. P. Human Reliability Data Bank for nuclear power plant operation, Volume 2: A data bank concept and system description (NUREG/CR-2744, Vol. 2; SAND82-7057, Vol. 2). Albuquerque, NM: Sandia National Laboratories, February 1983.

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

Efron, B. The jackknife bootstrap and other resampling plans (CBMS38). National Science Foundation, SIAM, 1982.

Embrey, D.E. The use of performance shaping factors and quantified expert judgment in the evaluation of human reliability: An initial appraisal (NUREG/CR-2986, BNL-NUREG-51591). Upton, NY:, Brookhaven National Laboratory, May 1983.

Embrey, D.E., Humphreys, P., Rosa, E.A., Kirwan, B., & Rea, K. SLIM-MAUD: An approach to assessing human error probabilities using structured expert judgment (NUREG/CR-3518, BNL-NUREG-51716). Upton, NY:, Brookhaven National Laboratory, March 1984.

Seaver, D. A., and Stillwell, W. G. Procedures for using expert judgment to estimate human error probabilities in nuclear power plant operations (NUREG/CR-2743, SAND82-7054). Albuquerque, NM: Sandia National Laboratories, March 1983.

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

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 (NUREG/CR-1278). Washington, D.C.: U.S. Nuclear Regulatory Commission, August 1983.

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.

Distribution

U.S. NRC Distribution Contractor (CDSI) (400)
7300 Pearl Street
Bethesda, MD 20014
400 copies for RX
232 copies for Author-Selected Distribution

Dr. Lee Abramson
Division of Risk Analysis and Operations
Reactor Risk Branch
Mail Stop 5650 Nicholson Lane
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

Mr. John Beakes
General Physics Corporation
10650 Hickory Ridge Road
Columbia, MD 21044

Mr. Arthur Beare
General Physics Corporation
1770 The Exchange
Atlanta, GA 30339

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

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

Dr. Robert Blanchard
Naval Personnel R&D Center
San Diego, CA 92152

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

Mr. Lewie Booth
LATA
2834 Sunnygled Road
Torrance, CA 90505

Dr. Katrin 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

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

Mr. Donald Burgy
General Physics Corporation
10650 Hickory Ridge Road
Columbia, MD 21044

Dr. James Chinnis
President
Decision Science Consortium, Inc.
Suite 421
7700 Leesburg Pike
Falls Church, VA 22043

Dr. Julien M. Christensen
Universal Energy Systems
4401 Dayton-Xenia Road
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 (50)
Manager of Engineering Analysis
General Physics Corporation
10650 Hickory Ridge Road
Columbia, MD 21044

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

Mr. Robert Danna
General Physics Corporation
10650 Hickory Ridge Road
Columbia, MD 21044

Mr. Tom Davis
General Physics Corporation
10650 Hickory Ridge Road
Columbia, MD 21044

Mr. Mike Donovan
General Physics Corporation
1770 The Exchange, Suite 150
Atlanta, GA 30339

Prof. Yves Dutuit
LARSACT
Inst. Univ. de Technologie A
33405 Talence Cedex
FRANCE

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

Mr. Gary Engmann
Black & Veatch Consulting Engineers
P. O. Box 8405
Kansas City, MO, 64114

Dr. Baruch Fischhoff
Decision Research
1201 Oak
Eugene, OR 97401

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

Dr. Catherine Gaddy
General Physics Corporation
10650 Hickory Ridge Road
Columbia, MD 21044

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

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

Dr. Frank Gomer
General Physics Corporation
1010 Woodman Drive, Suite 240
Dayton, OH 45432

Mr. Hank Guttmann
418 Oak, N. E.
Albuquerque, NM 87106

Dr. G. W. Hannaman
NUS Corporation, Suite 250
16885 W. Bernardo Drive
San Diego, CA 92127

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

Prof. Yoshio Hayashi
Department of Adm. Engineering
Keio University
3-14-1 Hiyoshi, Kohoku
Yokohama 223, JAPAN

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

Dr. J. Roger Humphries
Manager, Licensing Branch
Atomic Energy of Canada Limited
Sheridan Park Research Community
Mississauga, Ontario L5K 1B2 CANADA

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

Dr. Helmut Jungermann
Institut fur Psycnologie
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. Richard Kelly
Navy Personnel Research and Development Center
Code 17
San Diego, CA 92152

Mr. Ernest Koehler
Navy Personnel Research and Development Center
Code 17
San Diego, CA 92152

Ms. Nancy Knight
General Physics Corporation
10650 Hickory Ridge Road
Columbia, MD 21044

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

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. Bruce Logan
Duke Power Company
P. O. Box 33189
Charlotte, NC 28242

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

Dr. James A. Mahaffey
Georgia Institute of Technology
Engineering Experiment Station
Atlanta, GA 30332

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

Dr. Harry Martz
Group S-1 MS F600
Los Alamos Laboratory
Los Alamos, NM 87545

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. Doug Morris
Black & Veatch Consulting Engineers
P. O. Box 8405
Kansas City, MO 64114

Mr. Reidar J. Mykletun
Rogaland Research Institute
P. O. Box 2503, Ullandhaug
N-4001 Stavanger, Norway

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

Ms. Karen Ness
Army Research Institute
ARI Field Unit
Fort Leavenworth, KS 66027

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

Mr. John O'Neill
Florida Power & Light Company
P.O. Box 14000
Juno Beach, FL 33408

Mr. Reider Østvik
SINTEF
N7034 Trondheim
NTH
Norway

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

Mr. Alan Passwater
Superintendent, Licensing
Union Electric Company
1901 Gratiot St.
P. O. Box 149, Code 470
St. Louis, MO 63166

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. Ken Rebeck
General Physics Corporation
10650 Hickory Ridge Road
Columbia, MD 21044

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

Mr. Gene Rosa
Bldg. 130 - Brookhaven National Lab
Upton, NY 11973

Mr. Larry Rose
Detroit Edison
2000 Second Ave., Room 516SB
Detroit, MI 48226

Dr. Marvin Rousch
Dept. of Chemical & Nuclear Engineering
University of Maryland
College Park, MD 20742

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
Brahegatan 5
11437 Stockholm
Sweden

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

Dr. Lothar Schroeder
General Physics Corporation
10650 Hickory Ridge Road
Columbia, MD 21044

Dr. David A. Seaver
General Physics Corporation
10650 Hickory Ridge Road
Columbia, MD 21044

Mr. Lee Sippel
Kansas City Power & Light
1330 Baltimore Avenue
Kansas City, MO 64105

Dr. Kurt J. Snapper
12604 Magna Carta Road
Herndon, VA 22070

Dr. Michael E. Stephens (10)
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. 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. V. R. R. Uppuluri
Mathematics & Statistics Research Dept.
Building 9704-1
Oak Ridge National Laboratory
P. O. Box 4
Oak Ridge, TN 37830

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

Mr. Charles Willard
Willard Associates, Inc.
3412 Contry Hill Drive
Fairfax, VA 22030

Mr. Jeremy C. Williams
National Center of Systems Reliability
UKAEA
Safety & Reliability Directorate
Wigshaw Lane
Culcheth
Warrington WA 3 4NE England

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
Batelle Columbus Laboratories
505 King Avenue
Columbus, OH 43201

Mr. Jan Wright
Bronnoyvn 20
1315 Nesoya
NORWAY

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

Internal Sandia Distribution

3141 L. J. Erickson (5)
3151 W. L. Garner (3)
6412 G. J. Kolb
6412 R. L. Iman (2)
7223 R. G. Easterling
7223 B. P. Chao
7223 K. V. Diegert
7223 D. P. Miller
7223 F. W. Spencer
7223 L. M. Weston (32)
7223 H. O. Whitehurst
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-3688, SAND84-7115

4. TITLE AND SUBTITLE (Add Volume No., if appropriate)
Generating Human Reliability Estimates Using Expert Judgment, Volume 1: Main Report; Volume 2: Appendices

2. (Leave blank)

3. RECIPIENT'S ACCESSION NO.

7. AUTHOR(S)
M. K. Comer, D. A. Seaver, W. G. Stillwell, and C. D. Gaddy

5. DATE REPORT COMPLETED

MONTH	YEAR
October	1984

DATE REPORT ISSUED

MONTH	YEAR

9. PERFORMING ORGANIZATION NAME AND MAILING ADDRESS (Include Zip Code)
General Physics Corporation
10650 Hickory Ridge Road
Columbia, Maryland 21044

6. (Leave blank)

8. (Leave blank)

12. SPONSORING ORGANIZATION NAME AND MAILING ADDRESS (Include Zip Code)
Division of Risk Assessment and 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 1983 - October 1984

15. SUPPLEMENTARY NOTES

14. (Leave blank)

16. ABSTRACT (200 words or less)

The U.S. Nuclear Regulatory Commission is conducting a research program to determine the practicality, acceptability, and usefulness of several different methods for obtaining human reliability data and estimates that can be used in nuclear power plant probabilistic risk assessments (PRA). One method, investigated as part of this overall research program, uses expert judgment to generate human error probability (HEP) estimates and associated uncertainty bounds. The project described in this document evaluated two techniques for using expert judgment: paired comparisons and direct numerical estimation. Volume 1 of this report provides a brief overview of the background of the project, the procedure for using psychological scaling techniques to generate HEP estimates and conclusions from evaluation of the techniques. Volume 2 provides detailed procedures for using the techniques, detailed descriptions of the analyses performed to evaluate the techniques, and HEP estimates generated as part of this project.

The results of the evaluation indicate that techniques using expert judgment should be given strong consideration for use in developing HEP estimates. In addition, HEP estimates for 35 tasks related to boiling water reactors (BWRs) were obtained as part of the evaluation. These HEP estimates are also included in the report.

17. KEY WORDS AND DOCUMENT ANALYSIS

17a. DESCRIPTORS

psychological scaling
probability assessment
expert opinion
human error probability
uncertainty bounds
paired comparisons
direct numerical estimation

17b. IDENTIFIERS/OPEN-ENDED TERMS

18. AVAILABILITY STATEMENT
Unlimited

19. SECURITY CLASS (This report)
Unclassified

20. SECURITY CLASS (This page)
Unclassified

21. NO. OF PAGES
233

22. PRICE
$

*U.S. GOVERNMENT PRINTING OFFICE: 1984-0-576-035/4398

120555078877 1 1AN1RX
US NRC
ADM-DIV OF TIDC
POLICY & PUB MGT BR-PDR NUREG
W-501
WASHINGTON DC 20555