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

Generating Human Reliability Estimates Using Expert Judgment

Volume 2. Appendices

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)

LIST OF TABLES

LIST OF FIGURES

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

INTRODUCTION TO VOLUME 2

Volume 1, the main report, contains an overview of psychological scaling techniques and a summary description of how they were evaluated. Volume 2 contains the three appendices to the main report. They are: Appendix A - Instructions for the Use of Psychological Scaling Procedures, Appendix B - Evaluation Results, and Appendix C - Human Error Probability Estimates.

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 statistically evaluating the data and the techniques are relatively advanced, Appendix B is more complex than the other sections of the report. It 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 collected as part of this project. The appendix is intended to be used by those who have an interest in or need for estimates of human error probabilities (HEPs) and the associated uncertainty bounds.

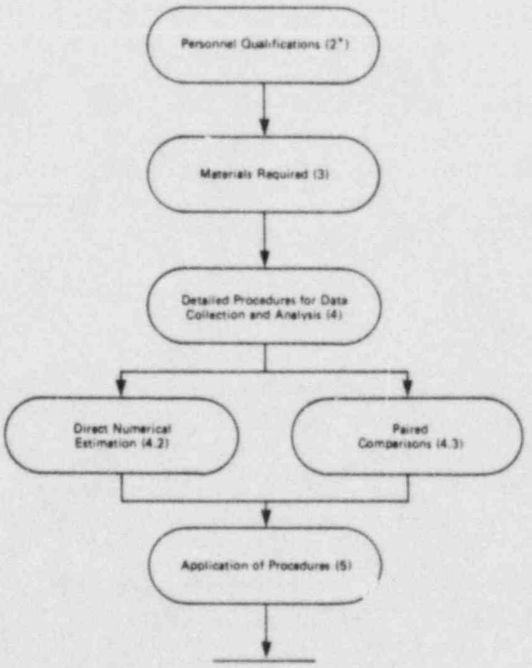
APPENDIX A

INSTRUCTIONS FOR THE USE OF PSYCHOLOGICAL SCALING TECHNIQUES

1. INTRODUCTION

Instructions for using psychological scaling techniques are presented in this appendix. Anyone wishing to use psychological scaling to generate human reliability estimates can use these instructions by following the instructions outlined in this appendix, and shown in Figure A.1, without reference to the main report presented in Volume 1 or any of the other appendices in this volume.

Personnel qualifications, materials, and data collection and analysis are described in the following sections. Specifically, qualifications are provided for the four types of personnel needed to implement the techniques. Descriptions and samples of the materials needed are given. Detailed, step-by-step procedures for collecting and analyzing direct numerical estimation or paired comparison data are given. A final section specifies cautions to be considered when analyzing the data and factors to consider in selecting a technique (i.e., direct numerical estimation or paired comparison).



*Numbers refer to section in Appendix A.

Figure A.1 Overview of psychological scaling.

2. PERSONNEL QUALIFICATIONS

Four types of personnel are needed to implement psychological scaling procedures as shown in Figure A.2: human reliability analysts, subject matter experts, a data collection session administrator, and data analysts. The qualifications of each are discussed in this section. The number of each of these types of personnel needed to implement direct numerical estimation or paired comparisons is shown in Table A.1.

2.1 Human Reliability Analyst

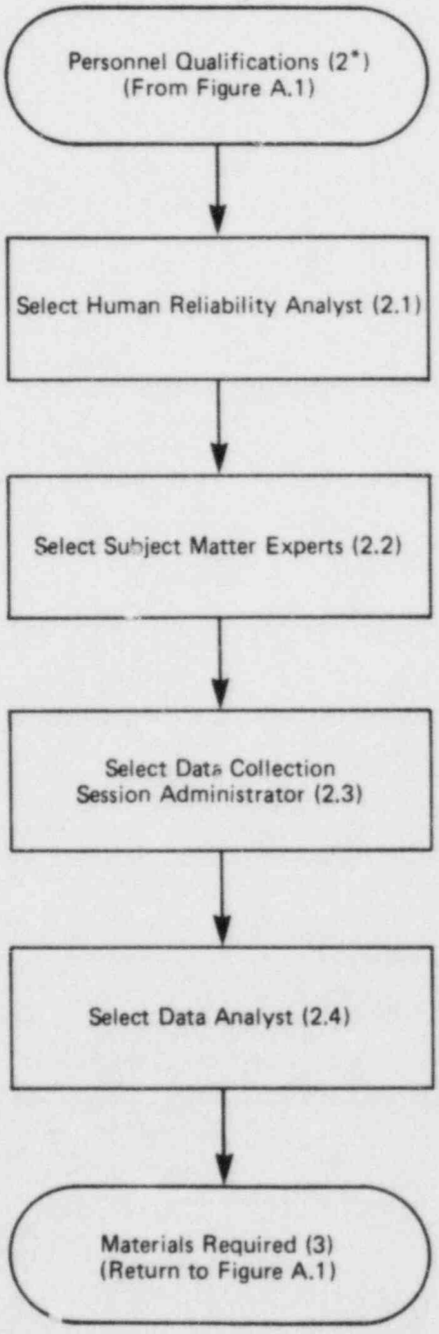
A human reliability analyst, or someone who is very familiar with the tasks to be judged and their application in a probabilistic risk assessment (PRA), is needed to assist in the task definition development. As stated in Volume 1, it is essential that the tasks be thoroughly defined and presented to the experts in their own language. Therefore, it is essential that someone be able to translate task definitions from PRA language to nuclear power plant operations language.

2.2 Subject Matter Experts

The experts who make the judgments must be familiar with the tasks to be judged. If the tasks involve nuclear power plant control room operations, the experts must have an in-depth knowledge of plant systems, operations, and control room procedures. Individuals who are currently, or were formerly, licensed as control room operators by the U.S. Nuclear Regulatory Commission, or certified as operations instructors, meet these requirements.

If the tasks to be judged include accidents or other infrequently occurring events, certified instructors are better qualified than individuals who are licensed operators to judge the likelihood of various operator actions under these conditions. Instructors have had the opportunity to witness many different operators and their reactions to simulated accident scenarios. Therefore, instructors fulfill the criterion of being familiar with operator actions under a variety of plant conditions.

The number of experts needed to make the judgments cannot be stated explicitly. However, as many experts as practical should participate. Generally, direct numerical estimation can be used with fewer experts than can paired comparisons. Seaver and Stillwell (1983) suggest six experts would be sufficient for direct estimation, though, of course, more would be better. They also indicate that under certain circumstances, eight experts could be used for paired comparisons. We recommend having at least 10 to 12 to ensure the necessary statistical reliability.



*Numbers refer to section in Appendix A.

Figure A.2 Process for selection of personnel.

2.3 Data Collection Session Administrator

The session administrator need not have any special qualifications if the tasks are well defined and assumptions are made explicit. The administrator should be familiar with the instructions and be given the opportunity to rehearse the instructions in some form of pretest. The administrator should not have to provide impromptu answers to questions concerning the task definitions. The administrator should, however, be prepared to answer questions about the instructions, and be familiar enough with the procedures to ensure that participants are responding appropriately.

2.4 Data Analyst

If the procedures described in the remainder of this appendix are followed, an individual with a background in mathematics or statistics can perform the calculations. Knowledge of a computer language, such as FORTRAN, is optional but would be helpful so that calculations can be automated. Also, knowledge of a statistical package, such as the Statistical Package for the Social Sciences, can be useful, but is not required.

Table A.1 Number of personnel needed to implement psychological scaling

Personnel	Minimum Number Required	
	Direct Numerical Estimation	Paired Comparisons
Human Reliability Analyst	1	1
Subject Matter Experts	6	10
Data Collection Session Administrator	1	1
Data Analyst	1	1

3. MATERIALS REQUIRED

Three types of materials are required to collect psychological scaling data: task statements, response booklets, and data collection session instructions. Task statements and response booklets are discussed in Sections 3.1 and 3.2, respectively. Sample instructions to be given to experts before they make direct estimates or paired comparisons are discussed in Section 3.3. Figure A.3 provides an overview of the materials.

Materials required to analyze psychological scaling data include: coding sheets (discussed in Section 3.4.1 of this appendix), a calculator with the capability to compute logarithms (or a table of logarithms), and a standard statistics textbook. A computer is optional but recommended. Each of these materials is shown in Figure A.3 and will be discussed in Section 3.4 of this appendix.

3.1 Task Statements

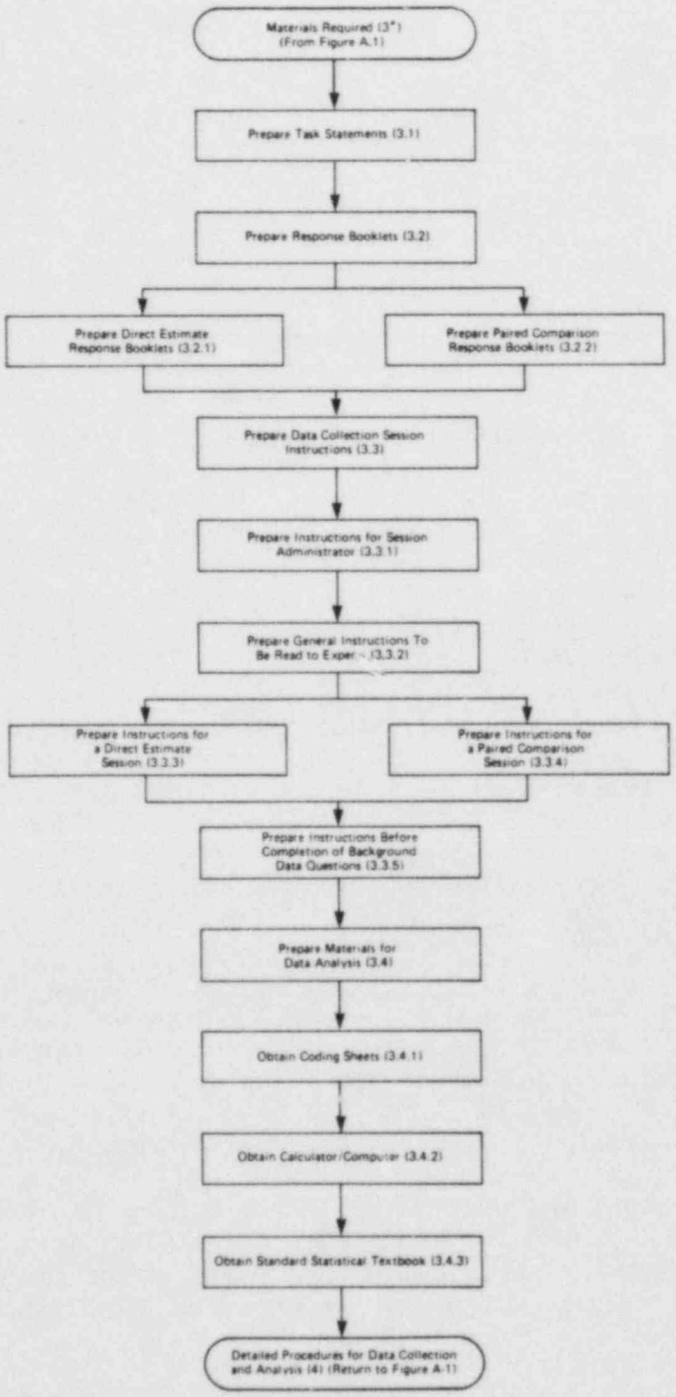
Well-defined task statements are probably the most critical aspect of psychological scaling. The more fully the tasks are specified, the less they will be open to individual interpretation by the experts while making their judgments.

The level of detail of task definitions will vary based on the task itself and the end use of the human error probability (HEP) estimate. Nuclear power plant tasks are categorized at three levels by the Human Reliability Data Bank [NUREG/CR-2744, Volume 2 (Comer et al., 1983)]: systems, components, and displays/instruments/controls. Tasks statements that address a human action involving a higher level (i.e., system) will probably be more complex but less specific than tasks at a lower level of detail.

The clarity of task definitions should be examined in a pretest setting. A few experts should be consulted to ensure that the level of detail of the task definition is sufficient and that explicitly stated assumptions for a task or group of tasks are clear. Specifically, as Seaver, Stillwell, and Schwartz (1982, p. A-1) delineated:

- The role of performance shaping factors (PSFs) in the task should be defined. (PSFs are external or internal factors, such as equipment design or stress, that affect the performance of an individual. See Swain and Guttmann (1983) for a more detailed discussion.)

- Tasks not under consideration, but which might be confused with the tasks to be judged, should be clearly separated from the task under consideration.



*Numbers refer to section in Appendix A.

Figure A.3 Process for material preparation.

- Larger sets of tasks to which this task belongs should be described.

- Causes of the task, e.g., sets of mutually exclusive initiating tasks and sequences of tasks, should be identified.

A final consideration for tasks that will be judged using the paired comparison procedure is that HEP estimates for some of the tasks (at least two) be available from another source, e.g., direct estimates, simulator research data, or the Handbook (Swain and Guttmann, 1983), for use as anchors. Anchors are needed to relate positions on the subjective scale, derived from the experts' judgments, to positions on the probability scale. In most cases, anchors will come from tasks, included among the task statements, for which independent probability estimates are known. These estimates can come from simulator research in which operator performance is observed during simulated tasks, and the frequency of actual errors is recorded and compared to the number of opportunities for error. An HEP can then be calculated based on actual operator performance rather than expert judgment. Estimates can also be taken from the Handbook (Swain and Guttmann, 1983). In some cases, however, none of the tasks will have independent probability estimates, so direct estimates of the anchors must be used. Calculations using anchors are fully described in Section 4 of this appendix.

3.2 Response Booklets

Response booklets of task statements must be prepared. While the wording of the task descriptions will not vary based on the scaling procedure, the response method the expert uses will vary. Thus, response booklets are discussed separately for direct estimates and paired comparisons.

3.2.1 Direct Estimate Response Booklets

A key consideration when using the direct estimate procedure is the type of scale on which experts will indicate their judgments. A scale such as the one in Figure A.4 may be used. Other types of scales are described by Seaver, Stillwell, and Schwartz (1982). It is important that the chosen scale be of sufficient detail that the sensitivity of the expert to differences can be indicated. The scale values must also reflect the estimated range of the true probabilities of the tasks.

Having prepared the task statements and the scale design, response booklets can be prepared. Instructions, assumptions for the task set, and sample items should appear first in the booklet. Sample instructions for direct estimates and sample assumptions for a set of tasks are provided in Figures A.5 and A.6, respectively. Then, the tasks are presented in random order to minimize any effects of task presentation sequence. Thus each expert has a different order of tasks.

EXAMPLE OF COMPLETED DIRECT ESTIMATE

Estimate the chances that:

An operator will read information from a graph incorrectly.

What assumptions did you make that impacted your answer?

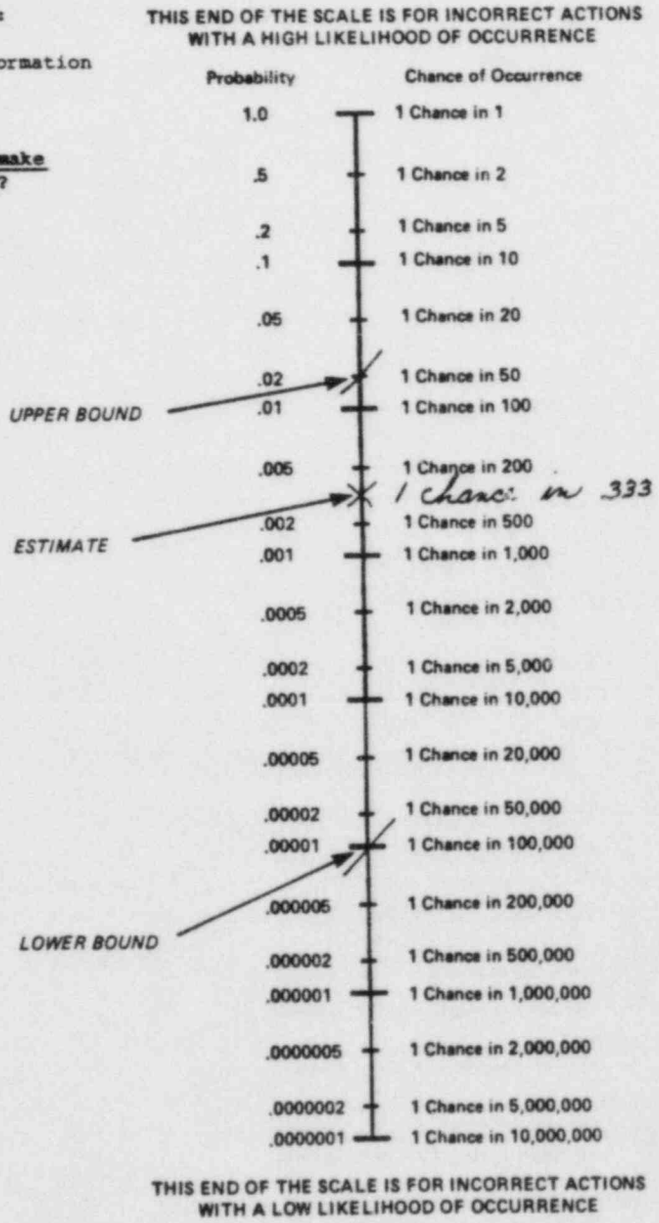


Figure A.4 Sample task statement and response scale for direct estimate.

INSTRUCTIONS FOR COMPLETION OF DIRECT ESTIMATE
AND UNCERTAINTY BOUNDARY JUDGMENTS

Once you have read and understood the task on the left side of the page, put an X on the point on the scale on the right that represents your best estimate of the chances of the incorrect action occurring. Remember, you are to assume that the operator does not have an unlimited amount of time in which to take action. Next, place slash marks to indicate upper and lower bounds so that you are 90 percent certain that the value will fall within those bounds. If a mark or exact value that represents your estimate does not appear on the scale (e.g., 1 chance in 3,500), place your X or slash at the approximate position on the scale and write your estimate to the right of the scale.

Figure A.5 Sample instructions to be included in response booklets for direct estimates.

You are to assume the following for the tasks that follow:

- A senior reactor operator and a reactor operator are in the control room at all times.
- Everything in the task statement that is not underlined is "given" and sets the stage for the underlined question.
- The person(s) performing the action in each task has been in his current job position for at least six months.
- No one involved in performing these tasks is wearing any type of protective clothing.
- The operator(s) does not have an unlimited amount of time in which to take action.

Figure A.6 Sample assumptions to be included in a response booklet.

3.2.2 Paired Comparison Response Booklets

As with direct estimates, instructions, assumptions, and sample items should appear on the first pages of the booklet. Sample instructions and examples are shown in Figure A.7. Assumptions for paired comparisons can also be presented (see Figure A.6). Then, the statement "check the task that is the more likely to occur" should appear on each page as a reminder to the experts.

The pairs should be presented in random order to minimize any effects of presentation order; that is, each expert will have a different order of pairs with all possible orders of pairs being equally likely. Randomization should also be used to alter which of two tasks within a pair comes first on the page to minimize any bias toward tasks in the first or second position.

3.3 Data Collection Session Instructions

In addition to the brief instructions provided in the front of response booklets, detailed instructions are needed for the data collection session administrator. Specifically, instructions provide the administrator with background on psychological scaling procedures. Then, instructions for the administrator to read to the experts for direct estimate or paired comparison sessions are given.

3.3.1 Instructions for Session Administrator

The following sample instructions provide background on a psychological scaling session for the data collection session administrator:

> *The purpose of this session is to allow experts to judge the likelihood that certain incorrect actions will occur during nuclear power plant operation. The data collected during this session will be used to make estimates of human error probability. These estimates will be used in probabilistic risk assessments (PRAs) of nuclear power plants.*
>
> *In this session, the experts will be asked to make judgments about sets of incorrect actions. Each of these actions is part of the more complex behavioral sequences undertaken by an operator in a nuclear power plant. An example of a specific incorrect action is "read the wrong meter in a group of meters that all look very similar and are identified only by labels." This simple action is part of many behavioral sequences that the operator performs. Depending upon the tasks that have been defined, other types of actions may consist of complex behavioral sequences that require several individual actions to be performed correctly for the entire sequence to be successful. For example, the experts could be asked to make judgments about the likelihood of the following situation:*

EXAMPLES OF COMPLETED PAIRED JUDGMENTS

Of the two possible tasks listed below, check the task that is more likely to occur.

X 1. An operator chooses the wrong switch from a set of switches that all look similar and are grouped according to their functions.

_____ 2. A locally operated valve does not have a rising stem or a position indicator. An auxiliary operator, while using written procedures to check a valve lineup, fails to realize that the valve is not in its proper position after a maintenance person has performed a procedure intended to restore it to its proper position after maintenance.

_____ 1. During a loss-of-off-site-power transient, several failures have rendered the high pressure coolant injection (HPCI) and the reactor core isolation cooling (RCIC) systems inoperable. Core cooling can be established with either low pressure coolant injection or low pressure core spray, but pressure must be reduced first. Procedural guidelines specify manual actuation of the automatic depressurization system (ADS) to reduce pressure. What is the likelihood that the operator will fail to actuate the ADS manually within 10 minutes?

X 2. During a loss-of-off-site-power transient, the generator has tripped, the reactor has scrammed, and the normal feedwater system is inoperable. According to the procedures, the reactor water level should be recovered and maintained by manually operating the reactor core isolation cooling (RCIC) system. What is the likelihood that the operator will fail to operate the RCIC system correctly?

Figure A.7 Sample instructions and examples to be included in a response booklet for paired comparisons.

During a loss-of-off-site-power transient, several failures have rendered the high pressure coolant injection (HPCI) and the reactor core isolation cooling (RCIC) systems inoperable. Core cooling can be established with either low pressure coolant injection or low pressure core spray, but pressure must be reduced first. Procedural guidelines specify manual actuation of the automatic depressurization system (ADS) to reduce pressure. What is the likelihood that the operator will fail to actuate the ADS manually within 10 minutes?

If the data collection session involves paired comparisons, the experts will be asked to determine which of a pair of incorrect actions is more likely to occur. The experts will be asked to make judgments about all possible pairs from within each set. If the data collection session involves direct estimates, the experts may make two types of judgments. The first type of judgment will be a direct estimate of the probability of the incorrect action. The expert will be asked to express an estimate of the chances that the incorrect action will occur out of some number of opportunities. For example, the expert will be asked, "What do you think the chances are that an operator will choose the wrong switch from a set of switches that all look similar and are identified only by labels?" The experts will be provided a scale that shows successively lower chances of occurrence of the event, from 1 chance in 1 to 1 chance in 10,000,000, and will be instructed to place a mark on the scale that corresponds to their estimate of the chances that the incorrect action will occur.

The second type of judgment involved in a direct estimate data collection session may be an estimate of the uncertainty about a direct estimate of the probability of an action. The experts will be asked to place bounds around their estimates of the chances of an action's occurrence so that they are certain that 90 percent of the time the actual chances of an incorrect action's occurrence will be within those bounds. For each expert's judgment on each scale, these bounds should surround the mark placed for their exact estimate. These bounds will provide information about the experts' uncertainty about their judgments.

In the final portion of the session, the experts may be asked for information about their experience and training.

Sample questions are provided for the experts in the response booklets. You can use these to ensure that the procedures are correctly understood. With these questions, you are only seeking to determine whether the experts understand the use of the judgmental procedures, not whether they agree with what you think is the "correct" probability. Make no attempt to change their answers except to explain further the type of judgment being asked for if their judgments are inconsistent with what is required by a procedure. An example of inconsistent judgments that should be pointed out to the expert is a case where the mark for the upper uncertainty bound is put below the mark for the error probability. By definition the bounds should surround the mark for the error probability with the upper bound always above and the lower bound below. This sort of inconsistency should be pointed out to the expert and an attempt should be made to reexplain the judgment required.

During the session, if the experts ask questions about the content of the task descriptions, you should not provide impromptu answers. Responses should reiterate any previous instructions, or, as a last resort, the experts should be told that if they must make special assumptions in order to respond, these assumptions should be written in their response booklets. You should explain to the experts that it is important that all experts have the same information so their responses can be compared. This is particularly important if data are collected from different experts in different sessions. If additional guidance is needed to clarify the instructions, please provide it.

3.3.2 General Instructions To Be Read to Experts

The following are sample instructions to be read to experts by the data collection session administrator:

The purpose of this session is to gather judgments of the likelihood of certain events. The events concern various incorrect actions performed in the process of operating a nuclear power plant. During any specific action or operation, for example, closing a valve, there will be a chance that the operator will make an error, that is, fail to close the valve correctly. As experienced instructors, you have as much as or more firsthand knowledge about the chances of incorrect actions than anyone else does. For this reason, we have asked you to participate in the session.

We will be asking you to make judgments about the likelihood of various incorrect actions that might occur during the operation of a nuclear power plant. You should try to incorporate all your knowledge of power plant operations and the likelihood of the various actions into these judgments. As an example, you may know that some of these actions are more difficult or complex. Thus, one might expect the chance of incorrectly performing that action to be higher. Some actions may occur during more stressful situations, so those actions might have a higher likelihood of being performed incorrectly. As you make each judgment, try to think of all information that is relevant to the chances of performing that action incorrectly. You are to assume that the operator does not have an unlimited amount of time in which to take action. He must respond to the system demands prior to the onset of consequences that would result from his inaction. In other words, he must respond within the period of time required by the situation and his specific plant design.

Specific instructions for the judgments will be given as needed, along with examples. You are to make the following assumptions regarding these tasks: These assumptions were used for the system operation tasks in this study. Other assumptions might be appropriate in other situations. For example, the assumptions for other tasks are also given below.

- *A senior reactor operator and a reactor operator are in the control room at all times.*
- *Everything in the task statement that is not underlined is "given" and sets the stage for the underlined question.*

- *The person(s) performing the action in each task has been in his current job position for at least six months.*

- *No one involved in performing these tasks is wearing any type of protective clothing.*

- *The operator(s) does not have an unlimited amount of time in which to take action.*

Another set of assumptions was used in this study for operation of components, instruments, and controls:

- *There is a one-man team in the control room during the performance of these tasks.*

- *These tasks take place during routine operations.*

- *The person performing the action in each task has been in his current job position for at least six months.*

- *No one involved in performing these tasks is wearing any type of protective clothing.*

The assumptions associated with the tasks are clearly labeled in the response booklets.

It is important to have independent judgments from each of you, so please do not discuss your judgments with each other. If you have any questions, please let me know. I will try to answer your questions in a way that does not lead to differences between your judgments and those of others who have not heard your questions and my responses.

3.3.3 Instructions To Be Read to Experts for a Direct Estimate Session

In addition to the introductory instructions, the following sample instructions could be read to experts before beginning a data collection session involving direct estimates:

Review the assumptions on the first page of the response booklet. (Pause.) In addition, assume that typical control room conditions exist. When making the judgments, remember that we are only interested in operator errors, not in any additional equipment failures. You will be giving numerical estimates of the chances that an operator will perform a single action incorrectly. The response booklet shows an example of this type of judgment. The action in this case is

"An operator will incorrectly read information from a graph that is in a procedure."

Your first judgment will be an estimate of the chances that such an error will be made. In making this estimate, you should consider all possible operators

and all circumstances that fit the task description. Taking these possibilities into account, we want your best estimate of how likely this incorrect action is. Would you expect such an incorrect action to occur once out of every ten times these circumstances occur? once out of a thousand? once out of a million? or something in between?

The scale on the right side of the page has been provided for you to mark your estimate. See Figure A.4. The scale is marked with both the chance of occurrence and the corresponding probability. For example, one chance in 100 corresponds to a probability of point zero one. A probability of point zero five is the same as five chances in 100 or one chance in 20. You should put an X on the scale at the point that corresponds to your estimate of the chances or the probability that the given incorrect action will occur.

If the scale does not include the exact chances or probability that you estimate, mark the scale with an X in approximately the correct position and write your estimate to the right of the scale. For example, if you think the given incorrect action would occur about three times in a thousand, you should put an X between one chance in 200 and one chance in 500. This estimate corresponds to one chance in 333 or point zero zero three. In addition to your X, you should write either "1 in 333" or ".003" to the right of the scale. The X labeled "estimate" on the example corresponds to this judgment.

We recognize that you cannot know for sure exactly what the chances of these incorrect actions are. Your response is simply your best estimate. Therefore, we also want to get estimates from you about what you think the range of chances for this incorrect action is. You might think, for example, that while your best estimate is one chance in 333, the actual chances may be quite a bit higher or lower than this estimate, depending on circumstances. Therefore, we will also ask you for upper and lower estimates or bounds that represent the range over which this estimate may vary. Specifically, you should indicate an upper and lower bound so that you think there is a 90 percent chance that in any circumstances the probability of error is between these bounds. In determining these bounds, you should consider the range of circumstances in which this task is performed. This includes different operators (e.g., with different capabilities or training), the physical and mental condition of operators (e.g., tired versus rested, under stress), the quality of instructions, and the physical conditions of the plant (e.g., temperature, layout of controls).

Suppose the upper bound is one chance in 50 and the lower bound is one chance in 100,000. This would indicate that you are quite certain—90 percent sure--that the actual chance of this incorrect action occurring is between these bounds. These bounds would also be marked on the scale as indicated in the example.

The scale provided goes as low as one chance in 10 million. You do not need to use the entire scale unless you think the chances of error are really that low. The scale is provided only so that you may respond as you think appropriate, and not as any guide to what we consider appropriate responses. However, if you use the very top of the scale, where the probabilities are between .5 and 1.0, it is particularly important that you write in the actual probability.

Each page has a place for you to list any assumptions that you might have made when making your estimate. You are not required to fill in this information for each task. Factors that might be listed include such things as time of day, environmental conditions in the control room, and quality of procedures. Indicate for each assumption whether it applies to the best estimate or to the uncertainty bounds or both.

Now, if you have any questions about how you are to give these estimates, I will try to answer them.

(ADMINISTRATOR: Answer questions.)

If there are no further questions, on the next two pages of the booklet are examples for which you should mark your best estimate and your uncertainty bounds. After you have completed these examples, I will check your responses to be sure they are consistent with the kinds of responses we are looking for.

After I have examined your responses to the sample questions, you will be free to proceed through the booklet. The tasks appear, one per page, with a scale to mark your responses. Estimate the chances of occurrence and upper and lower uncertainty bounds for each. You are free to turn back to previous pages once you have completed them. The assumptions for the tasks are presented prior to the questions on the tasks. If you do not have any questions, proceed with the judgments.

3.3.4 Instructions To Be Read to Experts for a Paired Comparison Session

In addition to the introductory instructions given above, the following sample instructions could be read to experts before beginning a data collection session involving paired comparisons:

Review the assumptions on the first page of the response booklet. (Pause.) In addition, assume that typical control room conditions exist. When making the judgments, remember that we are only interested in operator errors, not in any additional equipment failures. You will be shown tasks in pairs. Each task involves an incorrect action that an operator could take. For each pair, decide which of the two incorrect actions is more likely to occur. Thus, a very difficult action, even though the operator might not perform it often, should have a higher relative chance of being performed incorrectly than an easier action. Remember that you are not trying to determine which task describes a better or worse operating situation or control design. Rather, you are simply judging which task an operator is more likely to perform incorrectly. Mark your choice with a checkmark in the space provided.

Examine the completed example in your response booklet. See Figure A.7 . The first incorrect action was checked. For our hypothetical respondent, this reflects the belief that action 1 is more likely to occur out of the chances it has to occur than action 2. The second example shows that our hypothetical respondent believes that action 2 is more likely to occur than action 1.

We would like you to make a choice for each pair of actions. Do not leave any pair of actions unchecked, and do not check both actions of any one pair. If you

are unsure of the relative likelihood of the two actions, make your best guess as to which of the two is more likely.

At this time, please turn to the next page in your response booklet. You should find two uncompleted examples. Mark these examples as you have been instructed. Are there any questions about the procedure?

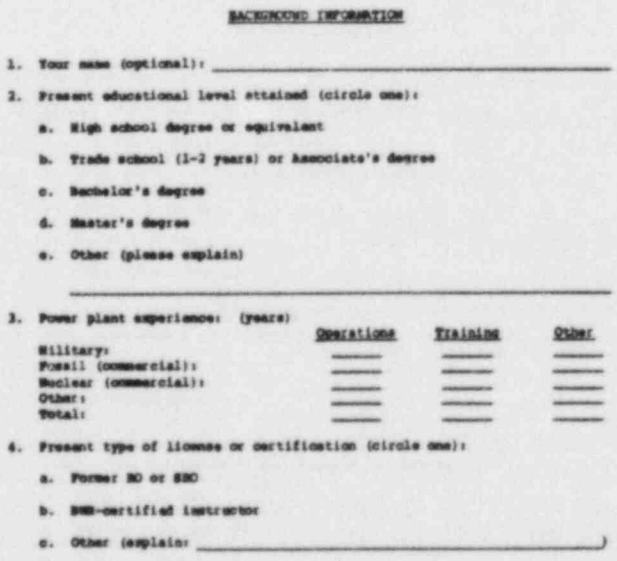
After you have completed all responses, please give me your response booklet. If you have any questions while you are making the judgments, please let me know.

3.3.5 Instructions To Be Read to Experts Before Completion of Background Data Questions

The experts' education and experience may provide useful information on the level of expertise of the participants. In addition to providing documentation on the source of judgments, the data can also be useful for research projects involving psychological scaling. A sample set of items is provided in Figure A.8.

The following sample instructions could be read to experts at the conclusion of a data collection session:

You have now finished all the judgments on the incorrect actions. We have one final request, which is that you answer the questions provided. You do not have to give your name, but it would be helpful to us so that we can follow up on any of your comments and ask questions if we need to. If you do give your name, it will be kept confidential. The questions about your past experience are for our information only. Any additional comments you have are welcome and can be entered in the space provided. If you have any questions, feel free to ask them. Otherwise, proceed with the questions.



BACKGROUND INFORMATION

1. Your name (optional): _____

2. Present educational level attained (circle one):

   a. High school degree or equivalent

   b. Trade school (1-2 years) or Associate's degree

   c. Bachelor's degree

   d. Master's degree

   e. Other (please explain)

   _____

3. Power plant experience: (years)

	Operations	Training	Other
Military:	_____	_____	_____
Fossil (commercial):	_____	_____	_____
Nuclear (commercial):	_____	_____	_____
Other:	_____	_____	_____
Total:	_____	_____	_____

4. Present type of license or certification (circle one):

   a. Former RO or SRO

   b. NRR-certified instructor

   c. Other (explain: _____)

Figure A.8 Sample items concerning expert background.

3.4 Materials Required for Data Analysis

Materials to be used in support of data analysis include coding sheets, a calculator, a computer (optional), and a standard statistical textbook. Each of these is discussed below.

3.4.1 Coding Sheets

Sample coding sheets for direct estimates and paired comparisons are shown in Figures A.9 and A.10, respectively. For direct estimates, actual probabilities chosen by each expert can be entered in Figure A.9 along with lower and upper uncertainty bounds for each task.

For paired comparisons, a matrix can be completed for each expert (see Figure A.10). A "1" is entered in the matrix if the task listed horizontally across the top of the matrix was selected by the expert as more likely than the task listed vertically. The "1" is placed at the intersection of the task numbers.

A "0" is placed in the matrix if the task listed horizontally was not selected as more likely than the task listed vertically. The diagonal of the matrix is not filled in because it makes no sense to compare a task with itself. The lower half of the matrix, which is darkened in Figure A.9, does not need to be filled in, since it would contain information equivalent to that in the upper half of the matrix. For example, if Task 2 was chosen more likely than Task 1, a "1" is filled in at the intersection of Task 2 and Task 1 in the upper half. At the intersection of Task 1 and Task 2 in the lower half, a "0" would be placed because if an expert judged Task 2 more likely than Task 1, then Task 1 must necessarily be less likely than Task 2.

The coding sheets format the data so that it can be used easily in the statistical calculations described in Section 4 of this appendix.

3.4.2 Calculator and Computer

Logarithms/antilogarithms and proportions of area under the normal curve (involved in paired comparison calculations only) can be found in tables provided in most standard statistical textbooks. However, the use of a calculator or computer with these capabilities will be less time consuming. A calculator or computer will also decrease the time required for calculation of other descriptive statistics needed to derive HEP estimates from data collected using psychological scaling.

3.4.3 Standard Statistical Textbook

A standard statistics textbook will also be needed for the tables it provides. In particular, tables for the normal distribution will be needed if a calculator or computer subroutine is not used to obtain normal deviates corresponding to proportions of area under the normal curve. Two other tables will also be needed for determining whether the consistency of judgments is adequate: a table showing the statistical significance of chi-squared values and a table showing the statistical significance of correlation coefficients.

Subject # _____

Level A Task #	Estimate	Uncertainty Bounds: Lower	Uncertainty Bounds: Upper
1.	_.________	_.________	_.________
2.	_.________	_.________	_.________
3.	_.________	_.________	_.________
4.	_.________	_.________	_.________
5.	_.________	_.________	_.________
6.	_.________	_.________	_.________
7.	_.________	_.________	_.________
8.	_.________	_.________	_.________
9.	_.________	_.________	_.________
10.	_.________	_.________	_.________
11.	_.________	_.________	_.________
12.	_.________	_.________	_.________
13.	_.________	_.________	_.________
14.	_.________	_.________	_.________
15.	_.________	_.________	_.________

Figure A.9 Sample coding sheet for direct estimate data.

Subject # _____

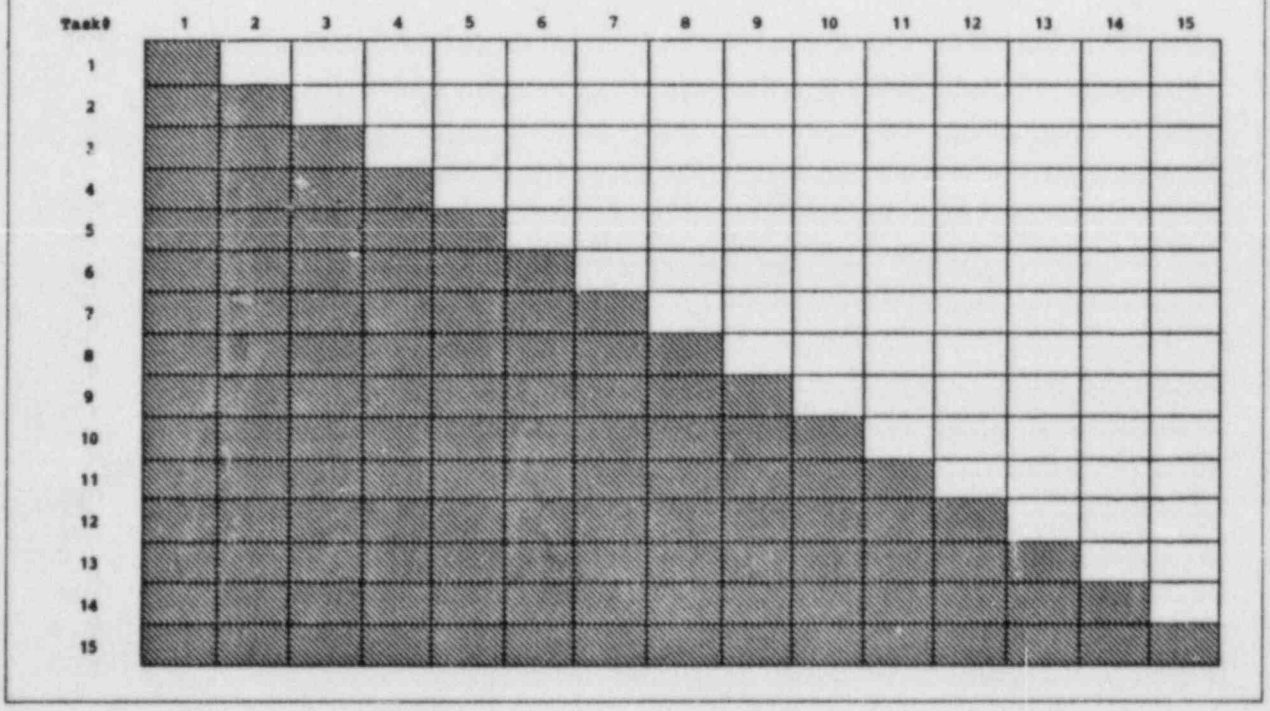


Figure A.10 Sample coding sheet for paired comparison data.

4. DETAILED PROCEDURES

In this section, the procedures for collecting the required judgments and analyzing them to obtain HEP estimates are described. These procedures draw heavily upon the materials described in Section 3. Section 4.1 discusses the data collection procedures. Section 4.2 describes the process by which direct numerical estimates are analyzed, and Section 4.3 provides a similar description for paired comparison scaling.

4.1 Data Collection

The session administrator should arrive at the site at which data are to be collected well ahead of the scheduled time. The administrator should ensure that there is an adequate number of tables and chairs for all experts. The administrator should bring:

- enough sharpened pencils for each expert to have two, plus several extra;
- instructions to be read to the experts;
- enough response booklets for each expert to have one, plus a few extras.

As the experts arrive, the administrator should space them out in the room as much as possible to provide adequate working space and reduce the potential for experts inadvertently disrupting each other. Since the tasks have been randomly ordered in the response booklets, there is no reason to separate the experts to avoid sharing of information. Therefore, if the experts cannot be separated, the impact on the results should be minimal.

When all experts have arrived, the administrator should read the general instructions describing the purpose of the data collection session. These instructions indicate that experts can ask questions regarding the data collection, so the administrator should respond to all questions. It is important, however, that all responses to questions only explain the process and not convey any new information that might influence interpretations of the task descriptions. This is particularly important when data are collected in multiple sessions with different experts. When questions cannot be answered by simply reiterating or rewording instructions, phrases such as "use your best judgment," "take into account everything you know," and "if you need to make specific assumptions, do so, but be sure to write them in your response booklet" may be used.

After these initial instructions, the administrator should pass out response booklets and two pencils to each expert. Then, specific instructions (i.e., either for paired comparisons or for direct numerical estimation) should be read. After example responses have been completed, each expert's responses should be checked by the administrator to be sure

they are appropriate. For paired comparisons, the administrator should reiterate that the marked task in the pair should be more likely to produce a human error. For direct estimates, in checking responses, if uncertainty bounds are being estimated, the administrator should be sure they have been estimated, as well as the nominal HEP, and that the nominal estimate is between the bounds. Once all example responses have been checked, and any questions have been answered, the experts should be told to continue making judgments until all required judgments have been made.

As experts complete their judgments, the administrator should collect the response booklets. Each booklet should be checked to be sure all responses have been made. Experience indicates that particular attention should be given to ensuring that all uncertainty bound estimates have been made.

After all the experts have completed the judgments, they should be encouraged to remain for a few minutes to discuss the procedures. This discussion should identify any difficulties they encountered, anything they felt particularly uncomfortable with, and any other thoughts they might have on the procedures and their use.

Following the session, the administrator should be responsible for delivering the response booklets to the data analyst for coding and analysis.

4.2 Direct Numerical Estimation

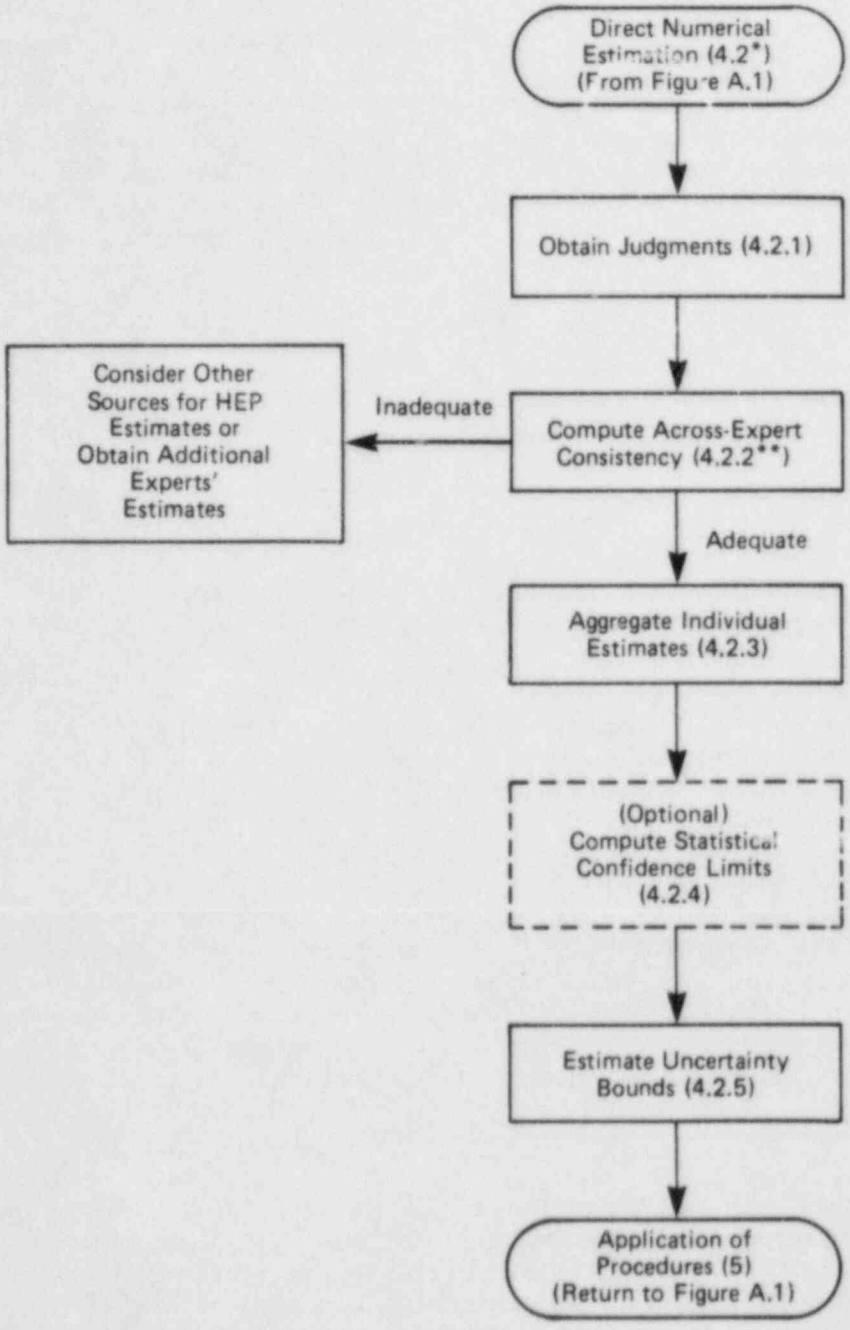
Direct numerical estimation is a relatively straightforward procedure. The major steps are listed in Figure A.11 and described in detail below. These steps can be used to obtain both HEP estimates and uncertainty bounds on the estimates. The procedure for estimating HEPs is described first, followed by a description of adaptations necessary for uncertainty bounds and some possible variations on obtaining uncertainty bounds.

4.2.1 Judgments Required

Each expert must estimate the HEP for each of the tasks under consideration. These estimates should be made on a scale such as described in Section 3.2.1. This scale allows experts to think in terms of either probabilities or "chances," e.g., one chance in a thousand. While these judgments appear to be straightforward, adequate instructions such as those given in Section 3 are important, and the data collection session administrator should ensure that the type of judgment required is understood by each expert.

4.2.2 Across-Expert Consistency

The HEP estimates derived as described above should be used only if there is reasonable agreement among the experts regarding the estimates. Lack



*Numbers refer to section in Appendix A.

**Within-expert consistency is not calculated since each expert provides one estimate for each task.

Figure A.11 Major steps in using direct numerical estimation.

of agreement suggests that the computed HEP estimate may not be appropriate because different experts may be interpreting the tasks differently or making very different assumptions regarding the tasks. If the extent of agreement as determined below is inadequate, task definitions and assumptions should be reviewed with the experts. Then additional data should be obtained, or another source for estimating the HEP should be used.

An appropriate, relatively simple measure of agreement is the Kendall coefficient of concordance, W. This measure can intuitively be thought of as an average correlation among the various experts. It is based only on the rank order of the HEP estimates for each expert, so these estimates must be converted into rank orders.

The formula for the Kendall coefficient of concordance is

$$W = \frac{12SS}{m^2 (n^3 - n)} \qquad (1)$$

where m is the number of experts, n is the number of tasks, and SS is the sum of squares:

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

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

Follow the instructions below to compute the Kendall coefficient of concordance (see Table A.2 for computation example.)

1. Put all HEP estimates into a table with the estimates of each expert in one row.

2. Derive a similar table of rank orders for each expert.

3. Calculate the sum of ranks for each task.

4. Calculate the average sum of ranks as m(n+1)/2.

5. Compute the difference of the sum of ranks for each task (Step 3) and the average sum (Step 4).

6. Square the differences computed in Step 5.

7. Sum the squared differences computed in Step 6.

8. Use the sum of squared differences from Step 7 and m and n in equation (4) to compute the coefficient of concordance.

Table A.2 Computation of coefficient of concordance to measure across-expert consistency

	HEP Estimates					Rank Order				
	Task					Task				
Expert	1	2	3	4	5	1	2	3	4	5
1	.02	.0001	.001	.1	.5	3	5	4	2	1
2	.005	.003	.02	.06	.35	4	5	3	2	1
3	.009	.008	.1	.02	.06	4	5	1	3	2
4	.015	.0004	.003	.01	.2	2	5	4	3	1
5	.004	.0023	.006	.15	.15	4	5	3	1.5	1.5
6	.01	.0009	.01	.008	.4	2.5	5	2.5	4	1
7	.006	.0013	.006	.05	.3	3.5	5	3.5	2	1
sum of ranks, R_j						23	35	21	17.5	8.5
average sum of ranks, $\bar{R} = \frac{m(n+1)}{2} = 21$										
$R_j - \bar{R}$ =						2	14	0	-3.5	-12.5
$(R_j - \bar{R})^2$ =						4	196	0	12.25	156.25

$$SS = \sum_{j=1}^{n} (R_j - \bar{R})^2 = 4 + 196 + 0 + 12.25 + 156.25 = 368.5$$

$$W = \frac{12(368.5)}{7^2(5^3 - 5)} = \frac{4422}{(49)(125-5)} = \frac{4422}{(49)(120)} = \frac{4422}{5880} = .75$$

$$T = m(n-1)W = 7(5-1)(.75) = 21.0$$

This Kendall coefficient of concordance provides a measure on a zero-to-one scale where 0 is no agreement and 1 is complete agreement. A value, T, related to W can also be tested for statistical significance. The value

$$T = m(n-1)W \qquad (3)$$

is distributed approximately like chi-square with n - 1 degrees of freedom. (This approximation is true only if there are more than seven tasks as would be true in most applications. Although this example has only five tasks, it is interpreted as if the approximation were true.) The significance of this value can be determined from chi-square tables

found in most statistics texts. In this example, T is equal to 21 with 4 degrees of freedom. Examination of a chi-square table indicates that a value of 14.86 is needed for significance at the .005 level. Thus, since the value of 21 is greater than 14.86, the significance is less than .005. As a general rule a significance level of .05 or less should be acceptable. Thus, we conclude that there is sufficient agreement among the experts because the significance level is less than .05. If the number of tasks is seven or less, the statistical significance of W can be determined from Appendix C of Seaver and Stillwell (1983).

4.2.3 Aggregating Individual Experts' Estimates

The individual experts' HEP estimates must be aggregated into a single estimate for each task. The lower and upper uncertainty bounds for these HEP estimates must also be aggregated. The procedure described in this section applies to aggregation of both HEP estimates and uncertainty bound estimates, even though the example only illustrates aggregation of HEP estimates. Prior to this aggregation, however, it is desirable to eliminate any unusually large or small estimates that are out of line with the estimates of most other experts. To identify these "outliers," the standard deviation of the estimates is first computed. This computation should be performed on logarithms of the HEP estimates. The formula for this computation is

$$\text{s.d.} = \sqrt{\frac{m \sum_{i=1}^{m} (\log \text{HEP}_i)^2 - \left(\sum_{i=1}^{m} \log \text{HEP}_i \right)^2}{m(m-1)}} \qquad (4)$$

where m is the number of experts and $\log \text{HEP}_i$ is the logarithm of the HEP estimate of expert i.

Table A.3 gives an example of how this standard deviation is calculated using data for one HEP, and only six of the seven experts from the previous example for computing the coefficient of concordance. The steps in this calculation are:

1. Find the logarithm of each expert's HEP estimate.
2. Calculate the sum of the logarithms of HEP estimates.
3. Find the square of the logarithm of each expert's HEP estimates.
4. Compute the sum of the squares of logarithms.
5. Use these computed values in equation (4) to compute the standard deviation, as shown in the example.

This standard deviation is then used in the following steps to determine which estimates are outliers.

6. Compute two times the standard deviation.

7. Compute the mean of the logarithms of HEPs, which is the sum computed in Step 2 divided by the number of experts.

Table A.3 Calculation of standard deviation for direct estimates of HEPs

Expert	HEP	log HEP	$(\log HEP)^2$
1	.02	-1.699	2.887
2	.005	-2.301	5.295
3	.009	-2.046	4.186
4	.015	-1.824	3.327
5	.004	-2.398	5.750
6	.01	-2.000	4.000
sum		-12.268	25.445

$$\text{s.d.} = \sqrt{\frac{6(25.445) - (-12.268)^2}{6(6-1)}}$$

$$= \sqrt{\frac{152.670 - 150.504}{(6)(5)}}$$

$$= \sqrt{\frac{2.166}{30}}$$

$$= \sqrt{.0722}$$

$$= .269$$

$$2\text{s.d.} = .538$$

$$\text{mean} = \frac{-12.268}{6} = -2.045$$

$$\text{mean} + 2\text{s.d.} = -2.045 + .538 = -1.507$$

$$\text{mean} - 2\text{s.d.} = -2.045 - .538 = -2.583$$

8. Compute the mean (Step 7) plus two times the standard deviation (Step 6).

9. Compute the mean (Step 7) minus two times the standard deviation (Step 6).

10. Throw out estimates for which the logarithm of the HEP estimate is either above the mean plus two standard deviations from Step 8, or below the mean minus two standard deviations from Step 9.

In the example in Table A.3, none of the estimates are thrown out by this procedure, so all are retained and aggregated into a single estimate.

The formula used to aggregate the individual HEP estimates is

$$HEP = \frac{\left(\prod_{i=1}^{m} HEP_i\right)^{1/m}}{\left(\prod_{i=1}^{m} (1 - HEP_i)\right)^{1/m} + \left(\prod_{i=1}^{m} HEP_i\right)^{1/m}} \qquad (5)$$

where again m is the number of experts and HEP_i is the estimate of expert i. Actual computation is simpler if logarithms of HEP estimates are used. The formula then becomes

$$HEP = \frac{\text{antilog}\left(\left(\sum_{i=1}^{m} \log HEP_i\right)/m\right)}{\text{antilog}\left(\left[\sum_{i=1}^{m} \log (1 - HEP_i)\right]/m\right) + \text{antilog}\left(\left(\sum_{i=1}^{m} \log HEP_i\right)/m\right)} \quad (6)$$

Using the HEP estimates from Table A.3, Table A.4 demonstrates this computation, which is composed of the following steps:

1. Find the logarithm of the HEP estimate for each expert.
2. Compute the sum of the logarithm from Step 1.
3. Compute one minus the HEP estimate for each expert.
4. Find the logarithm of one minus the HEP estimate for each expert.
5. Compute the sum of the logarithms from Step 4.
6. Divide the sum from Step 2 by m, the number of experts.
7. Divide the sum from Step 5 by m.
8. Find the antilogarithm of the value computed in Step 6.
9. Find the antilogarithm of the value computed in Step 7.
10. Compute the HEP estimate by dividing the value from Step 8 by the sum of the values from Steps 8 and 9.

The estimate produced in Step 10 is the HEP estimate to be used.

Table A.4 Aggregation of individual experts' estimates into a single estimate

Expert	HEP	log HEP	1-HEP	log (1-HEP)
1	.02	-1.699	.98	-.00877
2	.005	-2.301	.995	-.00218
3	.009	-2.046	.991	-.00393
4	.015	-1.824	.985	-.00656
5	.004	-2.398	.996	-.00174
6	.01	-2.000	.99	-.00436
sum		-12.268		-.02754
sum/m		- 2.0447		-.00459
antilog		.00902		.989

$$HEP = \frac{.00902}{.989 + .00902}$$

$$= .00904$$

4.2.4 Computing Statistical Confidence Limits

This is an optional step that may be taken to determine what degree of statistical variation can be expected in the HEP estimates. This variation is based on the variability of the estimates of different experts. Thus, it will be larger as there is more variation in experts' estimates and smaller with less variation.

The approximate 95% statistical confidence limits are based on the standard deviation (in logarithms) as computed in Section 4.2.3. In this case, however, any estimates that were identified as outliers by the Section 4.2.3 procedure are not included in the computation.

The basic value used to determine statistical confidence limits is the standard error,

$$s.e. = s.d./\sqrt{m}, \qquad (7)$$

where s.d. is the standard deviation calculated as described in Section 4.2.3 (with outliers excluded) and m is the number of experts. This value is calculated using the following steps.

1. Compute the standard deviation, s.d., as described in Section 4.2.3 and shown in Table A.3 with outliers excluded.

2. Use equation (7) to calculate the standard error (in logarithms).

In the example, this is $.269/\sqrt{6} = .110$. Statistical confidence limits are then derived using the standard error as follows.

3. Multiply the standard error by two (2s.e. = .220).

4. Subtract the value in Step 3 from the log of the HEP found in Step 10 associated with Table A.4 (-2.044 - .220 = -2.264).

5. Find the antilogarithm of the value from Step 4 (antilog -2.264 = .0054). This is the lower statistical confidence limit.

6. Add the value from Step 3 to the mean log HEP found in Step 7 associated with Table A.4 (-.2.044 + .220 = -1.824).

7. Find the antilogarithm of the value from Step 6 (antilog -1.824 = .015). This is the upper statistical confidence limit.

These statistical confidence limits provide what can be intuitively considered the probable range of variation that would be found if the same experts, without remembering their previous responses, or similar groups of experts made these same judgments many times.

4.2.5 Estimating Uncertainty Bounds

Uncertainty bounds can also be estimated using direct estimation. These uncertainty bounds represent the range of HEPs that might occur under varying performance shaping factors, e.g., different levels of operator training, different plant designs, and varying quality of written instructions. They differ from statistical confidence limits in that the statistical confidence limits pertain to HEPs under typical conditions, while the uncertainty bounds include the variation associated with more extreme, a typical conditions.

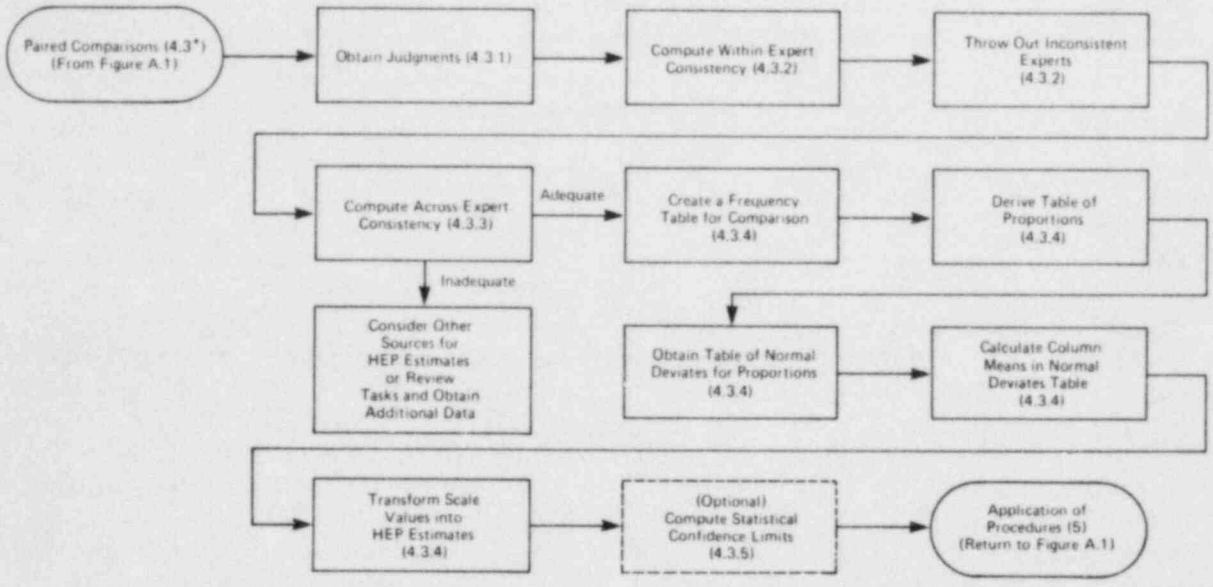
The uncertainty bounds are estimated using the same basic procedures described above for HEPs. Each expert is asked to estimate upper and lower uncertainty bounds using instructions such as those given in Section 3.3. Upper and lower bounds are then derived by aggregating the individual experts' estimates using the procedures described in Section 4.2.3. Across-expert consistency should also be checked as in Section 4.2.2 for the uncertainty bounds estimates. If the consistency is not acceptable, a conservative approach to estimating the uncertainty bounds may be used based on statistical confidence limits on the uncertainty bounds. With inadequate consistency, statistical confidence limits for the bounds should be computed as described in Section 4.2.4. Then, for the lower uncertainty bound, the lower statistical confidence limit of the estimated lower bound should be used. Similarly, for the upper uncertainty bound, the upper statistical confidence limit of the estimated upper bound should be used.

Depending on the required use, some variations on the uncertainty bounds are possible. For example, in some uses a worst-case or near worst-case HEP may be desired. In this instance, a worst-case scenario could be created in advance, or the experts could be asked to create such scenarios as they make the estimates.

4.3 Paired Comparison Scaling

Paired comparison scaling uses relatively simple judgments of experts regarding which of two tasks is more likely to produce an error. The HEP estimates are derived from these simple judgments. Since tasks are compared to other tasks within the same set, there may be some variation in estimates depending on the particular set. The types of tasks that should be included together in a set are described in Section 5.2.

A summary of the procedure for paired comparison scaling is shown in Figure A.12. The computation of the HEP estimates, while not difficult, is time consuming and for any relatively large number of tasks (e.g., 10 or more) should probably be done by computer. It can also involve a regression analysis on a few data points (e.g., four), which can also be best performed using a statistical package, although it can be done by hand using the procedures described below. In addition, the computation of statistical confidence limits (which is optional) as described in Section 4.3.5 requires somewhat more knowledge of statistics than any other aspect of either direct estimation or paired comparison scaling, and cannot be performed efficiently without a computer.



*Numbers refer to section in Appendix A.

Figure A.12 Major steps in using paired comparisons.

With paired comparisons, the same set of judgments used to generate HEP estimates cannot be used to obtain uncertainty bound estimates. It can, however, be used to estimate worst-case and/or best-case HEPs for various tasks in separate sets of comparisons. To do so, the procedures described below are simply repeated for the best- and/or worst-case scenarios, which should be defined prior to obtaining judgments to ensure that all experts are responding to the same tasks. If such scenarios are used, they should be included in different sets of tasks than the "typical" conditions tasks; that is, for example, worst-case scenario tasks should only be compared with other worst-case scenario tasks.

4.3.1 Judgments Required

Each expert is presented with several pairs of tasks and makes a discrete choice of which of each pair is the more likely to produce a human error. The judgment "equally likely" is not allowed. Examples of the type of judgment required were shown in Section 3.2.

An important consideration for the use of paired comparison techniques is the large number of pairs for even a moderate numbers of tasks. The number of pairs to be judged by each expert is $\frac{n(n-1)}{2}$ where n is the number of tasks. For example, for 20 tasks there are 190 pairs of tasks $[(20 \times 19)/2]$. All of these judgments are not necessarily required to get good estimates of the scale values, however, and several suggestions have been made for reducing the number of judgments required (see Torgerson, 1958; or Seaver and Stillwell, 1983).

Probably the most appropriate procedure for the judgment of human error probabilities is to select a limited number of tasks as standards. As much as possible, standards selected should be spaced out over the range of the HEPs. Each of the remaining tasks is then compared with each standard, giving $mn - m\frac{(m+1)}{2}$ independent judgments where n is the number of tasks and m is the number of standards. For example, with 20 tasks, 5 of which are taken as standards, the required judgments would be reduced from 190 to 85 $[(5)(20)-5(6/2)]$. It should be noted that to achieve the same degree of statistical confidence in estimates derived using this procedure, more experts may be required (Seaver and Stillwell, 1983).

4.3.2 Within-Expert Consistency

Before an expert's judgments are included in the data to be analyzed for estimating HEPs, a check should be made to be sure that the expert's judgments are internally consistent. Lack of consistency usually indicates that the expert does not understand the judgments required or does not have enough information to make the judgments. In such cases, the data of that expert should be disregarded and not used in determining HEP estimates.

For paired comparison judgments, internal consistency can be measured by the number of intransitive triads in the expert's judgments. 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. The coefficient of consistency, k, can be used to measure internal consistency (David, 1963). This coefficient ranges from 0 for a completely random (maximum number of intransitive triads) set of judgments to 1 for a completely consistent set (no intransitive triads).

The formula for the coefficient of consistency is

$$k = 1 - \frac{24c}{n(n^2-1)} \quad \text{if } n \text{ is odd} \tag{8}$$

and

$$k = 1 - \frac{24c}{n(n^2-4)} \quad \text{if } n \text{ is even.} \tag{9}$$

In these formulas, n is the number of tasks and C and T are intermediate quantities that are calculated as:

$$c = \frac{n}{24}(n^2 - 1) - \frac{T}{2}, \tag{10}$$

$$T = \sum_{i=1}^{n} (a_i - \bar{a})^2, \tag{11}$$

$$\text{and } \bar{a} = \frac{(n-1)}{2}. \tag{12}$$

The values of a_i are the number of times event i was judged more likely than any other task.

To compute this coefficient, each expert's judgments are put into a table such as Table A.5 where a "1" indicates that the column task was judged more likely than the row task, and a "0" indicates the reverse judgment was made. As illustrated in this table, the steps in computing the coefficient of consistency are:

1. Sum each column. These sums are the values of a_i.
2. Compute $\bar{a}$ as given by the formula.
3. Subtract $\bar{a}$ (Step 2) from each a_i (Step 1).
4. Square each of the values found in Step 3.
5. Compute T, which is the sum of the values from Step 4.
6. Compute c from the formula given in the table.
7. Determine whether n is odd or even and select the appropriate formula from the two given above.
8. Compute k using formula 8 or 9 depending on whether n is odd or even.

The coefficient of consistency does not have an exact statistical test for significance. It can, however, be interpreted as a measure similar to a correlation coefficient. A rule for determining whether the expert is sufficiently consistent is then to treat this coefficient as a correlation and find its significance. Tables of significance levels for correlations can be found in most statistics texts. A significance level of .05 or lower should be obtained or the expert's data should be disregarded.

Table A.5 Example of computation of the coefficient of consistency to measure within-expert consistency in paired comparison judgments

TASK	1	2	3	4	5	6
1	-	1	0	1	0	0
2	0	-	1	1	0	0
3	1	0	-	1	1	1
4	0	0	0	-	0	1
5	1	1	0	1	-	1
6	1	1	0	0	0	-
a_i =	3	3	1	4	1	3

$\bar{a} = (n - 1)/2 = (6 - 1)/2 = 2.5$

$a_i - \bar{a}$ =	.5	.5	-1.5	1.5	-1.5	.5
$(a_i - \bar{a})^2$ =	.25	.25	2.25	2.25	2.25	.25

$$T = \sum_{i=1}^{n} (a_i - \bar{a})^2 = .25 + .25 + 2.25 + 2.25 + 2.25 + .25 = 7.5$$

$$c = \frac{n}{24}(n^2-1) - \frac{T}{2} = \frac{6}{24}(6^2-1) - \frac{7.5}{2} = \frac{1}{4}(36-1) - 3.75 = \frac{1}{4}(35) - 3.75 = 5.0$$

$$k = 1 - \frac{24c}{n(n^2-4)} = 1 - \frac{(24)(5)}{6(6^2-4)} = 1 - \frac{120}{6(36-4)} = 1 - \frac{120}{192} = .375$$

(n is even)

4.3.3 Across-Expert Consistency

As with direct estimation, there should be reasonable agreement among the experts if the HEP estimates are to be used. (See Section 4.2.3.) The measure of across-expert consistency is the same as that used for direct estimation, the coefficient of concordance. Computation of this coefficient requires that the tasks being judged be rank-ordered for each expert. The rank order for an expert is derived by a count of the number of times each task was judged by that expert to be more likely than another task. The task with the largest count is ranked first, and so on.

Once these rank orders have been determined, the procedure for computing the coefficient of concordance is the same as for direct estimation. Section 4.2.2 describes this procedure.

4.3.4 Computing HEP Estimates

Once within-expert and across-expert consistencies have been established, the first step in deriving HEP estimates is to create a frequency table with tasks listed across the top and down the side. Each cell entry in the table is the number of times an expert judged the task listed at the top of the column to be more likely than the task listed at the side of the row. That is, this table is simply the sum of the coding sheets for each individual expert. Table A.6 is an example of a frequency table with judgments from 20 experts.

Table A.6 Frequency table for paired comparison judgments

Task	1	2	3	4	5	6
1	-	15	13	11	17	19
2	5	-	9	5	12	16
3	7	11	-	8	13	17
4	9	15	12	-	16	18
5	3	8	7	4	-	12
6	1	4	3	2	8	-

The frequencies in Table A.6 must then be converted to proportions, as shown in Table A.7. To do this, each entry in Table A.6 is divided by m, the number of experts (20 in this example) to produce a table such as the one shown in Table A.7.

Table A.7 Table of proportions*

Task	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 task listed across the top was more likely than the task listed down the side.

Although it does not occur in this example, in some instances the proportions of 0 or 1 may occur in the table of proportions. The z values or normal deviates for these proportions, which are required in the next step, are plus and minus infinity, respectively, which clearly cannot be used in the calculations. In such cases, the proportion $1/2(m+1)$ should be substituted for 0 and $\frac{(2m+1)}{2(m+1)}$ should be substituted for 1, where m is the number of experts.

The next step is to convert the proportions in Table A.7 into normal deviates, or z values, reflecting the assumption that the proportions 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 (or a calculator or computer, if available). For example, cell entry row 4, column 1, shows that 9 of 20, or 45 percent, of the judges stated that task 1 was more likely than task 4, while 11 of 20, or 55 percent, said the opposite. A table of the normal distribution shows that a z value of -.13 leaves 45 percent of the area of the normal distribution to the left. This z value represents the relative distance between events 4 and 1. Transforming each of the proportions in Table A.7 into unit normal deviates in this manner gives the values shown in Table A.8.

These normal deviates are summed and the mean calculated for each column as shown in Table A.6. This column mean is the value for the event on the newly created subjective scale. For example, the scale now looks like this:

Task No.	1	4	3	2	5	6
Scale Value	-.64	-.49	-.15	.06	.38	.84

Table A.8 Values of proportions under normal curve*

Task	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 Values $= \frac{z}{n}$	-.64	.06	-.15	-.49	.38	.84

where n = the number of tasks (six in this case).

*Each cell entry represents the normal deviate value (z) corresponding to the proportion for the cell shown in Table A.7.

This subjective scale of relative distances must now be converted into a scale of probabilities. A pair of anchors is required that relates positions on the subjective scale to those on the probability scale. In most cases, these anchors will come from tasks, placed for judgment among the others, for which the other probability estimates are available. In some cases, however, none of the tasks will have such probability estimates available, and direct estimates of the anchors must be made using the direct numerical estimation procedure described in Section 4.2. These direct estimates should be made after the paired comparisons so that the tasks used for the anchor judgments are appropriately selected. When just two tasks are to be used as anchors, they should be the tasks with the lowest and highest scale values. In this example, estimates should be obtained for Task 1 (lowest value) and Task 6 (highest value).

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

$$\log HEP = as + b \qquad (13)$$

where s is the scale value derived above, and a and b are constants, determined by simultaneous solution of the two variations of the above equation that result from the two anchors. In this example, anchor values are assumed to be known to be .0004 for Task 1 and .01 for Task 6, and thus the following two equations would be solved:

$$\begin{aligned} \log(.0004) &= a(-.64) + b \\ \log(.01) &= a(.84) + b \\ \hline \log(.0004) - \log(.01) &= -1.48a \\ -1.3979 &= -1.48a \\ a &= .94. \end{aligned} \tag{14}$$

These equations are solved as shown by substituting the known value of s and the HEP estimate for Task 1 into one equation, and s and the HEP estimate for Task 6 into a second equation. The second equation is then subtracted from the first and the resulting equation (below the line in the example) is solved for a.

Substituting a back into the first equation allows this equation to be solved for b:

$$\log(.0004) = .94(-.64) + b \tag{15}$$

$$b = -2.7963$$

The formula:

$$\log HEP = .94s + (-2.7963) \tag{16}$$

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

Task No.	1	4	3	2	5	6
log HEP	-3.3979	-3.2569	-2.9373	-2.7399	-2.4391	-2.0
Probability	.0004	.0006	.001	.002	.004	.01

It is often desirable to use more than two anchor tasks to determine the values of a and b. This reduces the effect of any single task. Four anchor tasks should be used if possible. With more than two anchor tasks, regression can be used to find a and b. The logarithms of the HEP estimates for each of the anchor tasks should be regressed onto the scale values for those tasks. The value of a is then the slope of the regression, and b is the intercept of the regression.

When more than two anchor tasks are to be used, they should be selected so as to be spaced relatively evenly across the range of scale values. If possible, the anchor tasks should again include the tasks with the lowest and highest scale values. If four anchor tasks are used, the other two selected should have scale values approximately one-third and two-thirds of the range of scale values.

A regression can most easily be performed using any of many statistical packages. If, however, no such package is available, it can be done by hand because of the few data points involved (e.g., four pairs of data points with four anchor tasks).

To illustrate the regression computations, Tasks 1, 2, 3, and 6 from the example will be used. Although with only six tasks, four anchors would usually not be necessary, this example will show the steps in performing the regression to determine a and b. For this illustration the known HEPs for Tasks 2 and 3 are assumed to be .003 and .001, respectively.

The steps in the regression computation are given below and illustrated in Table A.9.

Table A.9 Illustration of regression to obtain parameters for transformation of scale values to HEP estimates

	x		y	x^2	xy
Task	Scale Value	HEP	log HEP	Scale^2 Value	Scale Value x log HEP
1	-.64	.0004	-3.40	.410	2.18
2	.06	.003	-2.52	.004	-0.15
3	-.15	.001	-3.00	.023	0.45
6	.84	.01	-2.00	.706	-1.68
Sums	0.11		-10.92	1.143	0.80
Means	.0275		- 2.73		

$$SS_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = .80 - \frac{(.11)(-10.92)}{4} = 1.10$$

$$SS_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 1.143 - \frac{(.11)^2}{4} = 1.14$$

$$a = \text{slope} = \frac{SS_{xy}}{SS_{xx}} = \frac{1.10}{1.14} = .96$$

$$b = \text{intercept} = \text{mean}_y - (\text{slope})\ (\text{mean}_x) = -2.73 - (.96)(.0275) = -2.76$$

1. List the tasks and their corresponding scale values. In typical regression notation, these are the x values.
2. List the HEPs for the anchor tasks.
3. Find the logarithms of the HEPs in Step 2. These are the y values.
4. Compute the square of the x values from Step 1 for each task.
5. Compute the product of each x value from Step 1 and its corresponding y value from Step 3.

6. Sum each of the columns x, y, x^2, and xy.

7. Find the means of x and y by dividing the sum of x and the sum of y (from Step 6), respectively, by n, the number of anchor tasks (in this case four).

8. Compute the intermediate value SS_{xy} by using the formula shown in Table A.9.

9. Compute the intermediate value SS_{xx} by using the formula shown in Table A.9.

10. Compute the value for a, which is the slope of the regression, by the indicated formula.

11. Compute the value for b, which is the intercept of the regression, by the indicated formula.

Using this example, log HEP = as + b can again be used to compute HEP estimates. In this case, the estimates are .0004, .003, .001, .0006, .004, and .01 for Tasks 1 through 6, respectively.

4.3.5 Computing Statistical Confidence Limits

This is an optional step in the estimation of HEPs using paired comparisons -- one that provides useful information, but is not necessary to derive HEP estimates. The procedures described here require a higher level of statistical knowledge than other parts of either the direct estimation or paired comparison procedures. These procedures are based on the standard theory of the distributions of functions of random variables. To understand these procedures will require some knowledge of mathematical statistics. Potential users of the paired comparison procedure should either have access to someone with this background or should not perform this specific part of the procedure.

In addition, the procedure described below is far too time consuming to perform without a computer. Thus, to compute these statistical confidence limits, a computer program is necessary.

The procedure for obtaining approximate 95% statistical confidence limits is an application of the "bootstrap" method (Efron, 1982). It involves first obtaining a table of variances related to a table of proportions such as Table A.7. This table of variances has entries $V[F^{-1}(r'_{jk})]$, where F^{-1} is the inverse of the cumulative normal distribution and $r'_{jk} = x/m$, for x = 1, ..., m-1, and r'_{jk} is defined as shown in Table A-10 for x = 0 and x = m, where x is the number of experts selecting task k over task j and m is the total number of experts. The

variance of $F^{-1}(r'_{jk})$ is calculated assuming x is binomially distributed with a selection probability equal to r_{jk}, the observed proportion of experts selecting k over j. The necessary calculations are the following:

$$V[F^{-1}(r'_{jk})] = \sum_{x=0}^{m} [F^{-1}(r'_{jk})]^2 P(x|r_{jk}) - \mu^2 \tag{17}$$

where

$$\mu = \sum_{x=0}^{m} F^{-1}(r'_{jk}) P(x|r_{jk}) \tag{18}$$

and

$$P(x|r_{jk}) = \binom{m}{x} (r_{jk})^x (1 - r_{jk})^{m-x}, \quad x = 0, 1, \ldots, m \tag{19}$$

An example of these calculations with $r_{jk} = .6$ and $m = 5$ is based on Table A.10.

Table A.10 Sample calculations for statistical confidence limits

x	r'_{jk}	$F^{-1}(r'_{jk})$	$P(x\|r_{jk} = .6)$
0	.08*	-1.41	$\binom{5}{0} .6^0 .4^5 = .0102$
1	.2	-.84	$\binom{5}{1} .6^1 .4^4 = .0768$
2	.4	-.25	$\binom{5}{2} .6^2 .4^3 = .2304$
3	.6	.25	$\binom{5}{3} .6^3 .4^2 = .3456$
4	.8	.84	$\binom{5}{4} .6^4 .4^1 = .2592$
5	.92**	1.41	$\binom{5}{5} .6^5 .4^0 = .0778$

* $1/2(m + 1)$

** $\frac{(2m + 1)}{2(m + 1)}$

Then, from this table

$$\mu = (-1.41)(.0102) + (-.84)(.0768) + (-.25)(.2304) + (.25)(.3456) + (.84)(.2592) + (1.41)(.0778) = .2773 \tag{20}$$

and

$$V[F^{-1}(r_{jk})] = (-1.41)^2(.0102) + (-.84)^2(.0768) + (-.25)^2(.2304) + (.25)^2(.3456) + (.84)^2(.2592) + (1.41)^2(.0778) - .0769 = .3712 \quad (21)$$

This procedure is used to compute all the entries in the table of variances. Table A.11 is an example of such a table of variances. From this table, the standard error, s.e., of the scale values can be derived as shown in the table according to the following steps.

1. Sum the variances in each column.

2. Divide the sums from Step 1 by n - 1.

3. Compute the standard error by taking the square root of the values in Step 2 divided by n - 1.

These standard errors provide an estimate of the variability in scale values.

Table A.11 Table of variances for paired comparison data

Task	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	-
Sum =	.885	.999	.964	.94	.969	.831
Variance = sum/(n - 1) =	.177	.200	.193	.188	.194	.166
s.e.= $\sqrt{\text{variance}/n-1}$ =	.188	.200	.196	.194	.197	.182

In addition to the variability in scale values, variability in HEP estimates can also be produced by variability in the estimates of a and b. When more than two anchor tasks are used, estimates of this variability can be obtained from a regression in which log HEPs for the anchor tasks are regressed on scale values expressed as deviations from the mean of the scale values. For example, with four anchor tasks used in the example in Section 4.3.4, the mean scale value is .0275, so the deviations are -.64 - .0275 = -.6675 for Task 1, -.15 - .0275 = -.1775 for Task 3, .06 - .0275 = .0325 for Task 2, and .84 - .0275 = .8125 for Task 6. Log HEP values for these tasks should be regressed onto these deviations.

Deviations from the mean are used so that the slope and intercept of the resulting regression are statistically independent. This fact is used in the following development. For the logarithmic relationship used to transform scale values into HEP estimates, the following equation is used to estimate the variance of the HEP estimates:

$$V(\log HEP_i) = V(as_i + b). \tag{22}$$

Because the estimates of a and b are independent,

$$\begin{aligned} V(\log HEP_i) &= V(b) + V(as_i) \\ &= V(b) + E(a^2 s_i^2) - E(as_i)^2 \\ &= V(b) + E(a^2)E(s_i^2) - E(a)^2\, E(s_i)^2 \\ &= V(b) + [V(a) + E(a)^2][V(s_i) + E(s_i)]^2 - E(a)^2\, E(s_i)^2 \\ &= V(b) + V(a)V(s_i) + V(a)E(s_i)^2 + E(a)^2 V(s_i) \\ &= V(b) + V(a)\,[V(s_i) + E(s_i)^2] + E(a)^2\, V(s_i). \end{aligned} \tag{23}$$

Then, using s_i^2 and a^2 to estimate $E(s_i)^2$ and $E(a)^2$, respectively, and $\hat{V}$ to indicate other variance estimates,

$$\hat{V}(\log HEP) = \hat{V}(b) + \hat{V}(a)[\hat{V}(s_i) + s_i^2] + a^2\, \hat{V}(s_i) \tag{24}$$

$\hat{V}(a)$ can be estimated by the variance in the estimate of a from the regression,

$$\hat{V}(a) = \frac{\sigma^2}{\sum_{i=1}^{n} s_i^2} \tag{25}$$

where n is the number of anchors, s_i is the scale value for anchor i, and σ^2 is the error variance from the regression. This latter quantity can be estimated as

$$\sigma^2 = \frac{SSE}{n-2} \tag{26}$$

where

$$SSE = SS_{yy} - (\text{slope})\ SS_{xy} \tag{27}$$

and

$$SS_{yy} = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n} \tag{28}$$

The notation here is the same as that used in Section 4.3.4 for the regression to determine a and b shown in Table A.9. V(b) can also be estimated from the regression as

$$\hat{V}(b) = \frac{\sigma^2}{n} \tag{29}$$

and $\hat{V}(s_i)$ is equal to the square of the standard error, $s.e._i^2$, from Table A.10.

These variance estimates can be substituted into equation (24) to produce

$$\hat{V}(\log HEP) = \frac{\sigma^2}{n} + \frac{\sigma^2}{\sum_{i=1}^{n} s_i^2}(s.e._i^2 + s_i^2) + a^2\ s.e._i^2 \tag{30}$$

Then statistical confidence limits on log HEP_i are plus and minus twice the square root of this variance estimate. These bounds on log HEP_i are converted to bounds on HEP_i by taking the antilogarithm of each bound.

5. APPLICATION OF PROCEDURES

The results of the evaluation conducted as a part of this project indicate that both direct numerical estimation and paired comparison scaling can be used to obtain HEP estimates for a wide range of operator tasks. There are some cautions, however, that apply to their use. In addition, certain situations may affect the selection of the technique to be used. In this section, these cautions are discussed and guidance is provided regarding when each of the techniques is appropriate.

5.1 Cautions

The primary caution regarding use of HEP estimates produced by either technique is satisfied if appropriate consistency checks have been made as described in Section 4. We reiterate here for emphasis that if the judgments are not sufficiently consistent, any resulting HEP estimates are highly questionable. Inconsistencies across experts may arise because of differences in interpretations of tasks, or in assumptions regarding performance shaping factors. If across-expert consistency is low, the experts should be queried with respect to their interpretation of tasks and their assumptions. If these queries identify possible reasons for inconsistencies, the tasks and/or assumptions should be clarified and additional judgments should be collected again. It may be that only a few tasks created the inconsistencies so that only judgments regarding those tasks need to be collected. If inconsistencies cannot be resolved, other sources of HEP estimates should be sought.

A second caution regards the estimates of uncertainty bounds collected with direct estimation. The discussion here should not be taken to indicate that uncertainty bounds collected by the procedures described in Section 4.2 are inappropriate. Rather, it simply indicates one possible problem that should be considered in future applications. Experience suggests that experts may estimate these bounds by simply applying some consistent factor (e.g., 10) to their estimated HEP with little regard for variations among tasks. While this may be appropriate, it is more likely to reflect a simple response strategy used to make these generally difficult judgments much easier. Some variations in the procedure used to collect uncertainty bound estimates were discussed in Section 4.2.5 that could reduce the possibility of such a simple response strategy. Additionally, a very conservative approach could be taken by computing statistical confidence limits for uncertainty bound estimates, and using the upper confidence limit on the upper bound as the upper bound estimate, and the lower confidence limit on the lower bound as the lower uncertainty bound estimate.

A final caution, specific to paired comparison scaling, concerns the number of anchor tasks used to determine the parameters in the transformation of scale values to HEP estimates. The experience in this study suggests that, if possible, more than two tasks should be used as anchors.

5.2 Selection of Technique

Selection of a technique may be based on the practical considerations associated with the data collection. These considerations include the number of experts available, the time available to the experts and the individuals collecting and analyzing the data, the need for uncertainty bounds, and the type of tasks for which HEP estimates are required.

Direct numerical estimation can generally be used with fewer experts than can paired comparison scaling. Seaver and Stillwell (1983, pp. 2-16) suggest that paired comparison scaling could be used with as few as eight experts, but to include a margin of safety, paired comparison scaling should probably have at least 10 to 12. Direct estimation, on the other hand could probably be used with as few as six experts. In using either technique, as many experts as practical should be included.

Direct numerical estimation also requires relatively less time to obtain the judgments. Based on data collection in this study, approximately 35 direct estimates including uncertainty bounds can be made using direct estimation in thirty minutes by each expert. Each expert can make approximately 100 paired comparisons in 30 minutes. Recalling that for n tasks, n direct estimation judgments are required and n(n-1)/2 paired comparisons are required, estimates of the time required for judgments can be made.

In addition to the time required for the judgments, the techniques all require time to prepare for data collection and to analyze the data collected. Paired comparisons require more preparation time because of the randomization required as described in Section 3.2.2 and the production of the response booklets.

Analysis of the data from the two techniques requires approximately the same amount of time if statistical confidence limits are not computed. If they are computed, more time and effort will be required for paired comparisons.

The need for uncertainty bounds may also affect the selection of a technique. In this study, uncertainty bounds were estimated only with direct estimation. To estimate uncertainty bounds using paired comparisons would require considerable additional time and effort, approximately twice the time if only upper bounds were estimated or three times the time if both upper and lower bounds were estimated. To collect uncertainty bounds using paired comparisons would require the identification of the best possible and most adverse PSFs for each task. (There could be considerable overlap in these PSFs across tasks, but complete overlap should not be assumed.) All the tasks with the best possible PSFs and all those with the most adverse PSFs would be grouped into separate sets for paired comparison judgments in each set. Thus, considerable additional expert time would be required as well as additional preparation and analysis time.

Both techniques appear to be applicable across a range of types of tasks. However, when using paired comparisons, care must be taken in the grouping of tasks into sets, with paired comparisons being made only on tasks within the same set. Only similar types of tasks should be included in a set. For example, system operation tasks should be grouped together, but not with tasks involving reading displays or operating instruments. Such groupings are necessary to ensure that the experts can make the needed comparisons. In addition, the expected HEP estimates for all tasks in the same set should be relatively comparable. For example, tasks with expected HEP estimates below .1 could be grouped, but should definitely not be grouped with tasks that have expected HEP estimates above .5. A single task with an HEP of .6 would probably be judged more likely by all experts than the next most likely task, e.g., HEP = .01. This would make the resulting estimate for the .5 HEP task relatively arbitrary. Generally, we recommend one set of tasks with expected HEP estimates below .1, another with expected estimates between .1 and .5, and a third with estimates above .5. To avoid sets with only a few tasks, the low HEP set could go as high as .25, and the high set as low as .25.

This section has provided a discussion of factors that may affect the choice of technique. Based on the results of this study, our recommendation is to choose the technique that is the most comfortable to use and meets any situational constraints that exist. Although in general, procedures for direct estimation appear to be more practical than procedures for paired comparison, results of this study suggest that the experts felt more comfortable making the paired comparison judgments (see Appendix B). Whichever technique is used, it should be used carefully, and should accurately follow the procedures described above, keeping in mind the appropriate cautions.

APPENDIX B

EVALUATION RESULTS

Appendix B presents the results of the test and evaluation that was conducted on the paired comparison and direct numerical estimation techniques. The appendix describes the issues that were investigated, the evaluation methods that were used, the data that were collected, the analyses that were performed, and the results of the evaluation.

1. ISSUES

As stated in Section 1 of the main report, 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 sets of issues were program issues and technical issues, as shown in Table B.1. Each of these issues was also 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 usable means of acquiring estimates?
- Will government and industry use psychological scaling techniques as part of the probabilistic risk assessment (PRA) process?

Table B.2 lists the issues that were considered during the project, identifies the characteristics of practicality, acceptability, and usefulness for each, and describes the method and type of analysis which was used to address each. One or more of the following methods were used 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 resulted from observation or experience. The quantitative type resulted in a numerical resolution of the issue; the comparative type was used to determine the similarities and differences between choices.

Table B.1 List of program and technical issues

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?

Technical Issues

T1. Based on measures of consistency and comparisons with other human reliability estimates, is there any difference in 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?

Table B.2 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 Table B.1 for example: P1 = program issue #1; T1 = technical issue #1

** Practicality, acceptability, usefulness

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

2. EVALUATION METHOD

This section provides a description of the methodology used for evaluation of psychological scaling techniques. Subjects, materials, and procedures are each described in detail.

2.1 Test Subjects

Experts selected for the project had to be familiar with the tasks to be judged, which involved nuclear power plant operations from a control room perspective. Thus, in-depth knowledge of plant systems, operations, and control room procedures was an essential criterion for selection of experts. Individuals who are currently, or were formerly, licensed by the NRC as boiling water reactor (BWR) control room operators or certified as BWR instructors meet these requirements. Other types of experts considered for this project were power plant operators, human factors engineers, psychologists, and human reliability analysts. They were not chosen in favor of certified instructors because the instructors had the most appropriate background for the tasks that were to be judged.

The amount of experience or education beyond that required for licensing or certification was a variable in the analysis rather than a criterion for the selection of experts. Later sections of this appendix describe the background data collected from each participant to assess the correlation between experience and characteristics of judgments.

Initially, current control room operators were considered as the subject matter experts. However, because the tasks to be judged included accidents and transient events, it was decided that certified instructors were better qualified to judge the likelihood of various operator actions under these conditions. Certified instructors have the opportunity to witness many different operators and their reactions to simulated accident scenarios. Therefore, they fulfill the criterion of being familiar with operator actions in a variety of circumstances. The 19 experts selected for participation in this project were certified as BWR instructors.

2.2 Materials Used

Three types of materials were used during data collection: task statements, response booklets, and data collection session instructions.

2.2.1 Task Statements

For the purpose of this study, BWR-related tasks were chosen. If tasks pertaining to another type of reactor were chosen, a different group of experts would have been needed.

The task selection criteria were chosen to ensure that the estimates generated from this project would be useful for PRAs. Four criteria were used to select the tasks to be judged in this project:

(1) Tasks must correspond to the data bank structure. The NRC and SNL have sponsored a program to develop a Human Reliability Data Bank for nuclear power industry PRA applications (Comer et al., 1983). In that project, a taxonomy was developed for classifying human reliability data. The tasks to be judged in this project were selected to correspond with that taxonomy. There were two separate sets of tasks. The first set consisted of tasks that corresponded to Level 1 of the Human Reliability Data Bank as described in Comer et al., (1983). The second set of tasks corresponded to Level 2 and Level 3 of the data bank.

Level 1 of the Human Reliability Data Bank structure combines power plant systems with human actions that represent job duties. For the purposes of this project, the Level 1 task set represented BWR systems and control room operator duties. Level 2 of the data bank structure combines equipment components with human actions defined as tasks. This project included those tasks associated with control room operators and equipment operators. Level 3 corresponds to controls, displays, instruments, and task elements.

(2) Tasks must be important to PRA practitioners. The objective of the project was to ascertain whether or not psychological scaling techniques could be used to generate human reliability estimates. Therefore, the second criterion for task selection was that the tasks chosen be at least recognized by PRA practitioners to be of value in a PRA.

(3) Because estimates obtained from this project must be compared with human reliability estimates presented in Swain and Guttmann (1983) and data obtained from simulator research (Beare et al., 1984), some of the tasks selected were taken from each source. The NRC has sponsored the following two projects that have resulted in human reliability data/estimates:

- Swain, A. D., and Guttmann, H. E. Handbook of human reliability analysis with emphasis on nuclear power plant applications (NUREG/CR-1278, 1983), hereafter called the Handbook.

- 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, 1984).

Because the estimates were to be compared, the tasks selected to be judged were compatible with tasks for which these other projects had collected data/estimates. Therefore, the second criterion for task selection was that at least some of the tasks chosen correspond to tasks in the Handbook and the simulator experiments.

(4) Tasks must represent a spread on the human error probability (HEP) continuum. Because it is difficult for subject matter experts to differentiate between items that are quantitatively very similar, and because it is difficult to analyze and draw conclusions from data that cluster around one area on a continuum, the desired result from the psychological scaling test was a range of HEP estimates. Therefore, a fourth criterion for task selection was that the tasks chosen represent a range across the HEP continuum.

Because the tasks in the second set were chosen to correspond to Handbook and simulator tasks, more assurance could be placed on the fact that they represented a range of HEPs. However, because there were no prior data to use to determine variance among Level 1 tasks, some judgment was used in selecting those tasks so that the resulting HEP estimates would vary.

The complete text of the Level 1 tasks as they were presented to the experts are contained in Attachment 1 to this appendix. Attachment 2 to this appendix lists the Level 2 and 3 tasks as they were presented to the experts. Because the tasks as defined by PRA analysts were not always written in a manner that could be easily understood by subject matter experts, some translation was necessary.

2.2.2 Response Booklets

Sample pages from the response booklets that the experts used to make the judgments are provided as Attachment 3 to this appendix. There were four separate parts to the response booklet, one for each of the data collection periods.

The first part of the booklet, for Period 1, contained assumptions for the tasks (either Level 1 or Levels 2 and 3) and examples of paired comparisons for instruction, and additional pages with the pairs resulting from the first task set (either Level 1 or Level 2 and 3 tasks, depending on the counterbalancing). The order of these pairs was randomized for each expert to minimize any effects of order of presentation; that is, each expert had a different order of pairs with all possible orders of pairs being equally likely. Randomization was also used to determine which of the two tasks within a pair appeared first on the page to minimize any bias toward tasks in the first or second position.

The second part of the booklet, for Period 2, contained the pairs from the task set not presented in the first part of the booklet and their assumptions.

The third part of the booklet, for the direct estimation period, consisted of assumptions, an example, the example completed, and another example for instructions; and additional pages, one task and scale per page, on which the expert provided HEP estimates with estimated uncertainty bounds. The experts were given a scale of probabilities on

which to respond ranging from p = 1 to p = .0000001. This range was at least two orders of magnitude lower than the lowest uncertainty bounds from the Handbook, so that the experts would not be constrained by the scale. Instructions indicated to the experts that they need not use the entire scale. The tasks from the task set in Period 1 came first, followed by those from the task set in Period 2. Within each task set, the order of the tasks was randomized for each expert.

The fourth part of the booklet consisted of two pages on which the expert provided additional information (e.g., experience) and ratings regarding the scaling techniques.

Participants' data remained anonymous.

2.2.3 Data Collection Session Instructions

Two sets of instructions were required and are contained in Attachment 4 to this appendix. The first set of instructions was for the session administrator. It provided guidance on administering the session, responding to questions, etc. These instructions were intended to allow a person who was not a psychological scaling expert to administer the sessions. A pretest (see Section 2.3.3) tested the effectiveness of these instructions, and revisions were made based on the results.

The second set of instructions was read by the session administrator to the experts during the data collection session. Instructions to be read at the beginning of each of the four data collection periods were included. They were also revised after the pretest.

2.3 Test Procedure

This section describes the data collection periods, the data collection team, and the pretest of data collection procedures.

2.3.1 Data Collection Periods

The data collection session was divided into four periods. In addition to instructions and training, each of the first two periods consisted of the paired comparisons for a set of tasks, i.e., Level 1 and Levels 2 and 3. The third period consisted of the direct estimates for both sets of tasks, and the fourth period was used to obtain additional information from the experts.

The paired comparison judgments were scheduled before the direct estimation judgments to reduce transfer effects between the two types of judgments. The ordering of judgment procedures took into consideration that the same experts would make both paired comparison judgments and direct numerical estimates of HEPs. There existed the potential that the quality, consistency, or some other aspect of the second set of judgments might be affected by the expert previously making the other set of

judgments about the same tasks. The expert might attempt to maintain consistency between paired comparison judgments and direct numerical estimates rather than correctly expressing the expert's subjective judgment of likelihood. This transfer effect was expected to be differential; that is, making direct estimates of HEPs first was likely to affect paired comparison judgments by focusing the expert on the full set of errors and the numerical estimates that the expert had attached to them. On the other hand, making paired comparison judgments first was less likely to affect the direct numerical estimates because the experts would not be likely to remember all their judgments and the implied ranking of those judgments. Furthermore, even if the experts were to remember the implied ranking, that would still not provide assistance in attaching a number to it. The design of the data collection plan attempted to minimize this potential effect by obtaining paired comparison judgments prior to direct numerical estimates.

During each of the data collection periods, described in detail in the sections that follow, the data collection team did not provide impromptu answers to questions about the task statements. The team might have provided inconsistent answers to different experts or in different sessions. Rather than risk this potential breach of standardization, the experts were asked to complete the items to the best of their knowledge. They were encouraged to give written comments in the booklet, adjacent to the item in question, or during the exit interview.

2.3.1.1 Period 1, Instructions and Paired Comparisons

Period 1 included a general description of the purpose of the session and the information that was to be obtained. The general introduction was followed by a review of the tasks and assumptions to be sure each expert understood them. Following this review, instructions with an example of paired comparison judgments were given (see Attachment 4 to this appendix). The experts then made the paired comparisons necessary for the first set of tasks. Half the experts responded to one set of tasks in Period 1 (e.g., Level 2 and 3 tasks), while the other half responded to the other set of tasks (e.g., Level 1 tasks). The counterbalancing of the two sets ensured that any differences in results were not dependent on the order in which task sets were considered.

This period was followed by a short break. The time required for Period 1 was about 45 minutes.

2.3.1.2 Period 2, Paired Comparisons

In Period 2, the experts made paired comparisons for the set of tasks they did not consider in Period 1. The actual paired comparisons and their assumptions were again preceded by a review of the tasks to be compared. The time for this period was also approximately 45 minutes. This period was also followed by a short break.

2.3.1.3 Period 3, Instructions and Direct Estimates

In Period 3, the experts made the direct estimates of HEPs and uncertainty bounds for tasks in both sets. Uncertainty bounds were defined in the instructions to the experts as shown in Section 5 of Attachment 4 to this appendix. The task sets were in the same order as they were for the paired comparison periods; that is, if Level 1 tasks were judged in Period 1, then direct estimates of Level 1 tasks were made before Level 2 and 3 tasks. Prior to making these judgments, the experts received instructions for making direct estimates and completed practice questions. Assumptions for Level 1 and Level 2 and 3 tasks were again reviewed. The experts did not have access to their paired comparison judgments when making direct estimates. The time for this period was 30 minutes.

2.3.1.4 Period 4, General Information Questions

In this brief period (about 15 minutes), the experts were asked to provide some personal background information and judgments regarding the procedures used. The sample pages from the response booklet in Attachment 3 show the questions that were asked during this period.

2.3.2 Data Collection Team

Two data collection team members were available during all sessions. One of these team members was the session administrator. By project design, the session administrator did not have any special knowledge or experience relating to the use of psychological scaling techniques. This person was familiar with the response booklet and had been briefed on the overall project. The role of the session administrator was to direct the sessions by reading instructions to the subject matter experts, ensuring that the subjects did not exchange information, and handling any questions that arose. Details on how questions were handled during the sessions were given in Section 2.3.1.

The second team member who was available during all sessions was a psychological scaling expert. The primary role of the psychological scaling expert was to ensure that the data were collected as designed. The psychological scaling expert was also able to answer any questions about how responses were to be made if the initial instructions were unclear.

2.3.3 Pretest

A pretest of the entire data collection session, including the paired comparison periods, direct numerical estimation period, and background information questions, was conducted with personnel trained in BWR operations. The purpose of the pretest, the procedures used, and a description of the participants are described below. The results of the pretest are then described.

2.3.3.1 Purpose of Pretest

The pretest was conducted to save time and resources during the actual data collection phase. The pretest was a dry run of everything planned to take place during the actual sessions, thereby giving the project team an opportunity to revise and refine any problem areas. Four primary areas were emphasized during the pretest:

- Response booklet and instructions
- Time requirements
- Session administrator qualifications
- Procedures

Issues of concern were the clarity, completeness, and accuracy of the instructions and the clarity of the task definitions contained in the response booklet. The pretest also aided in defining the time required for each response period and for the total session. As a result of the pretest, the time requirements for the actual data collection sessions were closely defined. In addition, the pretest helped define whether or not the session could be administered by someone other than a psychological scaling expert. It also provided information on any special qualifications that this person should have. Finally, the pretest provided information regarding the acceptance by the experts of the overall procedures used. Items such as frequency and length of breaks, ordering of response periods, and attitude and cooperation of session participants were assessed.

2.3.3.2 Pretest Procedure

The pretest participants were asked not only to participate in the data collection sessions but also to discuss their thoughts and reactions as well. For the pretest session, the instructions to the participants were written out, with examples, so that a psychological scaling expert would not be required. The session administrator read the instructions to the participants. As the participants listened to the instructions and responded to the questions in the booklets, any questions they had were recorded. The questions were reviewed after the pretest session to evaluate whether changes to the instructions or booklet were needed. The pretest was conducted in a group session, with each period timed. After the session was complete, the participants were asked to express their opinions and offer advice on ways to improve the session. Their comments were noted and, where possible, incorporated into the plans for data collection.

2.3.3.3 Experts Used

The pretest participants were two individuals familiar with BWR operations. Because the test subjects were to be operations and training personnel, the pretest participants had similar knowledge. The two

pretest participants were both BWR-certified instructors and one was also a former BWR-licensed operator.

2.3.3.4 Pretest Results

Changes incorporated for the actual data collection sessions as a result of the pretest are listed below:

- The word "test" was not used.
- The general information question dealing with years of experience was divided into number of years of operating experience versus number of years of training experience.
- The general information questions dealing with accuracy and ease were confusing and were therefore clarified.
- The list of assumptions for Level 1 and for Levels 2 and 3 was printed and bound in the appropriate sections of the response booklets.
- At the top of each page of paired comparisons, the following reminder was added, "Check the task that is the more likely to occur."
- Each person was provided with more than two pencils.
- The pages containing examples were numbered.
- Level 1 tasks and Level 2 and 3 tasks were renamed Level A and Level B, respectively, to avoid confusion with Periods 1 and 2.
- The wording of task descriptions was changed to make the tasks more understandable. However, the pretest experts did not list any new performance shaping factors (PSFs) that they assumed to have an impact on their judgments.

3. DATA COLLECTED

This section presents the data collected and describes the major intermediate steps necessary for data analyses. Using the procedures described in Section 2, both direct estimates and paired comparison judgments were obtained for each task set. In addition, uncertainty bounds were also estimated.

Tables B.3 and B.4 present the direct estimates for individual experts for Level 1 and Level 2 and 3 tasks, respectively. The uncertainty bound estimates are presented in Tables B.5 and B.6 for Level 1 and Level 2 and 3 tasks, respectively. All direct estimates were aggregated using the procedures described in Appendix A to produce single HEP and uncertainty bound estimates. For all estimates, outliers, which are defined as estimates more than two standard deviations (in logarithms) away from the mean, were not included in the calculation of the HEP estimates.

Experts' paired comparison judgments were first aggregated into a frequency matrix. All experts' data were included since each expert met the within-expert consistency requirements. The frequency matrices for Level 1 and Level 2 and 3 tasks are given in Tables B.7 and B.8, respectively. Each entry in these matrices represents the number of experts who judged the column task more likely than the row task. Then, using the procedures described in Appendix A, scale values were obtained. Table B.9 gives the scale values obtained for both Level 1 and Level 2 and 3 tasks.

These scale values were then transformed into probabilities using the procedures described in Appendix A. To make this transformation, at least two anchors were required. In this study both two and four anchors were used to examine the effects of the number of anchors. Anchors are tasks with HEP estimates from a source other than paired comparisons that are used to relate positions on the subjective scale, resulting from paired comparisons, to the probability scale. The anchor tasks from the Handbook and direct estimation were chosen so they would be spaced approximately equally across scale values for four anchors and at the ends of the scale for two anchors. Since there were only four tasks from the simulator experiments, they were used as anchor tasks.

A logarithmic transformation was used after analyses described in Section 4.7 indicated it to be appropriate. For HEP estimates based on two anchors, Tasks 6 and 9 were used for Level 1 tasks. For four anchors, these tasks and Tasks 1 and 12 were used. For Level 2 and 3 tasks, the two anchor tasks were 14 and 15 for direct estimates and Handbook anchors, and 1 and 3 for simulator anchors. With four anchors, Tasks 7 and 8 were added for direct estimates and Handbook anchors, and Tasks 2 and 4 for simulator anchors. The values of a and b in the transformation equation, log HEP = as + b, where s is the scale value and a and b are constants, are given in Table B.10. Using these parameters to transform scale values provided the HEP estimates needed for the analyses described in the following section.

Table B.3 Individual experts' direct HEP estimates for Level 1 tasks

Expert	Task							
	1	2	3	4	5	6	7	8
1	.0000006	.000005	.00001	.0000007	.000001	.33	.00003	.33
2	.005	.002	.0001	.001	.002	.02	.002	.001
3	.05	.05	.05	.01	.001	.10	.02	.05
4	.002	.01	.01	.00002	.0001	.05	.02	.05
5	.0000002	.0000002	.0000002	.00001	.0000002	.02	.000001	.0001
6	.02	.05	.005	.01	.0001	.05	.01	.05
7	.000005	.0005	.002	.000005	.0000005	.50	.50	.05
8	.10	.05	.01	.001	.001	.20	.10	.01
9	.0000001	.000005	.001	.000005	.02	.05	.000001	.01
10	.0002	.005	.005	.001	.05	.10	.0002	.50
11	.05	.01	.0000005	.02	.0000005	.20	.02	.50
12	.001	.005	.02	.001	.01	.10	.001	.10
13	.005	.0001	.05	.00001	.0001	.02	.005	.01
14	.001	.0005	.0001	.0002	.0005	.01	.005	.005
15	.002	.002	.005	.00005	.0001	.20	.05	.50
16	.0001	.005	.000005	.0001	.002	.0002	.0005	.01
17	.00001	.0001	.0001	.00001	.0001	.05	.01	.10
18	.005	.002	.002	.002	.02	.02	.02	.02
19	.002	.0002	.0002	.05	.007	.05	.0005	.02

Table B.3 Continued

Expert	Task						
	9	10	11	12	13	14	15
1	.0002	.50	.0002	.000005	.001	.00005	.50
2	.001	.05	.001	.02	.005	.002	.001
3	.0005	.50	.005	.02	.0005	.05	.05
4	.000001	.005	.002	.02	.0001	.0001	.01
5	.0000002	.00001	.0000002	.000001	.0000002	.0000002	.00001
6	.0002	.01	.0001	.002	.01	.0002	.10
7	.0000001	.01	.0001	.000001	.00005	.000005	.20
8	.10	.20	.0001	.001	.01	.01	.01
9	.001	.0001	.005	.000005	.000002	.000005	.001
10	.000005	.50	.0005	.01	.0001	.001	.10
11	.0000005	.20	.00001	.50	.20	.001	.50
12	.0001	.001	.001	.01	.05	.001	.10
13	.001	.02	.001	.01	.01	.01	.01
14	.0005	.0002	.0002	.001	.001	.001	.005
15	.20	.005	.00005	.002	.005	.002	.10
16	.0005	.002	.0000005	.0005	.0002	.0001	.01
17	.000001	.01	.00001	.0001	.01	.000001	.01
18	.005	.01	.002	.02	.02	.01	.10
19	.002	.001	.0001	.05	.0005	.005	.05

Table B.4 Individual experts' direct HEP estimates for Level 2/3 tasks

Expert	Task									
	1	2	3	4	5	6	7	8	9	10
1	.0002	.0002	.00002	.0000005	.003	.002	.000005	.0005	.001	.002
2	.001	.02	.02	.02	.002	.005	.005	.01	.01	.005
3	.001	.0005	.0002	.005	.01	.0002	.0002	.01	.01	.01
4	.02	.002	.002	.005	.0002	.0002	.005	.01	.01	.0002
5	.001	.0001	.000001	.0001	.01	.00001	.0001	.001	.01	.001
6	.01	.01	.002	.005	.02	.0001	.01	.005	.02	.002
7	.00005	.00005	.000001	.000001	.001	.000001	.000005	.00005	.00005	.000005
8	.05	.02	.01	.01	.05	.005	.01	.01	.10	.01
9	.00005	.00005	.0002	.000002	.05	.000002	.0005	.002	.005	.005
10	.05	.10	.05	.05	.10	.10	.005	.02	.10	.001
11	.001	.001	.10	.01	.20	.20	.01	.10	.05	.01
12	.005	.001	.0001	.0001	.01	.0001	.0001	.01	.01	.01
13	.002	.0002	.00002	.000005	.005	.00005	.0001	.002	.01	.002
14	.02	.005	.002	.0002	.02	.0005	.002	.002	.005	.002
15	.005	.05	.001	.002	.10	.001	.001	.01	.005	.0005
16	.005	.01	.005	.002	.01	.001	.002	.01	.005	.02
17	.01	.0001	.0002	.0005	.10	.000005	.001	.02	.01	.001
18	.05	.10	.005	.05	.005	.20	.005	.05	.10	.005
19	.04	.002	.0001	.0005	.02	.0001	.0001	.0005	.005	.02

Table B.4 Continued

Expert	Task									
	11	12	13	14	15	16	17	18	19	20
1	.07	.007	.00003	.0000001	.03	.0000002	.002	.0008	.00008	.00002
2	.0001	.02	.001	.000001	.01	.02	.01	.01	.02	.005
3	.002	.001	.001	.000001	.10	.01	.00005	.002	.01	.0001
4	.00005	.02	.02	.00005	.10	.0002	.01	.02	.005	.0002
5	.01	.10	.10	.0000002	.10	.001	.000001	.01	.001	.01
6	.002	.05	.01	.00001	.01	.002	.0005	.005	.005	.02
7	.000005	.05	.00005	.0000002	.10	.0000005	.0002	.000005	.0000005	.002
8	.05	.02	.10	.00001	.20	.001	.001	.005	.05	.05
9	.005	.05	.01	.0000001	.10	.0000002	.005	.005	.00001	.005
10	.05	.10	.05	.0000001	.10	.000005	.01	.01	.10	.00005
11	.001	.05	.05	.0000002	.10	.001	.01	.01	.50	.005
12	.01	.0001	.05	.0001	.01	.00001	.001	.01	.01	.02
13	.01	.005	.005	.000002	.10	.00001	.0002	.01	.001	.002
14	.002	.005	.01	.00005	.005	.0005	.0002	.005	.002	.005
15	.0005	.02	.005	.000005	.10	.000001	.0005	.05	.0005	.005
16	.002	.01	.01	.000005	.01	.0001	.001	.02	.005	.001
17	.000005	.02	.0001	.0000005	.01	.0000001	.00001	.05	.00001	.01
18	.01	.02	.02	.005	.05	.005	.10	.20	.10	.02
19	.0005	.001	.001	.000002	.02	.000002	.002	.005	.002	.01

Table B.5 Individual experts' uncertainty bound estimates for Level 1 tasks

Expert	Task									
	1		2		3		4		5	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
1	.0000005	.000001	.000001	.00001	.000001	.01	.0000005	.000001	.000001	.000001
2	.0001	.02	.00002	.02	.000005	.005	.00001	.05	.00001	.02
3	.02	.10	.02	.10	.02	.10	.005	.02	.0005	.002
4	.0002	.01	.001	.05	.001	.05	.000001	.0005	.00001	.001
5	.0000001	.0000005	.0000001	.0000005	.0000001	.0000005	.0000001	.10	.0000001	.0000005
6	.001	.05	.005	.10	.0005	.01	.001	.02	.000005	.0005
7	.0000002	.0001	.0001	.002	.001	.01	.000001	.0001	.0000001	.000005
8	.05	.50	.01	.10	.001	.10	.0001	.01	.0001	.01
9	.0000001	.0000002	.000001	.00001	.0001	.01	.000001	.00001	.01	.05
10	.000005	.02	.001	.02	.000005	.02	.000005	.01	.0005	.10
11	.02	.10	.005	.10	.0000002	.000001	.01	.05	.0000002	.000001
12	.0001	.005	.001	.02	.005	.10	.0001	.01	.002	.05
13	.001	.05	.00002	.0005	.01	.10	.000002	.0001	.00002	.001
14	.0005	.005	.0002	.001	.00005	.0002	.0001	.0005	.0002	.001
15	.0001	.01	.0002	.01	.001	.01	.00001	.0002	.00002	.0005
16	.00001	.005	.001	.01	.000001	.00001	.00001	.0005	.00001	.01
17	.0000002	.001	.000001	.01	.000001	.01	.000001	.0001	.000001	.01
18	.001	.05	.0005	.01	.0005	.01	0005	.01	.005	.10
19	.0002	.005	.00005	.001	.00005	.001	.11	.10	.002	.05

Table B.5 Continued

Expert	Task									
	6		7		8		9		10	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
1	.20	.50	.000005	.0001	.10	.50	.00001	.0005	.20	1.0
2	.0005	.10	.00002	.02	.00005	.05	.00005	.05	.0005	.10
3	.05	.20	.01	.05	.02	.10	.0002	.001	.20	1.0
4	.005	.20	.005	.10	.01	.10	.0000001	.00001	.001	.01
5	.000001	.05	.0000001	.0001	.0000001	.10	.0000001	.0000005	.0000001	.01
6	.01	.20	.0005	.05	.005	.10	.00001	.001	.001	.05
7	.20	1.0	.20	1.0	.005	.10	.0000007	.0000002	.0001	.10
8	.10	.50	.02	.50	.001	.10	.05	.50	.10	.50
9	.01	.10	.0000005	.000002	.002	.05	.0000001	1.0	.000001	.01
10	.00001	1.0	.0000001	.002	.002	1.0	.0000001	.0005	.02	1.0
11	.10	.50	.01	.05	.20	1.0	.0000002	.000001	.10	.50
12	.01	.50	.0001	.005	.02	.50	.00001	.001	.0001	.01
13	.005	.10	.001	.02	.002	.05	.0002	.01	.002	.10
14	.005	.02	.002	.01	.002	.01	.0002	.001	.0001	.0005
15	.05	.50	.01	.10	.10	.70	.07	.30	.0005	.01
16	.00001	.02	.00001	.005	.0001	.05	.00005	.01	.0002	.02
17	.01	.10	.0001	.10	.01	.50	.0000001	.0001	.0001	.10
18	.005	.05	.005	.10	.005	.10	.001	.02	.001	.10
19	.005	.10	.0001	.001	.005	.05	.0002	.005	.0001	.005

Table B.5 Continued

Expert	Task									
	11		12		13		14		15	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
1	.0001	.0005	.000001	.00002	.0002	.002	.00001	.0001	.33	1.0
2	.00005	.02	.00001	.05	.00005	.10	.0001	.02	.00002	.05
3	.002	.01	.01	.05	.0002	.001	.02	.10	.02	.10
4	.0005	.005	.005	.05	.00001	.001	.00001	.0005	.001	.10
5	.0000001	.0000005	.0000001	.10	.0000001	.0000005	.0000001	.0000005	.000001	.10
6	.000005	.0005	.0001	.01	.0005	.05	.00001	.001	.01	.50
7	.00001	.001	.0000001	.00001	.00001	.001	.0000005	.00002	.05	.50
8	.0000	.001	.0001	.01	.005	.02	.001	.05	.001	.05
9	.001	.01	.000002	.00001	.000001	.000005	.000002	.00001	.0002	.005
10	.000005	.01	.000001	.10	.0000001	.01	.000001	.01	.005	1.0
11	.0000	.00002	.20	1.0	.10	.50	.0005	.002	.20	1.0
12	.00005	.005	.001	.10	.01	.10	.0001	.005	.05	.50
13	.0002	.01	.002	.10	.002	.05	.001	.10	.002	.05
14	.0001	.0005	.0005	.005	.0005	.002	.0005	.002	.002	.01
15	.000005	.0005	.0005	.01	.0005	.02	.0002	.01	.02	.20
16	.0000001	.000001	.00001	.01	.00001	.005	.00001	.001	.0005	.05
17	.000001	.001	.000001	.0005	.0001	.10	.0000002	.0001	.001	.10
18	.0005	.01	.005	.10	.005	.10	.002	.05	.02	.50
19	.00002	.0005	.01	.10	.0001	.005	.0005	.02	.01	.10

Table B.6 Individual experts' uncertainty bound estimates for Level 2 and 3 tasks

Expert	Task									
	1		2		3		4		5	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
1	.0001	.0005	.0001	.0005	.00001	.00005	.0000002	.000001	.002	.005
2	.00002	.05	.005	.05	.005	.10	.005	.10	.0001	.01
3	.0002	.002	.0002	.01	.0001	.0005	.002	.05	.002	.10
4	.005	.10	.0002	.02	.0001	.01	.001	.02	.00002	.001
5	.000001	.10	.00001	.001	.0000001	.00001	.00001	.001	.0000001	1.0
6	.001	.05	.001	.05	.0005	.005	.0005	.01	.002	.05
7	.00001	.001	.00001	.001	.0000002	.00001	.0000001	.00001	.0002	.01
8	.01	.10	.001	.10	.001	.10	.001	.05	.01	.10
9	.00002	.0001	.00001	.0001	.00005	.0005	.000001	.000005	.0000001	1.0
10	.005	.10	.005	.20	.001	.20	.005	.10	.01	1.0
11	.0005	.01	.0005	.01	.05	.20	.005	.02	.10	.50
12	.0005	.02	.0001	.01	.00001	.01	.00001	.01	.001	.10
13	.0005	.01	.00005	.001	.000005	.0001	.000002	.00002	.002	.02
14	.01	.05	.002	.01	.0005	.005	.00002	.002	.01	.05
15	.0005	.02	.01	.50	.0001	.01	.0001	.01	.005	.50
16	.0005	.05	.001	.05	.0005	.02	.0002	.01	.001	.05
17	.002	.02	.00001	.001	.00002	.001	.00005	.005	.05	.50
18	.01	.20	.02	.20	.001	.02	.01	.20	.002	.02
19	.001	.10	.00001	.005	.000001	.001	.00001	.001	.001	.10

Table B.6 Continued

Expert	Task									
	6		7		8		9		10	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
1	.001	.005	.000002	.00001	.0002	.001	.0005	.002	.001	.005
2	.0002	.02	.0005	.20	.005	.20	.0001	.10	.00002	.05
3	.00002	.001	.0001	.0005	.0001	.02	.001	.02	.001	.02
4	.00005	.001	.001	.02	.002	.05	.002	.05	.00002	.001
5	.0000001	.0001	.000001	.001	.00001	.50	.00001	.10	.0000001	.10
6	.000005	.0002	.0005	.05	.0005	.02	.001	.05	.0002	.005
7	.0000002	.00001	.000001	.00001	.00001	.001	.00001	.001	.000001	.0001
8	.0005	.05	.005	.05	.001	.05	.05	.20	.001	.10
9	.000001	.000005	.00001	.01	.001	.01	.0005	.05	.0000001	1.0
10	.01	1.0	.001	.05	.001	.10	.005	1.0	.00001	.02
11	.10	.50	.005	.05	.05	.20	.02	.10	.005	.02
12	.00001	.01	.00001	.01	.001	.10	.001	.10	.001	.10
13	.00002	.0002	.00002	.001	.0005	.01	.002	.05	.0005	.01
14	.0002	.002	.0005	.01	.001	.005	.001	.02	.001	.005
15	.0001	.005	.0001	.01	.001	.05	.0005	.02	.0001	.002
16	.00005	.005	.0001	.01	.0005	.05	.0005	.02	.0005	.10
17	.000001	.00001	.0002	.005	.005	.05	.001	.10	.00001	.01
18	.05	.50	.002	.02	.01	.20	.02	.20	.001	.02
19	.00001	.001	.000001	.001	.00001	.001	.0001	.01	.002	.10

Table B.6 Continued

Expert	Task									
	11		12		13		14		15	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
1	.05	.10	.005	.01	.00002	.00005	.0000001	.0000002	.02	.05
2	.00005	.005	.01	.10	.00001	.01	.0000001	.001	.0005	.10
3	.001	.05	.00005	.01	.0005	.01	.0000001	.000002	.05	.50
4	.000005	.0005	.005	.20	.005	.10	.00001	.0002	.01	.20
5	.0001	.50	.001	1.0	.0001	.50	.0000001	.0000005	.0000001	.50
6	.0001	.005	.005	.10	.001	.02	.0000005	.00005	.001	.02
7	.000001	.00001	.001	.10	.00001	.001	.0000001	.0000005	.01	.50
8	.01	.10	.005	.05	.01	.20	.0000005	.00002	.05	.50
9	.0000001	1.0	.0001	.10	.00001	.10	.0000001	.0000002	.0000001	1.0
10	.001	.10	.005	.50	.002	.10	.0000001	.000005	.001	.50
11	.0005	.01	.02	.10	.02	.10	.0000001	.0000005	.05	.20
12	.001	.10	.00001	.001	.01	.20	.00001	.001	.001	.10
13	.002	.05	.001	.02	.0005	.01	.000001	.00002	.02	.20
14	.0005	.01	.002	.01	.002	.05	.00002	.0001	.002	.01
15	.00002	.005	.002	.10	.001	.01	.0000005	.00005	.001	.20
16	.0001	.05	.001	.05	.001	.10	.0000001	.00001	.001	.10
17	.0000002	.0001	.005	.10	.00005	.001	.0000001	.000001	.00001	.50
18	.001	.10	.005	.20	.005	.05	.0002	.02	.005	.20
19	.00001	.01	.0001	.005	.0001	.005	.000001	.00001	.002	.05

Table B.6 Continued

Expert	Task									
	16		17		18		19		20	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
1	.0000001	.0000005	.001	.005	.0005	.002	.00005	.0002	.00001	.00005
2	.0002	.10	.001	.05	.0001	.05	.0005	.10	.001	.10
3	.005	.02	.00002	.001	.001	.01	.002	.05	.00001	.001
4	.00005	.001	.001	.10	.005	.10	.001	.02	.00005	.001
5	.0001	.01	.0000001	.01	.0001	.10	.0001	.01	.001	.10
6	.0001	.005	.00002	.002	.0001	.02	.0005	.02	.0005	.05
7	.0000001	.00001	.00001	.001	.000001	.00005	.0000002	.000005	.0002	.01
8	.0001	.005	.0001	.005	.001	.01	.01	.10	.01	.10
9	.0000001	.00001	.0002	.05	.0002	.05	.0000001	1.0	.0001	.10
10	.0000001	.001	.0002	.05	.0005	.05	.005	.50	.0000001	.002
11	.0005	.002	.005	.02	.005	.02	.20	1.0	.002	.01
12	.000001	.0001	.0001	.01	.001	.10	.001	.05	.002	.10
13	.000005	.00005	.00002	.002	.001	.05	.0001	.005	.0005	.005
14	.0002	.001	.0001	.0005	.002	.01	.001	.005	.002	.02
15	.0000001	.000005	.0001	.002	.005	.20	.0001	.005	.0002	.02
16	.00001	.0005	.00005	.005	.002	.05	.0002	.02	.0001	.01
17	.0000001	.0000002	.000001	.0001	.01	.10	.01	.10	.005	.02
18	.002	.02	.02	.20	.02	.50	.02	.20	.005	.10
19	.000001	.00001	.0002	.01	.0002	.05	.0005	.01	.0005	.05

Table B.7 Frequency matrix for paired comparison judgments on Level 1 tasks

Task	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	-	10	7	10	8	19	17	19	4	14	8	16	13	10	18
2	9	-	5	6	5	19	13	18	6	13	8	11	10	6	19
3	12	14	-	9	9	19	13	15	7	13	10	12	12	10	17
4	9	13	10	-	11	18	16	19	6	16	11	16	14	8	19
5	11	14	10	8	-	19	16	17	8	13	12	15	12	11	18
6	0	0	0	1	0	-	2	11	0	0	1	2	2	1	9
7	2	6	6	3	3	17	-	17	3	8	3	6	7	6	14
8	0	1	4	0	2	8	2	-	1	1	0	3	2	1	6
9	15	13	12	13	11	19	16	18	-	13	10	17	13	15	18
10	5	6	6	3	6	19	11	18	6	-	4	10	10	7	17
11	11	11	9	8	7	18	16	19	9	15	-	17	14	12	19
12	3	8	7	3	4	17	13	16	2	9	2	-	6	5	18
13	6	9	7	5	7	17	12	17	6	9	5	13	-	8	15
14	9	13	9	11	8	18	13	18	4	12	7	14	11	-	18
15	1	0	2	0	1	10	5	13	1	2	0	1	4	1	-

Table B.8 Frequency matrix for paired comparison judgments on Level 2 and 3 tasks

Task	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	-	5	0	2	14	0	3	12	13	10	6	12	9	1	14	3	5	13	9	9
2	14	-	0	2	14	0	2	13	14	13	7	14	15	2	17	3	6	17	6	13
3	19	19	-	10	17	5	13	19	18	19	10	18	18	3	16	4	10	19	13	16
4	17	17	9	-	18	7	10	17	16	17	12	13	15	2	16	5	11	17	16	16
5	5	5	2	1	-	0	3	5	8	3	2	8	5	2	13	3	6	10	3	5
6	19	19	14	12	19	-	16	19	16	18	15	17	18	5	17	7	12	19	16	16
7	16	17	6	9	16	3	-	19	19	15	11	16	14	1	19	3	6	19	13	15
8	7	6	0	2	14	0	0	-	14	10	4	12	9	1	14	2	3	13	7	8
9	6	5	1	3	11	3	0	5	-	6	2	8	6	1	13	1	1	12	6	5
10	9	6	0	2	16	1	4	9	13	-	5	10	11	1	14	3	5	12	9	10
11	13	12	9	7	17	4	8	15	17	14	-	17	15	2	18	4	8	17	14	12
12	7	5	1	6	11	2	3	7	11	9	2	-	5	1	14	2	5	10	6	5
13	10	4	1	4	14	1	5	10	13	8	4	14	-	1	15	3	4	13	9	11
14	18	17	16	17	17	14	18	18	18	18	17	18	18	-	19	18	17	18	17	17
15	5	2	3	3	6	2	0	5	6	5	1	5	4	0	-	0	3	7	3	4
16	16	16	15	14	16	12	16	17	18	16	15	17	16	1	19	-	15	18	15	15
17	14	13	9	8	13	7	13	16	18	14	11	14	15	2	16	4	-	18	12	13
18	6	2	0	2	9	0	0	6	7	7	2	9	6	1	12	1	1	-	4	4
19	10	13	6	3	16	3	6	12	13	10	5	13	10	2	16	4	7	15	-	8
20	10	6	3	3	14	3	4	11	14	9	7	14	8	2	15	4	6	15	11	-

Table B.9 Scale values from paired comparison judgments

Task	Level 1 Tasks	Level 2/3 Tasks
1	-0.540	0.409
2	-0.271	0.142
3	-0.449	-0.875
4	-0.670	-0.605
5	-0.606	0.782
6	1.446	-1.089
7	0.327	-0.546
8	1.328	0.550
9	-0.827	0.774
10	-0.033	0.391
11	-0.664	-0.368
12	0.191	0.586
13	-0.064	0.348
14	-0.408	-1.417
15	1.241	1.024
16	-	-0.300
17	-	-0.408
18	-	0.948
19	-	0.074
20	-	0.182

Table B.10 Parameters for transforming scale values into HEP estimates

	a	b
Level 1 Tasks		
Two anchors	1.21105	-2.90120
Four anchors	1.12612	-2.86060
Level 2/3 Tasks		
Two direct estimate anchors	1.79676	-3.22355
Four direct estimate anchors	1.64314	-2.90808
Two Handbook anchors	1.39203	-2.02750
Four Handbook anchors	1.14486	-2.61444
Two simulator anchors	0.88982	-2.50540
Four simulator anchors	0.33608	-2.63935

4. RESULTS OF ANALYSES

Because of the wide range of issues being addressed in this implementation and evaluation study, a number of different data analyses were required. In this section, the results of these analyses are described in detail. The presentation of these results is organized around the issues themselves. We recognize that in many instances the analyses do not provide a definitive resolution of the issue. In some cases, this is because of some ambiguity in our results. In other cases, it was not possible for this single study to thoroughly address the issue. In Section 5 of this appendix, we discuss the extent to which the issues have been resolved by this study.

While a large number of analyses have been completed, including those that most directly bear on the issues, numerous additional analyses could be performed that would provide additional support for these findings. Complete data are provided here in the hope that others will take advantage of this very rich set of data to conduct other analyses.

Tables B.11 and B.12 show the HEP estimates obtained from the experts' judgments in this study for Level 1 and Level 2 and 3 tasks, respectively. This same information is also presented graphically in Figures B.1 through B.4 to show how well the different estimates agree. Also included in both the tables and the figures are HEP estimates that are currently available for some of these tasks from the Handbook and from simulator studies. The task numbers refer to the numbers of tasks described in Attachments 1 and 2 to this appendix.

These estimates have been derived from the experts' judgments using the procedures described in Appendix A. All estimates have been rounded to one significant figure, except those greater than .1, which are rounded to two significant figures. Inclusion of more significant figures would give a false sense of precision and would make comparisons among the estimates less obvious. However, in analyses described subsequently in this appendix the estimates were not rounded.

In addition to the HEP estimates, estimates were also obtained for upper and lower 90-percent uncertainty bounds. These uncertainty bound estimates are given in Tables B.13 and B.14, along with corresponding uncertainty bound estimates from the Handbook. These tables also show the 95-percent statistical confidence limits for these uncertainty bound estimates computed using the procedure described in Appendix A for computing statistical confidence limits on direct HEP estimates.

4.1 Discussion of Program Issues

The six Program Issues described in Section 1 of this appendix are addressed in the section.

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

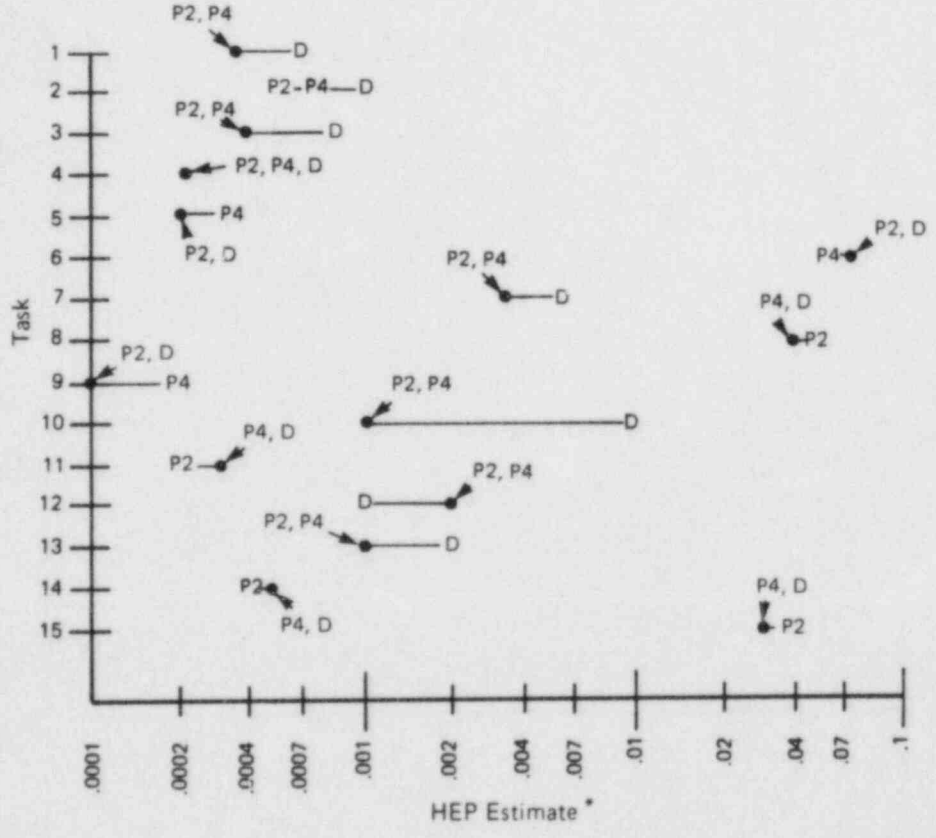
One of the major issues addressed in this study concerned whether procedures based on psychological scaling using expert judgment could produce consistent judgments and result in consistent HEP and uncertainty bound estimates. Such consistency is, of course, a prerequisite for any further use of these procedures.

Table B.11 Comparison of HEP estimates for Level 1 tasks

Task	Direct Numerical Estimation	Paired Comparisons	
		2 Anchors	4 Anchors
1	0.0007	0.0003	0.0003
2	0.001	0.0006	0.0007
3	0.0008	0.0004	0.0004
4	0.0002	0.0002	0.0002
5	0.0002	0.0002	0.0003
6	0.07	0.07	0.06
7	0.006	0.003	0.003
8	0.04	0.05	0.04
9	0.0001	0.0001	0.0002
10	0.01	0.001	0.001
11	0.0003	0.0002	0.0003
12	0.001	0.002	0.002
13	0.002	0.001	0.001
14	0.0005	0.0004	0.0005
15	0.03	0.04	0.03

Table B.12 Comparison of HEP estimates for Level 2 and 3 tasks

	HEP Estimates								
		Paired Comparisons							
		Direct Numerical Estimation		Handbook		Simulator			
Task	Direct Numerical Estimation	2 Anchors	4 Anchors	2 Anchors	4 Anchors	2 Anchors	4 Anchors	Handbook	Simulator
1	0.004	0.003	0.006	0.03	0.007	0.007	0.003	0.003	0.007
2	0.002	0.001	0.002	0.01	0.004	0.004	0.003	0.001	0.0006
3	0.0005	0.00002	0.00005	0.0006	0.0002	0.0005	0.001	0.0005	0.0005
4	0.0005	0.00005	0.0001	0.001	0.0005	0.0009	0.001	0.0005	0.006
5	0.02	0.02	0.02	0.12	0.02	0.02	0.004	0.01	
6	0.0004	0.000007	0.00002	0.0003	0.0001	0.0003	0.001	0.001	
7	0.001	0.00006	0.0002	0.002	0.0006	0.001	0.002	0.0005	
8	0.006	0.006	0.01	0.05	0.01	0.01	0.004	0.001	
9	0.01	0.01	0.02	0.11	0.02	0.02	0.004	0.003	
10	0.003	0.003	0.005	0.03	0.007	0.007	0.003	0.005	
11	0.003	0.0001	0.0003	0.003	0.0009	0.001	0.002	0.25	
12	0.02	0.007	0.01	0.06	0.01	0.01	0.004	0.10	
13	0.007	0.003	0.005	0.03	0.006	0.006	0.003	0.01	
14	0.000002	0.000002	0.000006	0.0001	0.00006	0.0002	0.0008	0.0001	
15	0.04	0.04	0.06	0.25	0.04	0.03	0.005	0.25	
16	0.00005	0.00001	0.00004	0.0005	0.0002	0.0005	0.001	0.001	
17	0.001	0.0001	0.0003	0.003	0.0008	0.001	0.002	0.002	
18	0.01	0.03	0.04	0.20	0.03	0.02	0.005	0.006	
19	0.003	0.0008	0.002	0.01	0.003	0.004	0.002	0.05	
20	0.003	0.001	0.002	0.02	0.004	0.005	0.003	0.001	

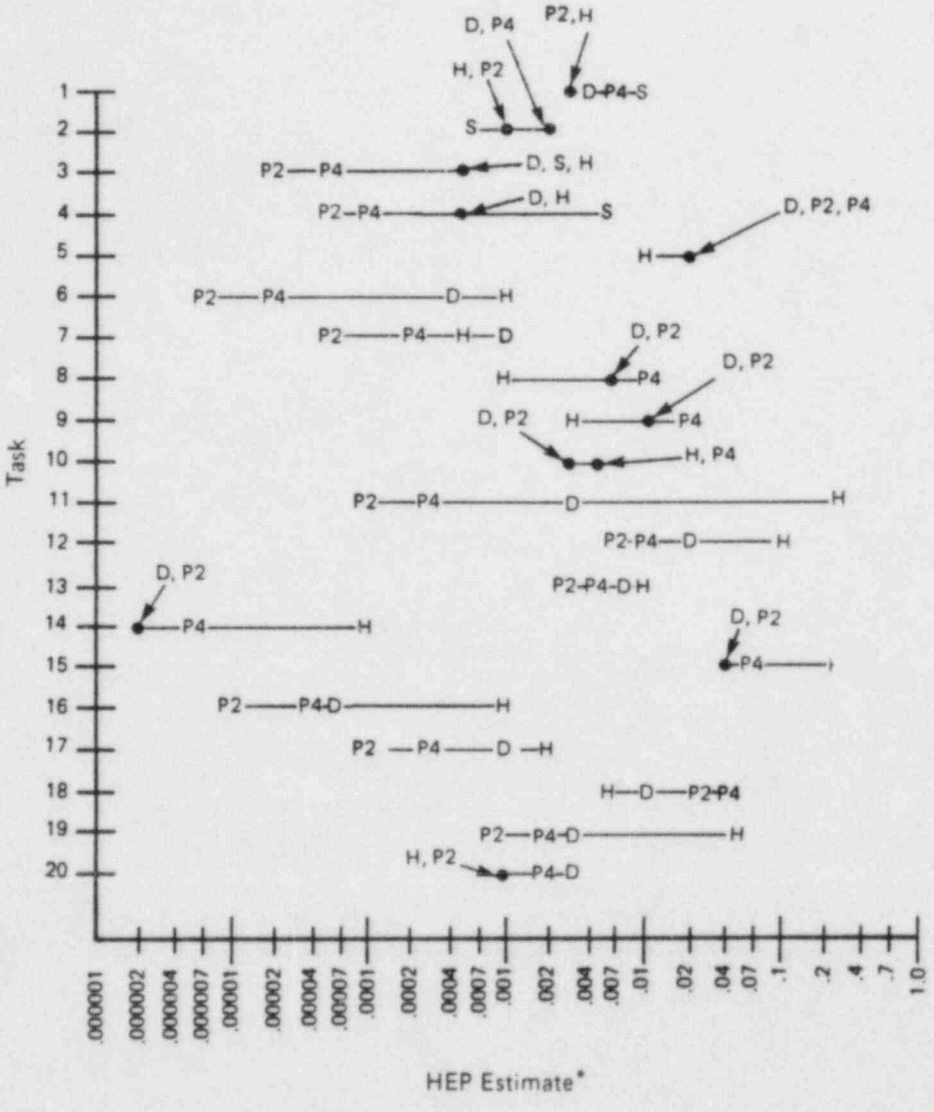


LEGEND

P2 = paired comparison estimates with two anchors
P4 = paired comparison estimates with four anchors
D = direct estimates

* Precise HEP estimates are given in Table B.11. This graph is for illustrative purposes only.

Figure B.1 Direct numerical estimates and paired comparison estimates for Level 1 tasks.

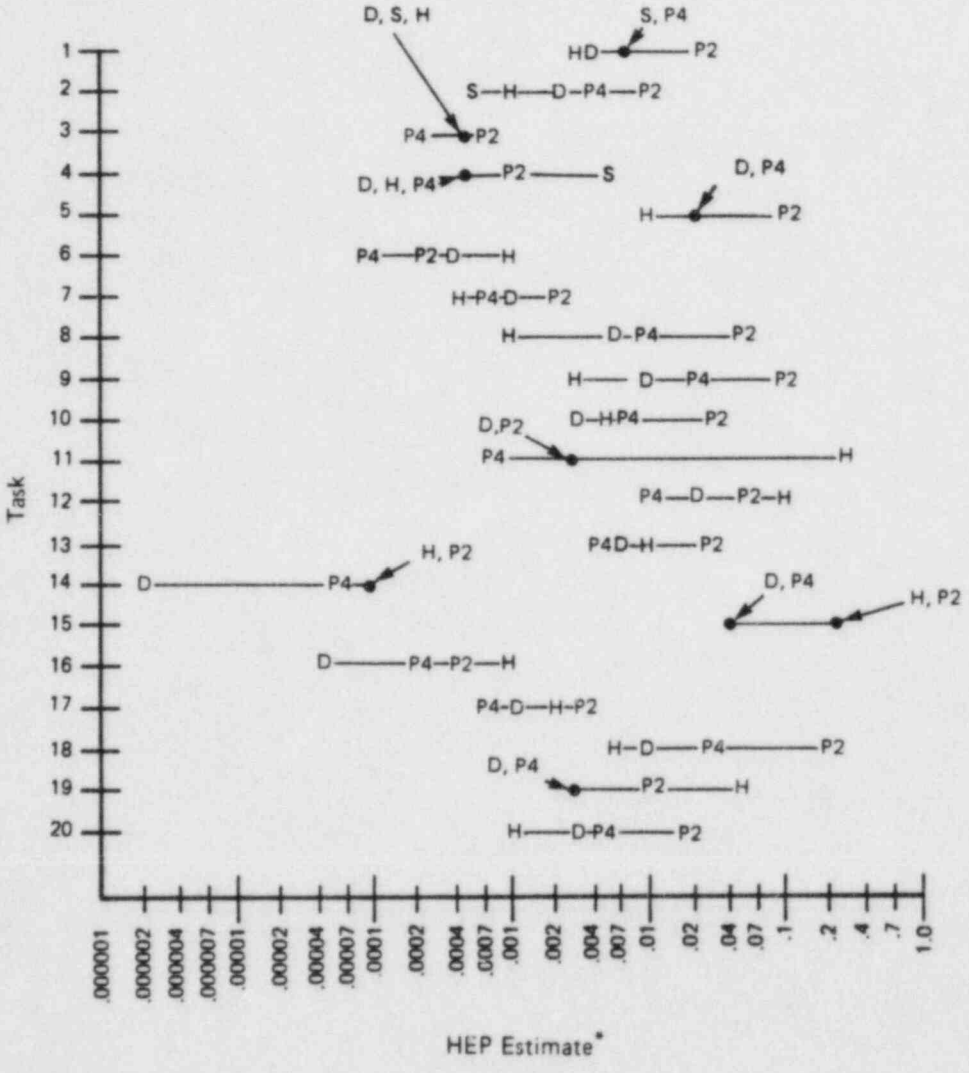


LEGEND

P2 = paired comparison estimates with two anchors
P4 = paired comparison estimates with four anchors
D = direct estimates
H = Handbook estimates
S = Similator estimates

* Precise HEP estimates are given in Table B.12. This graph is for illustrative purposes only.

Figure B.2 HEP estimates for Level 2 and 3 tasks with direct estimate anchors for paired comparison estimates.

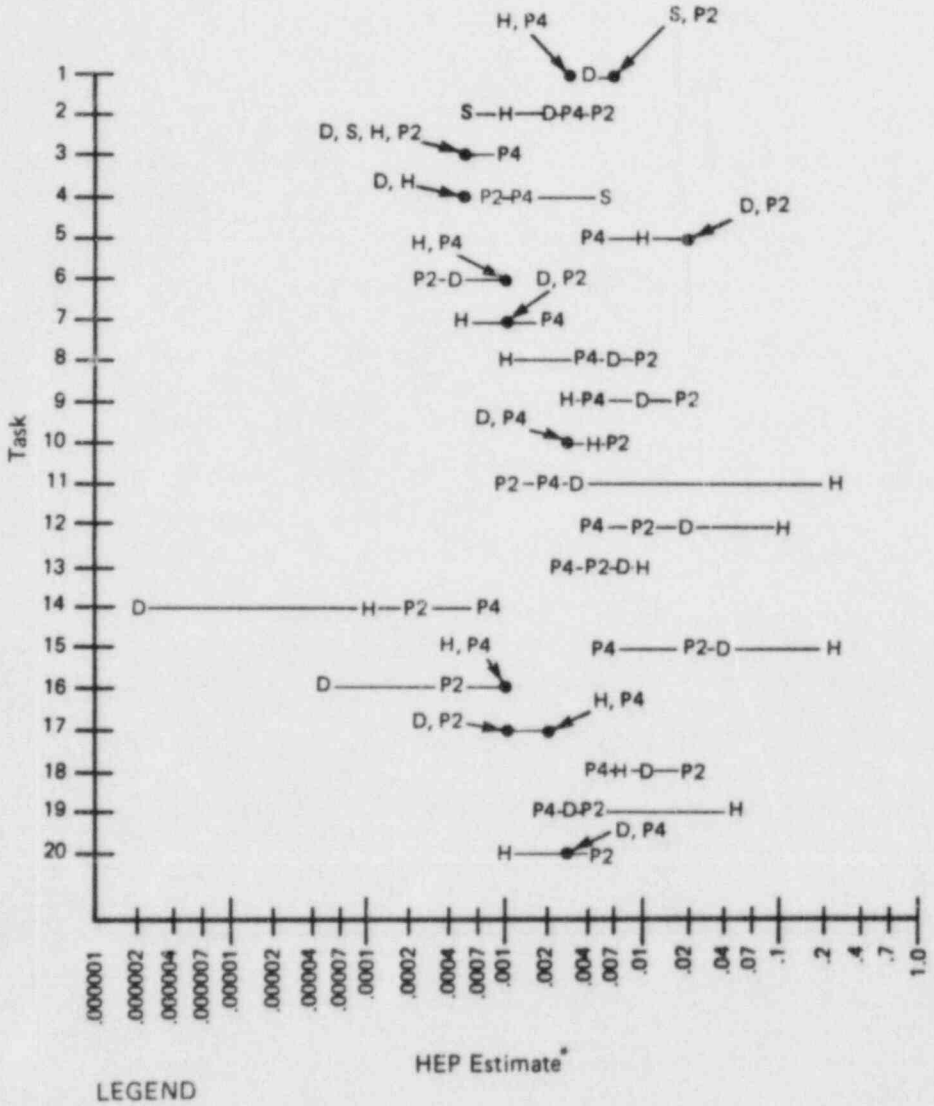


LEGEND

P2	=	paired comparison estimates with two anchors
P4	=	paired comparison estimates with four anchors
D	=	direct estimates
H	=	Handbook estimates
S	=	Simulator estimates

* Precise HEP estimates are given in Table B.12. This graph is for illustrative purposes only.

Figure B.3 HEP estimates for Level 2 and 3 tasks with Handbook anchors for paired comparison estimates.



LEGEND

P2 = paired comparison estimates with two anchors
P4 = paired comparison estimates with four anchors
D = direct estimates
H = Handbook estimates
S = Simulator estimates

* Precise HEP estimates are given in Table B.12. This graph is for illustrative purposes only.

Figure B.4 HEP estimates for Level 2 and 3 tasks with simulator anchors for paired comparison estimates.

Table B.13 90-Percent uncertainty bounds with associated statistical confidence limits for Level 1 tasks

Task	Lower Bound	Lower Limit	Upper Limit	Upper Bound	Lower Limit	Upper Limit	Upper Bound / Lower Bound
1	0.00006	0.000008	0.0004	0.008	0.002	0.04	133.3
2	0.0002	0.00004	0.0009	0.006	0.002	0.02	30.0
3	0.00007	0.00001	0.0004	0.009	0.003	0.03	128.6
4	0.00002	0.000005	0.0001	0.003	0.0007	0.01	150.0
5	0.00003	0.000005	0.0001	0.001	0.0002	0.008	33.3
6	0.007	0.002	0.03	0.31	0.19	0.51	44.3
7	0.0002	0.00003	0.002	0.03	0.008	0.09	150.0
8	0.005	0.002	0.01	0.30	0.18	0.50	60.0
9	0.00002	0.000002	0.0001	0.002	0.0002	0.01	100.0
10	0.001	0.0002	0.006	0.20	0.08	0.47	200.0
11	0.00002	0.000006	0.00008	0.002	0.0007	0.004	100.0
12	0.00009	0.00001	0.0006	0.03	0.01	0.11	333.3
13	0.0001	0.00002	0.0006	0.02	0.006	0.04	200.0
14	0.00004	0.000008	0.0002	0.003	0.0007	0.01	75.0
15	0.005	0.002	0.02	0.39	0.21	0.73	78.0

Table B.14 Uncertainty bounds with associated 90-percent statistical confidence limits for Level 2 and 3 tasks

Task	Direct Estimation							Handbook		
	Lower Bound	Lower Limit	Upper Limit	Upper Bound	Lower Limit	Upper Limit	Upper Bound / Lower Bound	Lower Bound	Upper Bound	Upper Bound / Lower Bound
1	0.0006	0.0002	0.002	0.03	0.01	0.06	50.0	0.001	0.009	9.0
2	0.0003	0.00008	0.0009	0.01	0.005	0.04	33.3	0.0003	0.003	10.0
3	0.0001	0.00002	0.0005	0.003	0.0007	0.01	30.0	0.00005	0.005	100.0
4	0.00008	0.00002	0.0004	0.004	0.001	0.02	50.0	0.00005	0.005	100.0
5	0.002	0.0008	0.006	0.26	0.11	0.57	130.0	0.003	0.03	10.0
6	0.00004	0.000007	0.0002	0.003	0.0006	0.02	75.0	0.0003	0.003	10.0
7	0.00009	0.00002	0.0003	0.01	0.005	0.02	111.1	0.00005	0.005	100.0
8	0.0006	0.0002	0.002	0.03	0.01	0.07	50.0	0.0003	0.003	10.0
9	0.001	0.0006	0.003	0.05	0.03	0.08	50.0	0.001	0.009	9.0
10	0.0002	0.00006	0.0006	0.02	0.009	0.04	100.0	0.002	0.02	10.0
11	0.0002	0.00003	0.0008	0.04	0.01	0.11	200.0	0.05	1.00	20.0
12	0.002	0.0008	0.004	0.10	0.05	0.19	50.0	0.02	0.50	25.0
13	0.0005	0.0001	0.001	0.03	0.01	0.08	60.0	0.003	0.03	10.0
14	0.0000004	0.0000002	0.0000009	0.000009	0.000002	0.00003	22.5	0.00001	0.001	100.0
15	0.003	0.001	0.009	0.29	0.18	0.47	96.7	0.03	1.00	33.3
16	0.000009	0.000002	0.00005	0.0003	0.00005	0.002	33.3	0.0003	0.003	10.0
17	0.0001	0.00005	0.0004	0.008	0.004	0.02	80.0	0.0007	0.006	8.6
18	0.001	0.0005	0.002	0.04	0.02	0.08	40.0	0.002	0.02	10.0
19	0.001	0.0004	0.003	0.08	0.03	0.21	80.0	0.01	0.25	25.0
20	0.0005	0.0002	0.001	0.02	0.009	0.04	40.0	0.0001	0.01	100.0

Consistency was examined in two ways. The internal consistency of the judgments of individual experts (within-expert consistency) was analyzed to determine the degree to which the judgments were systematic and not contradictory rather than random. Across-expert consistency was investigated to determine how well the judgments of the experts agreed with each other.

The coefficient of consistency (David, 1963) was used to assess the internal consistency of the experts' paired comparison judgments. This correlation-like statistic is based on the number of intransitive triads in the experts' judgments relative to the number of possible intransitive triads. An intransitive triad is one in which a is judged more likely than b, b more likely than c, and c more likely than a. The coefficient of consistency ranges from 0 where the maximum possible triads are intransitive to 1 where there are no intransitive triads.

Table B.15 gives the coefficient of consistency for each of the 19 experts for both task sets. As can be seen, all coefficients are quite high, indicating that the experts in this study were very consistent in their paired comparison judgments.

Across-expert consistency for both paired comparison judgments and direct estimates was measured by the coefficient of concordance (Siegel, 1956). This statistic describes the extent to which the rank orders of the tasks by the different experts tend to agree. For paired comparison data, tasks were ranked by counting the number of times each task was judged to be more likely than other tasks.

Table B.16 shows the coefficients of concordance for both paired comparison judgments and direct estimates. It also gives coefficients for the estimated lower and upper uncertainty bounds. These coefficients can range from 0, no agreement among experts, to 1, complete agreement. With values of n (the number of tasks) larger than 7, the coefficient of concordance can be tested approximately for significance using chi-square tables by noting that $\chi^2 = m(n - 1)W$, with $n - 1$ degrees of freedom, where n is the number of experts and W is the coefficient of concordance. Using this test, all coefficients are significant at the .001 level as indicated in Table B.16.

The agreement among experts is also demonstrated to some extent by statistical confidence limits on the HEP estimates. Ninety-five percent confidence limits were obtained using the procedures described in Appendix A for all estimates. (Outliers were removed in the computation of statistical confidence limits.) These limits are shown in Tables B.17 through B.26. These tables also show the ratio of the upper limit to the

lower limit which is a measure of the width of the confidence limits. The confidence limits were relatively narrow for all Level 1 estimates and for the Level 2 and 3 direct estimates. All these widths were less than two orders of magnitude and many were less than one. Generally the statistical confidence limits for these estimates were narrower than the uncertainty bound estimates (Tables B.13 and B.14) as would be expected because the confidence limits are based on statistical variation in the nominal estimates for typical conditions while the uncertainty bounds should also reflect atypical conditions.

Table B.15 Coefficients of consistency

Expert	Level 1 Tasks	Level 2 and 3 Tasks
1	0.950	0.830
2	0.771	0.818
3	0.900	0.800
4	0.929	0.785
5	0.671	0.797
6	0.800	0.858
7	0.921	0.852
8	0.829	0.903
9	0.886	0.900
10	0.964	0.912
11	0.943	0.812
12	0.964	0.939
13	0.971	0.873
14	0.986	0.924
15	0.871	0.791
16	0.836	0.903
17	0.864	0.894
18	0.907	0.915
19	0.871	0.885
Mean	0.886	0.863

Table B.16 Coefficients of concordance

Source	Level 1 Tasks (χ^2)	Level 2 and 3 Tasks(χ^2)
Paired Comparison	0.542 (144*)	0.572 (206**)
Direct Estimation	0.390 (104*)	0.423 (153**)
Lower Uncertainty Bound	0.347 (92*)	0.342 (123**)
Upper Uncertainty Bound	0.399 (106*)	0.407 (147**)

* df = 14, $p \leq .001$
** df = 19, $p \leq .001$

Table B.17 Statistical 95-percent confidence limits for Level 1 direct estimation HEP estimates

Task	HEP	Lower Limit	Upper Limit	Upper Limit / Lower Limit
1	.0007	.0001	.004	40.0
2	.001	.0004	.005	12.5
3	.0008	.0002	.004	20.0
4	.0002	.00005	.0009	18.0
5	.0002	.00004	.001	25.0
6	.07	.04	.12	3.0
7	.006	.002	.02	10.0
8	.04	.02	.10	5.0
9	.0001	.00002	.0009	45.0
10	.01	.004	.05	12.5
11	.0003	.0001	.0008	8.0
12	.001	.0002	.007	35.0
13	.002	.0004	.006	15.0
14	.0005	.0001	.002	20.0
15	.03	.01	.08	8.0

Table B.18 Statistical 95-percent confidence limits for Level 1 paired comparison HEP estimates with two anchors

Task	HEP	Lower Limit	Upper Limit	Upper Limit / Lower Limit
1	.0003	.0002	.002	10.0
2	.0006	.0003	.001	3.3
3	.0004	.0002	.0009	4.5
4	.0002	.00008	.0005	6.3
5	.0002	.00009	.0006	6.7
6	.07	.02	.28	14.0
7	.003	.001	.007	7.0
8	.05	.01	.19	19.0
9	.0001	.00004	.0002	5.0
10	.001	.0005	.003	6.0
11	.0002	.00008	.0005	6.3
12	.002	.0009	.005	5.6
13	.001	.0005	.002	4.0
14	.0004	.0002	.001	5.0
15	.04	.01	.14	14.0

Table B.19 Statistical 95-percent confidence limits for Level 1 paired comparison HEP estimates with four anchors

Task	HEP	Lower Limit	Upper Limit	Upper Limit / Lower Limit
1	.0003	.0001	.0007	7.0
2	.0007	.0003	.001	3.3
3	.0004	.0002	.0009	4.5
4	.0002	.00008	.0005	6.3
5	.0003	.0001	.0006	6.0
6	.06	.02	.23	11.5
7	.003	.001	.007	7.0
8	.04	.01	.16	16.0
9	.0002	.00005	.0004	8.0
10	.001	.0005	.002	4.0
11	.0003	.00008	.0005	6.3
12	.002	.0009	.005	5.6
13	.001	.0005	.002	4.0
14	.0005	.0002	.001	5.0
15	.03	.01	.12	12.0

Table B.20 Statistical 95-percent confidence limits for Level 2/3 direct estimation HEP estimates

Task	HEP	Lower Limit	Upper Limit	Upper Limit / Lower Limit
1	.004	.001	.01	10.0
2	.002	.0006	.006	10.0
3	.0005	.0001	.002	20.0
4	.0005	.0001	.003	30.0
5	.02	.008	.03	3.8
6	.0004	.00008	.002	25.0
7	.001	.0005	.003	6.0
8	.006	.003	.01	3.3
9	.01	.007	.02	2.9
10	.003	.002	.006	3.0
11	.003	.001	.008	8.0
12	.02	.009	.03	3.3
13	.007	.002	.02	10.0
14	.000002	.0000006	.000005	8.3
15	.04	.02	.07	3.5
16	.00005	.000007	.0003	42.9
17	.001	.0004	.003	7.5
18	.01	.006	.02	3.3
19	.003	.0008	.01	12.5
20	.003	.001	.009	9.0

Table B.21 Statistical 95-percent confidence limits for Level 2 and 3 paired comparison HEP estimates with two direct estimate anchors

Task	HEP	Lower Limit	Upper Limit	Upper Limit / Lower Limit
1	.003	.0005	.02	40.0
2	.001	.0002	.007	35.0
3	.00002	.000002	.0002	100.0
4	.00005	.000006	.0004	66.7
5	.02	.002	.15	75.0
6	.000007	.0000005	.0001	200.0
7	.00006	.000008	.0005	62.5
8	.006	.0008	.05	62.5
9	.01	.002	.14	70.0
10	.003	.0004	.02	50.0
11	.0001	.00002	.0009	45.0
12	.007	.0008	.06	75.0
13	.003	.0004	.02	50.0
14	.000002	.00000007	.00004	571.4
15	.04	.003	.55	183.3
16	.00001	.000001	.0002	200.0
17	.0001	.00002	.0008	40.0
18	.03	.003	.36	120.0
19	.0008	.0001	.005	50.0
20	.001	.0002	.008	40.0

Table B.22 Statistical 95-percent confidence limits for Level 2 and 3 paired comparison HEP estimates with four direct estimate anchors

Task	HEP	Lower Limit	Upper Limit	Upper Limit / Lower Limit
1	.006	.0008	.04	50.0
2	.002	.0004	.01	25.0
3	.00005	.000004	.0005	125.0
4	.0001	.00002	.001	50.0
5	.02	.002	.24	120.0
6	.00002	.000001	.0003	300.0
7	.0002	.00002	.001	50.0
8	.01	.001	.08	80.0
9	.02	.002	.23	115.0
10	.005	.0008	.04	50.0
11	.0003	.00005	.002	40.0
12	.01	.001	.09	90.0
13	.005	.0007	.03	42.9
14	.000006	.0000003	.0001	333.3
15	.06	.005	.79	158.0
16	.00004	.000004	.0005	125.0
17	.0003	.00004	.002	50.0
18	.04	.004	.54	135.0
19	.002	.0003	.01	33.3
20	.002	.0004	.02	50.0

Table B.23 Statistical 95-percent confidence limits for Level 2 and 3 paired comparison HEP estimates with two Handbook anchors

Task	HEP	Lower Limit	Upper Limit	Upper Limit / Lower Limit
1	.03	.003	.39	130.0
2	.01	.002	.14	70.0
3	.0006	.00003	.01	333.3
4	.001	.0001	.02	200.0
5	.12	.007	1.0	142.9
6	.0003	.00001	.008	800.0
7	.002	.0001	.02	200.0
8	.05	.004	.70	175.0
9	.11	.007	1.0	142.9
10	.03	.003	.36	120.0
11	.003	.0003	.03	100.0
12	.06	.005	.82	164.0
13	.03	.003	.30	100.0
14	.0001	.000002	.005	2500.0
15	.25	.01	1.0	100.0
16	.0005	.00003	.01	333.3
17	.003	.0002	.03	150.0
18	.20	.009	1.0	111.1
19	.01	.001	.11	110.0
20	.02	.002	.16	80.0

Table B.24 Statistical 95-percent confidence limits for Level 2 and 3 paired comparison HEP estimates with four Handbook anchors

Task	HEP	Lower Limit	Upper Limit	Upper Limit / Lower Limit
1	.007	.0007	.08	114.3
2	.004	.0004	.03	75.0
3	.0002	.00001	.005	500.0
4	.0005	.00004	.007	175.0
5	.02	.001	.33	330.0
6	.0001	.000005	.004	800.0
7	.0006	.00005	.007	140.0
8	.01	.0008	.13	162.5
9	.02	.001	.32	320.0
10	.007	.0006	.07	116.7
11	.0009	.00009	.01	111.1
12	.01	.0009	.15	166.7
13	.006	.0006	.06	100.0
14	.00006	.000001	.003	3000.0
15	.04	.001	.91	910.0
16	.0002	.00001	.005	500.0
17	.0008	.00008	.009	112.5
18	.03	.001	.66	660.0
19	.003	.0003	.03	100.0
20	.004	.0004	.04	100.0

Table B.25 Statistical 95-percent confidence limits for Level 2 and 3 paired comparison HEP estimates with two simulator anchors

Task	HEP	Lower Limit	Upper Limit	Upper Limit / Lower Limit
1	.007	.0009	.06	66.7
2	.004	.0008	.02	25.0
3	.0005	.00002	.01	500.0
4	.0009	.00007	.01	142.9
5	.02	.0008	.29	362.5
6	.0003	.000008	.01	1250.0
7	.001	.0001	.01	100.0
8	.01	.0009	.10	111.1
9	.02	.0008	.28	350.0
10	.007	.0009	.05	55.6
11	.001	.0002	.01	50.0
12	.01	.0009	.12	133.3
13	.006	.0009	.05	55.6
14	.0002	.000002	.02	10000.0
15	.03	.0007	.91	1300.0
16	.0005	.J0002	.01	500.0
17	.001	.0002	.01	50.0
18	.02	.0008	.63	787.5
19	.004	.0007	.02	28.6
20	.005	.0008	.03	37.5

Table B.26 Statistical 95-percent confidence limits for Level 2 and 3 paired comparison HEP estimates with four simulator anchors

Task	HEP	Lower Limit	Upper Limit	Upper Limit / Lower Limit
1	.003	.0004	.03	75.0
2	.003	.0005	.01	20.0
3	.001	.00005	.03	600.0
4	.001	.0001	.02	200.0
5	.004	.0002	.08	400.0
6	.001	.00002	.04	2000.0
7	.002	.0001	.02	200.0
8	.004	.0003	.04	133.3
9	.004	.0002	.08	400.0
10	.003	.0004	.02	50.0
11	.002	.0002	.01	50.0
12	.004	.0003	.04	133.3
13	.003	.0004	.02	50.0
14	.0008	.000007	.08	11428.6
15	.005	.0001	.18	1800.0
16	.001	.00005	.03	600.0
17	.002	.0002	.01	50.0
18	.005	.0002	.14	700.0
19	.002	.0005	.01	20.0
20	.003	.0005	.02	40.0

The statistical confidence limits for Level 2 and 3 paired comparison estimates, on the other hand, were usually much wider. Many were more than two orders of magnitude. This is primarily because of variation in the estimates of a and b used in transforming scale values into HEP estimates rather than because of variation in the scale values themselves.

4.1.2 Do Psychological Scaling Techniques Produce Valid HEP Estimates?

Analyses with respect to validity focused on convergent validity, i.e, the degree to which the different approaches to estimating HEPs generally agree. Thus, agreement among direct estimates and paired comparison estimates, as well as among these estimates and Handbook and simulator estimates were analyzed.

The simplest measure of convergence is the correlation between different estimates. These correlations (for log HEPs) are shown in Table B.27; all are statistically significant. Correlations with simulator estimates are not given because of the small number (four) of such estimates available. Scatterplots of these HEPs are shown in Figures B.5 through B.8. The correlation between paired comparison HEP estimates and other sources of estimates is the same regardless of the anchors used to derive the paired comparison since the use of different anchors involves linear transformations that do not affect the correlation.

Convergence was also examined at the individual expert level. Table B.28 shows the correlations between the ranks for direct estimates and for paired comparisons for each task. These correlations are Spearman rank order correlations computed across experts, e.g. for one task the rank assigned by each expert, using each procedure (direct estimates and paired comparisons) is determined. Then within each procedure these assigned ranks are rank ordered and the correlation is computed. Generally, the correlations were higher for Level 1 tasks than for Level 2 and 3 tasks. These correlations can also be used to identify specific tasks on which the two types of judgments did not agree well. For example, on Task 14 for Level 2 and 3 tasks, the correlation was substantial and negative. Unfortunately, this was one of the tasks used as an anchor, which may contribute to some disagreement between HEP estimates from direct estimates and from paired comparisons with direct estimate anchors.

Although the correlations indicate a moderate-to-high degree of convergent validity, particularly at the aggregate level, correlation measures only the linear relationship, not absolute agreement. To further examine the convergence of estimates, analyses of variance (ANOVAs) were used. All dependent measures used in the analyses were log HEPs. For Level 1 tasks, the ANOVA had two factors: task and HEP source. The task factor had 15 levels, one for each task. The HEP source factor had three levels: direct HEP estimates, paired comparison estimates derived using two anchor tasks, and paired comparison estimates using four anchor tasks. A second similar ANOVA was performed with the four anchor tasks removed. The results of these ANOVAs are shown in Table B.29.

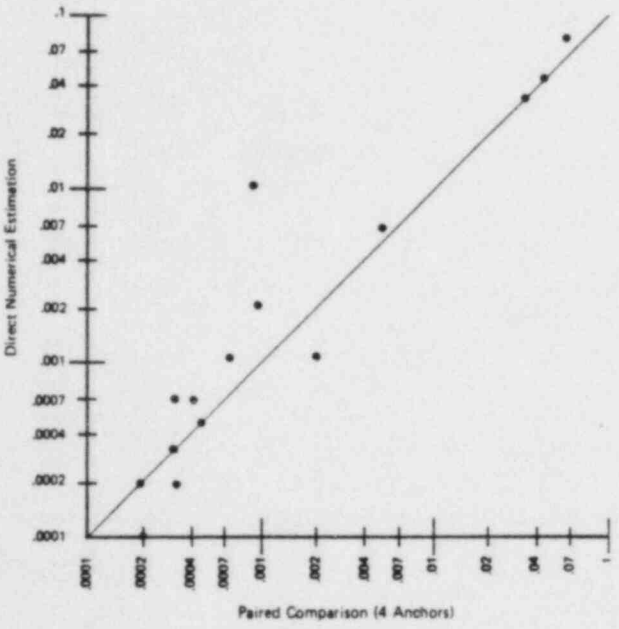


Figure B-5. Level 1 direct numerical estimates and paired comparison estimates with four anchors.

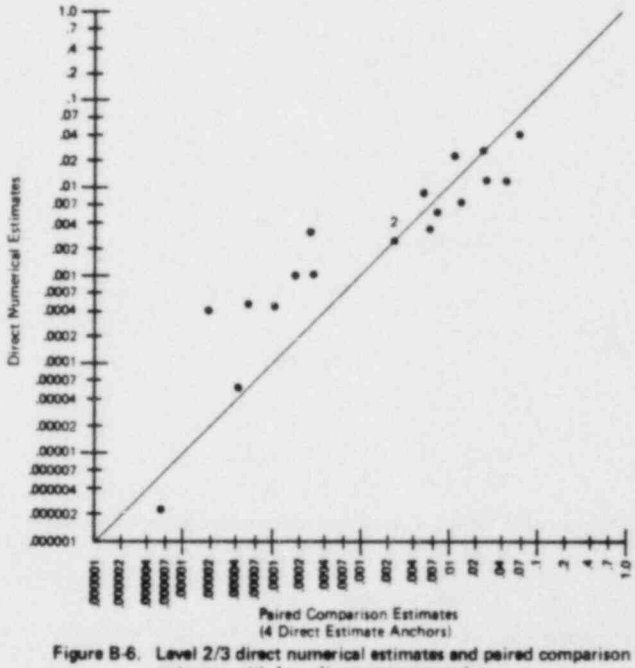


Figure B-6. Level 2/3 direct numerical estimates and paired comparison estimates with four direct estimate anchors.

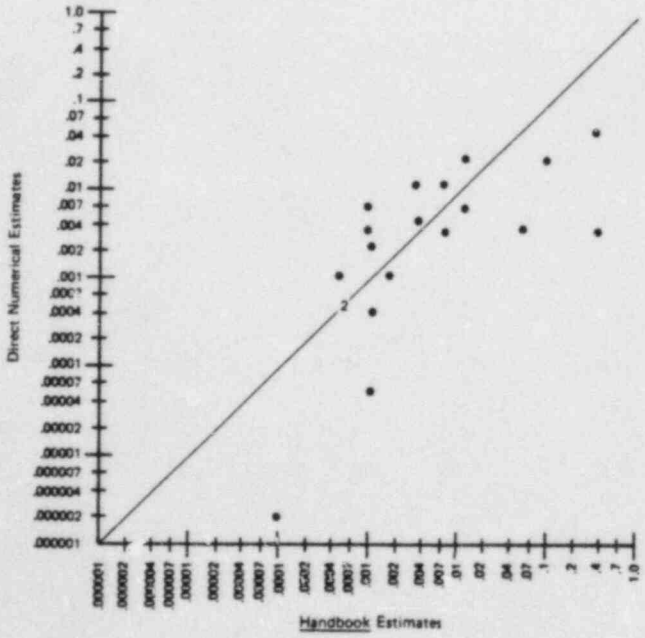


Figure B-7. Level 2/3 direct numerical estimates and Handbook estimates.

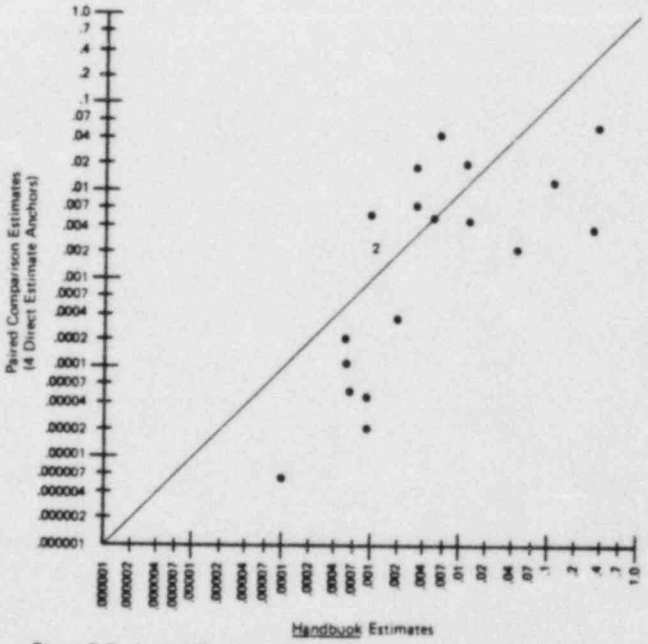


Figure B-8. Level 2/3 paired comparison estimates with four direct estimate anchors and Handbook estimates.

2 = two points coincide

Table B.27 Intercorrelations between HEP estimates

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*

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

Table B.28 Correlations between direct estimates and paired comparisons ranks for each task

Task	Level 1 Tasks	Level 2/3 Tasks
1	.38	.34
2	.65	.14
3	.75	-.18
4	.73	.48
5	.78	.38
6	.41	.14
7	.38	.38
8	.65	.16
9	.69	.47
10	.48	.72
11	.70	.57
12	.26	.70
13	.75	.35
14	.67	-.44
15	.12	.38
16		.70
17		.20
18		.54
19		.48
20		.63
Mean	.56	.36

Table B.29 Analyses of variance for source of estimates and task for Level 1 tasks

All Tasks	SS	df	F	p
Factor				
Source of Estimates	0.2	2	3.6	0.04
Task	33.0	14	73.2	0.0001
Residual	0.9	28		
Anchors Tasks Removed				
Factor				
Source of Estimates	0.3	2	4.1	0.03
Task	20.6	10	58.3	0.001
Residual	0.7	20		

These results indicate that removing anchor tasks had little effect, so additional analyses and subsequent discussion were based on ANOVAs with all tasks included. While the differences in estimates among different HEP sources was not as extensive as across-task differences, they were significant. Thus, planned comparisons were used to further identify the source of differences.

The results of the planned comparisons are shown in Table B.30. For Level 1 tasks, differences in HEP estimates appeared to result primarily from using two versus four anchors.

Table B.30 Planned comparisons for source of estimates for Level 1 tasks

Level 1 Tasks	F	df	p
Paired Comparisons vs. Direct Estimates	3.6	1,14	n.s.
Two Anchors vs. Four Anchors	5.8	1,14	.05

For Level 2 and 3 tasks, similar ANOVAs were performed on differences with Handbook and simulator estimates.

The dependent measures used in these ANOVAs were the differences between logarithms of HEP estimates derived in this study and logarithms of estimates from the Handbook or simulator studies. These measures include both the amount and the direction of differences.

The first ANOVA was performed on all 20 tasks for differences with Handbook estimates. It was a two-way ANOVA with one factor, tasks, having 20 levels; and the second factor, source of HEP estimates, having seven levels: direct estimates, and paired comparison estimates derived using two and four anchors with anchors from direct estimates, Handbook, and simulator sources. A similar ANOVA was performed with the four anchor tasks removed. The results of these ANOVAs are shown in Table B.32. Again, for these ANOVAs, removing anchor tasks has little effect on results. Both source of estimate and task factors were significant, although the F-ratios for the source factor were considerably lower.

Planned comparisons were again used to identify specific differences that created the significance of the source factor. The results of these planned comparisons are given in Table B.32. Again the paired comparison versus direct estimate comparison was not significant. The comparison of sources of anchors confirmed that the major source of differences among estimates resulted from the use of direct estimates versus Handbook estimates as anchors.

Another set of ANOVAs was performed on HEP estimates for those four tasks for which simulator estimates were available. Two ANOVAs were performed: one for the differences with each of Handbook and simulator estimates. The ANOVA designs were similar to those just previously described except the two levels on the source factor using simulator anchors were not included and the number of tasks was reduced. Table B.33 shows these results.

Table B.31 Analyses of variance for differences with Handbook estimates

	SS	df	F	p
Difference-All Tasks				
Factor				
Source of Estimates	16.2	6	17.3	0.001
Task	88.5	19	29.9	0.001
Residual	17.8	114		
Difference-Anchor Tasks Removed				
Factor				
Source of Estimates	12.6	6	17.6	0.001
Task	74.5	15	41.8	0.001
Residual	10.7	90		

Table B.32 Planned comparisons for differences with Handbook estimates

Comparison	F	df	p
Paired Comparisons vs. Direct Estimates	0.3	1,19	n.s
Two Anchors vs. Four Anchors	6.8	1,19	0.05
Simulator Anchors vs. D.E. Anchors	7.5	1,19	0.05
Simulator Anchors vs. Handbook Anchors	5.6	1,19	0.05
D.E. Anchors vs. Handbook Anchors	103.0	1,19	0.001

The pattern of results shown in Table B.33, for differences with simulator estimates was generally similar to that for differences with Handbook estimates.

Table B.33 Analyses of variance for differences with Handbook and simulator estimates for four tasks with simulator estimates

	SS	df	F	p
Difference-Handbook				
Factor				
Source of Estimate	3.7	4	12.5	0.001
Task	3.6	3	16.0	0.001
Residual	0.9	12		
Difference-Simulator				
Factor				
Source of Estimate	3.7	4	12.5	0.001
Task	10.5	3	47.1	0.001
Residual	0.9	12		

In order to investigate further the reason for the differences in the source of estimates factor, i.e., why the differences with Handbook and simulator estimates varied across sets of estimates, we took advantage of the factorial nature of the paired comparison estimates. Three additional three-way ANOVAs were performed using only paired comparison estimates. The first used differences between paired comparison estimates (in logarithms) and Handbook estimates (in logarithms) as dependent measures with the three factors: tasks (20 levels), source of anchors (three levels - direct estimates, Handbook, or simulator), and number of anchors (two-levels - two or four anchors). The second ANOVA was similar to the first with the four anchor tasks removed. Since the results for this ANOVA were similar to those for the first, they are not discussed further. The third ANOVA was performed on differences with simulator estimates. It had only four levels on the task factor and only two levels (direct estimate and Handbook) on the source-of-anchors factor. The results of these ANOVAs are shown in Table B-34. The most interesting result is the source of anchors by number of anchors interaction.

Mean differences in estimates by the source of anchors and number of anchors are plotted in Figures B.9 and B.10. These figures show clearly the interaction. Specifically, differences between sources of anchors were reduced as the number of anchors was increased.

Estimates with Handbook anchors go from being generally larger (positive difference) to being somewhat smaller (negative difference) but closer to zero than before. Estimates with direct estimate anchors go from being considerably smaller than Handbook or simulator estimates (large negative difference) to being less different though still smaller. These results suggest better convergence with four anchors than with two.

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

The most important way in which generalizability was addressed in this study was in the selection and specification of the tasks. They were selected and defined to be generic for any BWR plant, thereby ensuring some degree of generalizability to all BWR plants.

Data analyses played a minor role in addressing this issue. The consistency measures presented in Section 4.1.1 have some application to the issue of generalizability. The across-expert consistency measures provide an indication of the degree of generalizability to estimates from other, similar experts. The moderate, though significant, coefficients of concordance indicate that a reasonable degree of similarity in estimates could be expected from other similar experts. It must be noted, however, that these measures are based on individual estimates. The more important concern for generalizability is the aggregated estimates. By aggregating estimates across several experts, the effect of variation in individual experts is reduced, thus suggesting more generalizability in the aggregated estimates.

Table B.34 Three-way analyses of variance for differences of paired comparison estimates with Handbook and simulator estimates

	SS	df	F	p
Difference with Handbook				
Source				
Task (T)	80.4	19	12.9*	.0005
Source of Anchors(S)	11.5	2	17.5*	.0005
Number of Anchors(N)	0.5	1	47.4	.0001
T x S	12.4	38	28.3	.0001
T x N	1.5	19	6.9	.0001
S x N	4.1	2	175.9	.0001
Residual	0.4	38		
	110.9	119		
Difference with Simulator				
Source				
Task (T)	9.92	3	44.1*	.01
Source of Anchors(S)	2.89	1	38.5*	.01
Number of Anchors(N)	0.03	1	39.5	.01
T x S	0.22	3	93.2	.01
T x N	0.04	3	18.4	.02
S x N	0.78	1	963.8	.0001
Residual	0.002	3		
	13.88	15		

* T x S term was used as error term to provide a conservative test.

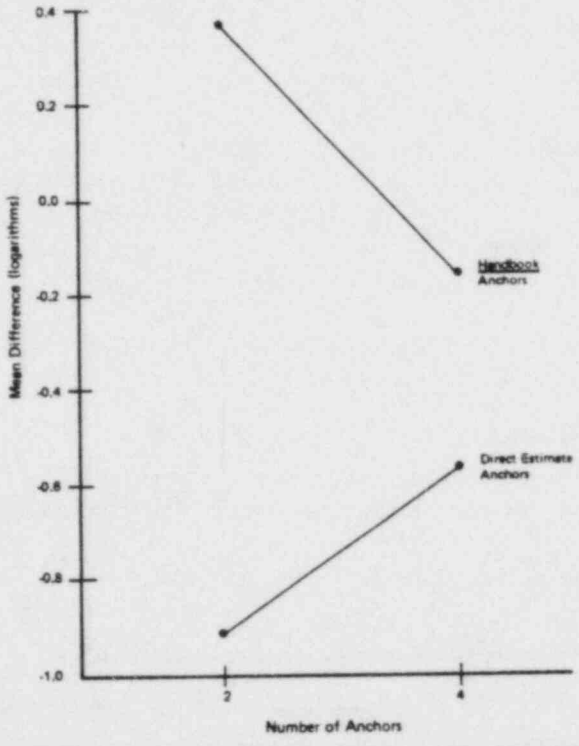


Figure B.9 Interaction of source and number of anchors for differences between paired comparison and simulator estimates (four tasks with simulator estimates)

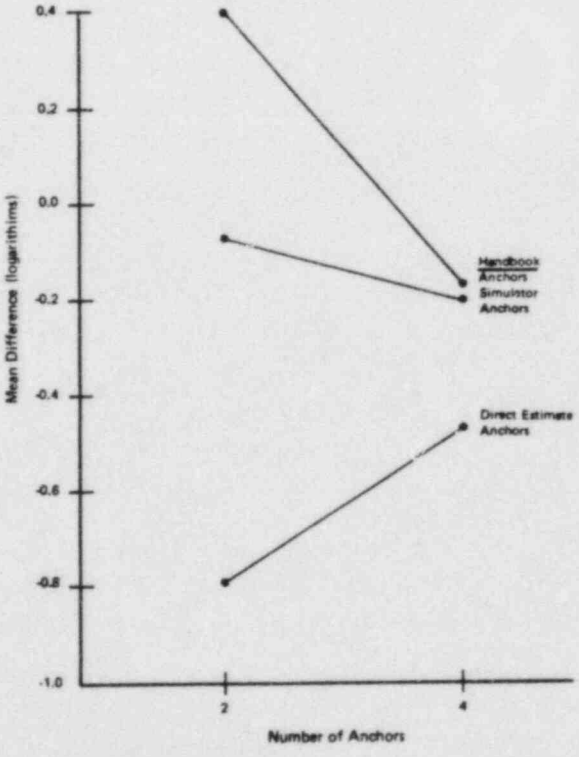


Figure B.10 Interaction of source and number of anchors for differences between paired comparison and Handbook estimates (all 20 tasks).

The ANOVA results presented in Section 4.1.2 also indicate the degree to which similar results can be expected to be obtained under similar circumstances. Significant results can generally be expected to be reproduced, although changes in conditions can have unexpected effects.

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

This issue was addressed less by analyses of data collected than by the design of the study and the specific tasks developed for which HEP estimates were obtained. All tasks were developed so that HEP estimates could be included in the data bank. In addition, PRA practitioners reviewed all tasks (with most attention to Level 1 tasks) to ensure that they were representative of tasks for which HEPs are needed in PRAs.

The one aspect of this issue that is addressed in analyses is the usefulness of the estimates of uncertainty bounds. As a practical matter, uncertainty bounds should play a significant role in PRAs, and if psychological scaling techniques can produce useful estimates of uncertainty bounds, their usefulness for PRAs is enhanced.

As noted in Section 4.1.1 of this appendix, the across-expert consistency for estimates of uncertainty bounds was significant. Coefficients of concordance ranged from approximately .34 to .41.

Two other types of analyses were also conducted to demonstrate the usefulness of the estimates of uncertainty bounds. The first is a comparison of the range between lower and upper uncertainty bounds among those estimated in this study and those from the Handbook. The second is a comparison of where HEP estimates from other sources fall with respect to the uncertainty bounds. Because other HEP and uncertainty bound estimates are available only for Level 2 and 3 tasks, only these tasks were used in these analyses.

Table B.35 shows the ratio of upper to lower uncertainty bounds as a measure of the range of the uncertainty bounds. Thus, for example, a value of 50 in the table indicates that the upper bound is estimated to be 50 times the lower bound. For 15 out of the 20 tasks, the uncertainty bounds estimated in this study were wider than those estimated in the Handbook. A sign test of the difference is significant at the .05 level (Siegel, 1956).

Comparing HEP estimates from sources other than direct estimation with the estimated uncertainty bounds indicates that, for the most part, these HEP estimates fell between the estimated bounds, as shown in Table B.36. To the extent that this was true, it substantiated the credibility of the bounds because the estimated uncertainty bounds were meant to include HEPs under varying conditions. The two sources of HEP estimates that had

a relatively large number of estimates outside the bounds were paired comparison estimates with two direct estimates as anchors, and paired comparison estimates with two Handbook estimates as anchors. This result was consistent with the previously described planned comparisons that indicate much of the differences in HEP estimates was produced by direct estimates compared with Handbook estimates as anchors.

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

This study was designed to minimize the expertise needed by administrators to use the psychological scaling techniques and to see if consistent and valid judgments could be collected under such conditions. The results discussed previously indicate that the judgments were adequately consistent and reliable, thereby implying that the techniques can be used by nonexperts in psychological scaling. In particular, the instructions and procedures described in Appendix A place no requirements on the person who administers data collection sessions other than an ability to understand the instructions. The expertise requirements for the people making the judgments are rather stringent, but do not include any previous knowledge of psychological scaling or probability estimation. Also, a human reliability analyst is needed to develop the tasks for which HEPs are to be estimated, but again no knowledge of psychological scaling is needed. Finally, a data analyst is needed to develop the HEP estimates from the experts' judgments. This person must be able to perform relatively simple computations, but with the step-by-step procedures given in Appendix A, does not require any experience with psychological scaling.

Table B.35 Ratios of upper to lower uncertainty bound estimates from different sources

Task	Estimated	Handbook
1	50.0	9.0
2	33.3	10.0
3	30.0	100.0
4	50.0	100.0
5	130.0	10.0
6	75.0	10.0
7	111.1	100.0
8	50.0	10.0
9	50.0	9.0
10	100.0	10.0
11	200.0	20.0
12	50.0	25.0
13	60.0	10.0
14	22.5	100.0
15	96.7	33.3
16	33.3	10.0
17	80.0	8.6
18	40.0	10.0
19	80.0	25.0
20	40.0	100.0

Table B.36 Comparison of HEP estimates from other sources with estimated uncertainty bounds

HEP Source	Below Lower Bound	Above Upper Bound	n
Level 1			
Paired Comparison-Two D.E. Anchors	0	0	15
Paired Comparison-Four D.E. Anchors	0	0	15
Level 2/3			
Handbook	0	3	20
Simulator	0	1	4
Paired Comparison-Two D.E. Anchors	6	0	20
Paired Comparison-Four D.E. Anchors	2	1	20
Paired Comparison-Two Handbook Anchors	1	7	20
Paired Comparison-Four Handbook Anchors	1	2	20
Paired Comparison-Two Simulator Anchors	1	3	20
Paired Comparison-Four Simulator Anchors	0	3	20

4.1.6 Do the Experts Used in the Psychological Scaling Process Have Confidence in Their Ability to Make the Judgments?

After the experts made all judgments necessary to obtain HEP estimates using direct estimation and paired comparisons, they were asked to provide ratings with respect to several aspects of their judgments. In particular, ratings were obtained regarding the accuracy of judgments, the ease of judgments, and the understandability of the task descriptions. The specific questions asked and the ratings are shown in Table B.37.

The experts tended to be rather neutral in their ratings with a few exceptions:

- Tasks were considered relatively easy to understand.
- Paired comparison judgments were considered accurate.
- The scale for direct estimates was easy to use.

Generally, the experts thought their uncertainty bound estimates were more difficult and less accurate than the HEP estimates, and direct estimates were more difficult and less accurate than paired comparisons.

4.2 Discussion of Technical Issues

The five Technical Issues described in Section 1 of this Appendix are addressed in this section. These issues concern how psychological scaling should be implemented.

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

In addition to the results discussed below, there are two practical differences between these two techniques. Direct estimation can be used when there are relatively few experts available; paired comparisons cannot. Direct estimation also provides a relatively easy procedure for estimating uncertainty bounds. Obtaining uncertainty bounds using paired comparisons is much more time consuming. Statistical confidence limits can be estimated for paired comparisons (as well as for direct estimates), but these limits have a different meaning than appears to be needed for uncertainty bounds.

Analyses previously described can be used to compare these two procedures. Within-expert consistency cannot be measured for direct estimation, but across-expert consistency was slightly higher for paired comparisons (Table B.16). Direct estimates had a higher correlation with handbook estimates than did paired comparisons (Table B.27). The planned comparisons from the ANOVAs, however, indicated that the HEP estimates from the two procedures were not significantly different for either task set (Tables B.30 and B.32).

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

Again, differences between task sets can be determined from the analyses presented above. Both task sets produced very high within-expert consistency, with Level 1 tasks slightly higher (Table B.15). Across-expert consistency was also quite similar for the two task sets with Level 2 and 3 tasks slightly higher (Table B-16).

Convergent validity for the two task sets was similar, with Level 1 tasks having a slightly higher correlation between direct estimates and paired comparisons (Table B.27). The ANOVAs confirmed this similarity because they produced generally similar results for both task sets, and the source of estimates factor was significant at a lower level for Level 1 tasks than for Level 2 and 3 tasks (Tables B.29 and B.31). Furthermore, the experts indicated no difference between the two task sets with respect to how easy the task descriptions were to understand (Table B.37).

Table B.37 Results of ratings indicating confidence level of experts in their judgments

	Questions	SA	A	AS	DS	D	SD	Mean
1.	I accurately judged which incorrect action was more likely.	3	11	2	2	0	0	2.1
2.	I found it easy to make the comparisons.	0	5	8	4	2	0	3.2
3.	My judgments are accurate estimates of the true chances of incorrect action.	3	5	4	3	3	1	3.1
4.	I found it easy to make the estimates of incorrect actions.	0	5	7	4	2	1	3.3
5.	The judgments of uncertainty bounds are accurate.	0	4	7	3	5	0	3.5
6.	I found the uncertainty bounds easy to estimate.	0	2	5	6	4	2	3.9
7.	The scale on which I made the estimates was easy to use.	4	9	2	2	2	0	2.4
8.	The task descriptions in Period 1 were easy to understand.	6	13	0	0	0	0	1.7
9.	The task descriptions in Period 2 were easy to understand.	6	12	1	0	0	0	1.7

Key:

SA = Strongly agree = 1
A = Agree = 2
AS = Agree slightly = 3
DS = Disagree slightly = 4
D = Disagree = 5
SD = Strongly disagree = 6

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

Experts were asked to provide information regarding their education, years of experience, and type of license or certification so that analyses could be performed to determine what effect, if any, these variables had on judgments. All experts had the same type of license or certification, so no effect could be determined for that variable.

Multiple regressions were conducted for five independent variables:

- Coefficient of consistency
- Average log direct estimate
- Average range of uncertainty bounds (in logarithms)
- Average absolute differences for direct estimates with Handbook estimates
- Average absolute difference of direct estimates with simulator estimates

Each of these was regressed onto the education and years of experience variables as well as the confidence ratings described previously in Section 4.1.6. Neither of these dependent variables was a significant predictor of any of the independent variables. This result was expected since there was little variation among experts' levels of education and years of experience.

4.2.4 How Should the Paired Comparison Scale Be Calibrated into a Probability Scale?

A major technical difficulty in the use of paired comparison is that the underlying model, the law of comparative judgment, only produces an interval scale from the judgments. Scale values for human error on this scale must be transformed into probabilities. There are two basic parts of this transformation process: the form of the transformation function and two parameters used in the transformation. These parameters are estimated from tasks with independent HEP estimates called anchors or anchor tasks. This study specifically investigated two questions: (1) should the transformation function be linear or logarithmic, and (2) does using two versus four anchor tasks affect HEP estimates?

To answer the first question, matched-pair t-tests were performed on the logarithm of HEP estimates derived using the linear and the logarithmic transformation, each with the same two direct estimates as anchors. Significant differences were found for both Level 1 and Level 2 and 3 tasks ($t = 6.59$, $df = 14$, $p = 0.001$ for Level 1 tasks; $t = 6.24$, $df = 19$, $p = 0.001$ for Level 2 and 3 tasks).

Because there were differences, it was necessary to determine which function should be used. To make this determination, scale values were correlated with HEP estimates from direct estimation and from the Handbook and also with the logarithms of these estimates. Table B.38 gives the resulting correlations. In all cases the correlations were higher for the logarithms of HEP estimates, indicating that the logarithmic transformation provided a better fit. For the Level 2 and 3 tasks the differences in the correlations were significant (for direct estimates $t(2.5) = 17$, $p \leq .05$; for Handbook estimates $t(2.6) = 17$, $p \leq .05$), but for the Level 1 tasks the difference was not significant. As a result of this analysis, the logarithmic transformation was used throughout for all other analyses.

Table B.38 Correlations of scale values with HEP estimates and log HEP estimates

	Estimates	
Level 1 Tasks	Linear	Logarithmic
Direct Estimates	0.89	0.94
Level 2/3 Tasks		
Direct Estimates	0.67	0.89
Handbook	0.23	0.57

The question of two versus four anchors was addressed primarily by the planned comparisons (Tables B.30 and B.32), and by the three-way ANOVAs (Table B.34). These results show that the number of anchors does have a significant effect for both Level 1 tasks and for Level 2 and 3 tasks.

An examination of the actual HEP estimates (Tables B.11 and B.12) confirms that for Level 1 there was very little difference in estimates. (No difference is more than one significant digit.) There is more variability in the Level 2 and 3 tasks. Although nowhere was the difference between HEP estimates using two versus four anchors (with the same source of anchors) as much as an order of magnitude, the difference did approach this for several tasks.

Differences appeared to be relatively more pronounced for direct estimates and Handbook estimates. Table B.36 provides some additional confirmation regarding where differences occurred for two versus four anchors. This table, which compares HEP estimates with the estimated uncertainty bounds, suggests that paired comparison estimates with two direct estimate anchors tend to be relatively smaller (several estimates

below the estimated lower bounds) and that paired comparison estimates with two Handbook anchors tend to be relatively larger (several estimates above estimated upper bounds). These results were confirmed by Figures B.9 and B.10.

Examination of the tasks used for two anchors suggests an explanation for these results. For two anchors, the tasks used as anchors were numbers 14 and 15. For direct estimation, the HEP estimate for Task 14 was extremely low (.000002), which tended to pull all estimates down when it was used as an anchor. For the Handbook, the HEP estimate for Task 15 was relatively high (.25), thereby pulling all estimates up to some degree when used as an anchor. Using four anchors appears to have lessened the impact of such extreme values. Therefore, more than two anchors should be used, if possible, to reduce the effect of any one anchor.

4.2.5 Can Reasonable Uncertainty Bounds Be Estimated Judgmentally?

Experts were able to estimate uncertainty bounds using direct estimation. The across-expert consistency of these estimates was moderately good, although not as good as the HEP estimates themselves. The estimated uncertainty bounds were generally wider than were the statistical confidence limits for direct estimates and for Level 1 paired comparison estimates (Tables B.13, B.17, B.18, and B.19), as they should be, but were not wider than the statistical confidence limits for the paired comparison estimates on Level 2 and 3 tasks (Tables B.14, and B.20 through B.26). This latter result was more likely to be caused by the procedure for estimating statistical confidence limits than by any problem in the uncertainty bound estimates. The bounds estimated in this study also tended to be somewhat wider than those from the Handbook (Table B.14). HEP estimates from other sources were generally between the bound estimates (Table B.36).

Although the widths of the estimated uncertainty bounds varied considerably, there were many that were quite similar, e.g., between 50 and 100 (Table B.35). Some experts may have adopted a simple strategy of making upper and lower bound estimates a constant multiple of the HEP estimates, e.g., the upper bound would be estimated to be five times the HEP estimate and the lower bound would be one fifth the HEP estimate. Our analyses were unable to determine if such a strategy was used, and, if so, to what extent.

5. DISCUSSION AND CONCLUSIONS

At the outset of this study, the study team identified several issues that needed to be resolved satisfactorily for psychological scaling to be used to estimate HEPs for PRAs. These issues were discussed in Section 1 of this appendix. They were also used as a framework for describing the results of analyses. Here we draw conclusions with respect to these issues and discuss the basis for these conclusions. These conclusions are not based solely on the results of the analyses. Several issues were addressed in the way the study was designed and were largely resolved by simply observing whether or not certain procedures could effectively be implemented. Again, this section is organized around the issues themselves.

5.1 Consistency of Human Error Probability Estimates

Analyses indicated that the judgments of the experts used in this study were consistent enough to support the use of these procedures. The internal consistency of individual experts' paired comparison judgments was exceptional. The across-expert consistency measures used to determine the extent of agreement in judgments across experts were less dramatic, though they still indicated statistically significant agreement.

5.2 Convergent Validity of Human Error Probability Estimates

Given that the judgments required to obtain HEP estimates were sufficiently consistent, their validity then had to be established. As indicated in the discussion of issues, validity can take many forms. Predictive validity could not be established for these procedures because there were no "true" HEPs to be predicted. Therefore, the validity that was investigated here was convergent validity, i.e., the extent to which different procedures for measuring the same concept agreed in their measurements. In investigating convergent validity, we were fortunate to have not only the HEP estimates produced in this study, but also some estimates from other sources, namely the Handbook (Swain and Guttmann, 1983) and simulator studies (though only four estimates were available from the latter; Beare et al., 1984).

Plots of the estimated HEPs suggest that with the exception of a few tasks, the HEPs from various sources were in general agreement. Most differences were less than an order of magnitude, which, given the use of these estimates and the statistical error in estimation, was a satisfactory level of agreement.

Convergent validity was further established by the moderate-to-high correlations between the estimates from various sources. Correlation, however, was not a sufficient measure because it measured only the linear relationship, not the absolute relationship that was of interest with absolute probability measures. Therefore, ANOVAs were also used to determine the degree of convergence.

Generally, the ANOVAs showed that while there were significant differences resulting from different HEP sources, these differences were relatively small compared to differences produced by different tasks. In addition, planned comparisons and ANOVAs on just paired comparison estimates indicated that the differences in sources of HEPs were caused primarily by the anchors used for the paired comparison estimates. Examination of the data suggested that the specific tasks used for anchors had one quite low HEP estimate from direct estimates (Task 14, .000002) and one relatively high estimate from Handbook estimates (Task 15, .25).

Taken together, these results suggest that the convergent validity of the HEP estimates derived from psychological scaling was relatively high, although affected somewhat by the anchors used for paired comparisons. This effect was much less pronounced when four rather than two anchors were used, indicating that more than two anchors should be used if possible.

5.3 Use of Estimates in Probabilistic Risk Assessment and the Data Bank

This issue was addressed by the way the study was designed and conducted rather than by the analyses. Tasks were defined in such a way as to be useful for PRA and to permit inclusion in the data bank. The fact that we could obtain consistent and valid HEP estimates for these tasks indicates that psychological scaling can be used.

Analyses did address the part of this issue related to the need for uncertainty bounds in PRAs. Experts were asked to estimate uncertainty bounds as well as nominal HEPs. These estimates were reasonably consistent across experts, although slightly less consistent than the HEP estimates. The estimated uncertainty bounds were also generally wider than those from the Handbook, a result we interpret as positive since experts often have a tendency to make uncertainty bound estimates too narrow (Stillwell, Seaver, and Schwartz, 1982). An additional comparison of HEP estimates from various sources with the estimated uncertainty bounds showed the positive result that most HEP estimates fell between the bounds. Those that did not could be largely accounted for by the differences in HEPs produced by the anchors in paired comparison estimates, i.e., specifically the same problem with anchors just discussed above.

Our primary concern with respect to the estimation of uncertainty bounds was that experts might have estimated them by simply adjusting their nominal estimates up and down by some relatively consistent factor. While the data show considerable variation in the ranges of uncertainty bounds across tasks, the ranges of a large number of tasks were quite similar, so we cannot tell whether such a strategy was used.

Some changes in procedures might alleviate this possibility, if it exists. Experts could be asked to make the uncertainty bound estimates at a different time than when nominal estimates are made, preferably before. A second possibility would be to ask for worst-case scenarios (e.g., define performance shaping factors so as to produce the highest possible HEP) rather than for bounds, and again to get these estimates at a different time. These approaches should help ensure that the experts would be thinking about the range of possible HEPs.

5.4 Generalizability of Human Error Probability Estimates

Again, this issue was addressed primarily in the study design. Tasks were defined to be generic for BWR plants, thus providing generalization to all such plants with the possible need for some adjustment based on plant-specific conditions.

Generalizability was also supported by the across-expert consistency. These measures suggest that similar estimates would be obtained from other, similar experts. To the degree that other experts differ from those used in this study, some differences in estimates could occur. This study did not address the degree, if any, of these possible differences. The use of HEPs based on aggregating judgments across experts, however, helps to ensure reasonable generalizability to other groups of experts.

5.5 Confidence of Experts in Judgments

The only judgments in which the experts expressed more than a modest degree of confidence were paired comparisons. They had the least confidence in uncertainty bound judgments. These results were not surprising. Usually experts without experience in estimating probabilities will not be particularly confident in their judgments, even though the judgments are reasonably accurate. Confidence can be expected to increase as experience with this type of judgment increases. While a lack of confidence may be a stumbling block in the acceptance of psychological scaling for use in PRAs, it should not be interpreted as suggesting the judgments are not sound.

5.6 Use of Data Collectors Without Expertise in Psychological Scaling

The data collection procedures used to obtain the needed expert judgments were designed to be used by someone who does not have any experience with psychological scaling. These procedures were pretested with a nonexpert, revised as needed, and then used by a nonexpert in actual data collection. Since the actual data collection went smoothly, and the needed judgments were obtained and were consistent, we can reasonably conclude that the techniques can be used without a psychological scaling expert.

5.7 Technical Issues

In addition to the issues just discussed, this study was designed to address several technical issues regarding how psychological scaling should be implemented.

This study tested two techniques: direct numerical estimation and paired comparisons. Our results indicated that there were few differences in the HEP estimates resulting from these procedures. The choice between them should be made primarily on practical grounds. For example, in this study, obtaining the required paired comparison judgments took about 30 minutes, and obtaining the direct estimates including uncertainty bounds took about 90 minutes. This difference is probably relatively unimportant, but if more tasks are included, the time for direct estimates will go up approximately linearly with the number of tasks, while the time for paired comparisons goes up with the square of the number of tasks. (See Seaver and Stillwell, 1983, and Stillwell, Seaver, and Schwartz, 1982, for some ways to reduce the number of paired comparisons required.) Other practical considerations include the need for uncertainty bound estimates (obtainable most efficiently with direct estimates although paired comparisons could be used to estimate worst-case scenario HEPs) and the number of experts available. Another consideration could be that the experts considered their paired comparison judgments more accurate than their direct estimates.

A second technical issue was to determine what, if any, differences in results were due to the type of tasks (Level 1 and Levels 2 and 3). Results were basically similar for both types of tasks, indicating that HEP estimates for Level 1 tasks could be obtained for use in PRAs.

We were also interested in identifying anything in the experts' backgrounds (e.g., education, experience, type of license/certification) that might affect judgments. None of the background measures obtained was related to the experts' judgments for this homogeneous group of experts.

The transformation of scale values into HEPs for the paired comparisons appears to be the most critical step in the use of paired comparisons. Results indicated conclusively that a logarmithmic relationship was appropriate. They also suggested that using four anchors rather than two could reduce potential differences in HEP estimates created by extreme estimates in the anchors used.

Finally, the experts were able to estimate uncertainty bounds, but these estimates were subjected to only limited analysis. There was some possibility, however, that a simple response strategy was used for these estimates.

ATTACHMENT 1 TO APPENDIX B

LEVEL A* TASK DESCRIPTIONS

[*Level A refers to Level 1]

ATTACHMENT 1 TO APPENDIX B

LEVEL A TASKS

(1) During a loss-of-off-site-power transient, several failures have rendered the high pressure coolant injection (HPCI) and the reactor core isolation cooling (RCIC) systems inoperable. Core cooling can be established with either low pressure coolant injection or low pressure core spray, but pressure must be reduced first. Procedural guidelines specify manual actuation of the automatic depressurization system (ADS) to reduce pressure. What is the likelihood that the operator will fail to actuate the ADS manually within 10 minutes?

(2) During a loss-of-off-site-power transient, the generator has tripped, the reactor has scrammed, and the normal feedwater system is inoperable. According to the procedures, the reactor water level should be recovered and maintained by manually operating the reactor core isolation cooling (RCIC) system. What is the likelihood that the operator will fail to operate the RCIC system correctly?

(3) During a loss-of-off-site-power transient, the generator has tripped, the reactor has scrammed, and the normal feedwater system is inoperable. According to the emergency procedures, the operator must operate the nuclear instrumentation system by inserting the source and intermediate range monitors to verify that reactor power is decreasing following the scram. What is the likelihood that the operator will fail to operate the nuclear instrumentation system correctly?

(4) One of the main steam relief valves inadvertently opens. The operator, after successfully closing the valve, is monitoring the suppression pool temperature. The indicated temperature of the suppression pool is 95°F. According to procedures, this requires that the residual heat removal (RHR) system be manually placed in the suppression pool cooling mode. What is the likelihood that the operator will fail to actuate the suppression pool cooling mode of RHR?

(5) One of the main steam relief valves inadvertently opens. The operator mistakenly thinks he has reclosed the valve; however, the valve is still open. The operator properly places the RHR system in the suppression pool cooling mode when the temperature reaches 95°F. The temperature eventually reaches 110°F. The procedure then specifies that the operator must scram the reactor manually. What is the likelihood that the operator will fail to scram the reactor?

(6) A transient has occurred, the high pressure coolant injection (HPCI) system is operating, and the suppression pool cooling is inoperable. The operator notices that the HPCI system has inadvertently switched to suppression pool suction. The condensate storage tank (CST) level and the suppression pool level are both normal. The operator checks and finds that the CST water is still plentiful. What is the likelihood

that the operator will not realize that high suppression pool temperature could ultimately fail HPCI due to loss of net positive suction head?

(7) A transient has occurred, the high pressure coolant injection (HPCI) system is operating, and the suppression pool cooling system is inoperable. The operator notices that the HPCI system has automatically switched to suppression pool suction. He checks and finds that the condensate storage tank (CST) water is still plentiful. The operator realizes that high suppression pool temperature could ultimately fail HPCI. What is the likelihood that he will fail to take the appropriate action to return the system manually so that the CST is the water supply?

(8) The plant is experiencing an extended station blackout (loss of on-site and off-site power) greater than 5 hours. Continued operation of the reactor core isolation cooling (RCIC) and high pressure coolant injection (HPCI) systems depends on sufficient room cooling for the equipment. What is the likelihood that the operator will fail to take precautions such as opening doors or providing other ventilation to ensure that the vital system equipment is being properly cooled?

(9) A transient has occurred, and the reactor has failed to scram. The operator, realizing what has happened, consults the emergency procedure for dealing with an anticipated transient without scram. The procedure states that he should attempt to trip the reactor manually. The operator attempts this but is unsuccessful. The procedure then calls for him to use the standby liquid control (SLC) system. What is the likelihood that the operator will fail to initiate SLC within 5-10 minutes after he reads the procedural step telling him to do so?

(10) A station blackout including total failure of the diesel generator system has just occurred. After the first immediate steps have been taken, the emergency procedures are referenced. What is the likelihood that the operator will attempt to restore off-site power before he attempts to restore power using the diesel generators?

(11) A transient has occurred, and the reactor protection system has failed to insert the rods. All attempts to manually scram the reactor have failed. According to the procedures, the operator is now required to manually insert the rods. What is the likelihood that the operator will fail to attempt to manually insert the rods using reactor manual control?

(12) A loss-of-coolant accident (LOCA) has occurred. The residual heat removal service water (RHRSW) system must be manually initiated within the first 30 minutes after the transient to obtain successful long-term decay heat removal. The emergency operating procedures contain detailed instructions on operating the RHRSW. What is the likelihood that the operator will fail to recognize that he should initiate RHRSW within 30 minutes?

(13) A loss-of-coolant accident (LOCA) has occurred. The residual heat removal service water (RHRSW) system must be manually initiated to obtain successful long-term decay heat removal. The emergency operating procedures contain detailed instructions on operating the RHRSW, but the operator has so much to do he fails to operate the RHRSW. After 40 minutes the operator gets a high suppression pool temperature alarm. What is the likelihood that he will then fail to diagnose the problem correctly and take steps to initiate RHRSW?

(14) The residual heat removal (RHR) system is providing shutdown cooling when the running RHR pump trips because of an electrical fault. The operator acknowledges that the pump tripped. Procedures state that the operator is to restore shutdown cooling. What is the likelihood that the operator will fail to attempt to restore RHR cooling within 10 minutes?

(15) The high pressure coolant injection (HPCI) system and the reactor core isolation cooling (RCIC) system have automatically initiated. The plant has experienced a total loss of instrument air. The pneumatic valves that control the cooling water to HPCI and RCIC room coolers do not open on demand because of the loss of instrument air. Opening these valves requires local operation. What is the likelihood that the operator will fail to open these valves within 1 hour?

ATTACHMENT 2 TO APPENDIX B

LEVEL B* TASK DESCRIPTIONS

[*Level B refers to Levels 2 and 3]

ATTACHMENT 2 TO APPENDIX B

LEVEL B TASKS

(1) An operator chooses the wrong switch from a set of switches that all look similar and are identified only by labels.

(2) An operator chooses the wrong switch from a set of switches that all look similar and are grouped according to their functions.

(3) An operator chooses the wrong switch from a set of switches that all look similar and are arranged with clearly drawn mimic lines.

(4) The controls in a control room are all designed so that they are moved to the right if the operator wants to turn on a component. The operator makes an error and turns a rotary control that has three or more positions to the left when he intends to turn the component on.

(5) Two or more locally operated valves are not clearly labeled. In addition, they are very similar in size and shape, they are in the same state (either open or closed), and they all have been tagged in a similar fashion. (The tags are all the same color, etc.) The operator attempts to place one of these valves back in service, but he mistakenly chooses the wrong one.

(6) A locally operated valve is clearly and unambiguously labeled and is not located near any similar-appearing valves. The operator intends to place the valve back in service, but he mistakenly chooses the wrong one.

(7) An operator reads the wrong meter in a group of meters that all look similar. They are arranged with clearly drawn mimic lines.

(8) An operator reads the wrong meter in a group of meters that all look similar. The meters are grouped according to their functions.

(9) An operator reads the wrong meter in a group of meters that all look similar and are identified only by labels.

(10) An equipment or auxiliary operator selects the wrong circuit breaker from a group of circuit breakers that are located outside the control room. The circuit breakers are densely grouped and identified only by labels.

(11) A locally operated valve has a rising stem and a position indicator. An auxiliary operator, while using written procedures to check a valve lineup, fails to realize that the valve is not in its proper position after a maintenance person has performed a procedure intended to restore it to its proper position after maintenance.

(12) A meter has jammed so that the pointer is stuck on the scale. When an operator reads the meter, he fails to realize that it is jammed even though the value displayed is erroneous.

(13) An operator incorrectly reads information from a graph that is in a procedure.

(14) Assume that five annunciators are alarming. An operator fails to act on any of them.

(15) Assume that ten annunciators have alarmed and an operator has responded to nine of them. The operator fails to act on the one remaining annunciator.

(16) An operator reads a digital indicator incorrectly.

(17) A chart recorder has normal bands indicated on the scale. An operator incorrectly interprets the value shown when he scans the recorder.

(18) A chart recorder does not have normal bands indicated on the scale. An operator incorrectly interprets the value shown when he scans the recorder.

(19) A meter has normal bands indicated on the scale. An operator does not notice that the meter is out of range after he performs an initial control room evaluation. No written materials are used.

(20) An operator intends to operate a 10-position rotary selector switch. He sets it to the wrong position.

ATTACHMENT 3 TO APPENDIX B

SAMPLE PAGES FROM RESPONSE BOOKLET

PERIOD 1

PAIRED COMPARISON JUDGMENTS: SET 1

You are to assume the following for the tasks that follow (Level B tasks):

- There is a one-man team in the control room during the performance of these tasks.
- These tasks take place during routine operations.
- The person performing the action in each task has been in his current job position for at least six months.
- No one involved in performing these tasks is wearing any type of protective clothing.

[Note: A different set of assumptions was used for Level A tasks.]

EXAMPLES OF COMPLETED PAIRED JUDGMENTS

Of the two possible tasks listed below, check the task that is the most likely to occur.

X 1. An operator chooses the wrong switch from a set of switches that all look similar and are grouped according to their functions.

_____ 2. A locally operated valve does not have a rising stem or a position indicator. An auxiliary operator, while using written procedures to check a valve lineup, fails to realize that the valve is not in its proper position after a maintenance person has performed a procedure intended to restore it to its proper position after maintenance.

_____ 1. During a loss-of-off-site-power transient, several failures have rendered the high pressure coolant injection (HPCI) and the reactor core isolation cooling (RCIC) systems inoperable. Core cooling can be established with either low pressure coolant injection or low pressure core spray, but pressure must be reduced first. Procedural guidelines specify manual actuation of the automatic depressurization system (ADS) to reduce pressure. What is the likelihood that the operator will fail to actuate the ADS manually within 10 minutes?

X 2. During a loss-of-off-site-power transient, the generator has tripped, the reactor has scrammed, and the normal feedwater system is inoperable. According to the procedures, the reactor water level should be recovered and maintained by manually operating the reactor core isolation cooling (RCIC) system. What is the likelihood that the operator will fail to operate the RCIC system correctly?

[Note: Applicable assumptions for Level A or Level B (Levels 2 and 3 of the Data Bank) preceded these examples.]

INSTRUCTIONS FOR COMPLETION OF DIRECT ESTIMATE AND UNCERTAINTY BOUNDARY JUDGMENTS

Once you have read and understood the task on the left side of the page, put an X on the point on the scale on the right that represents your best estimate of the chances of the incorrect action occurring. Remember, you are to assume that the operator does not have an unlimited amount of time in which to take action. Next, place slash marks to indicate upper and lower boundaries so that you are 90 percent certain that the value will fall within those boundaries. If a mark or exact value that represents your estimate does not appear on the scale (e.g., 1 chance in 3,500), place your X or slash at the approximate position on the scale and write your estimate to the right of the scale.

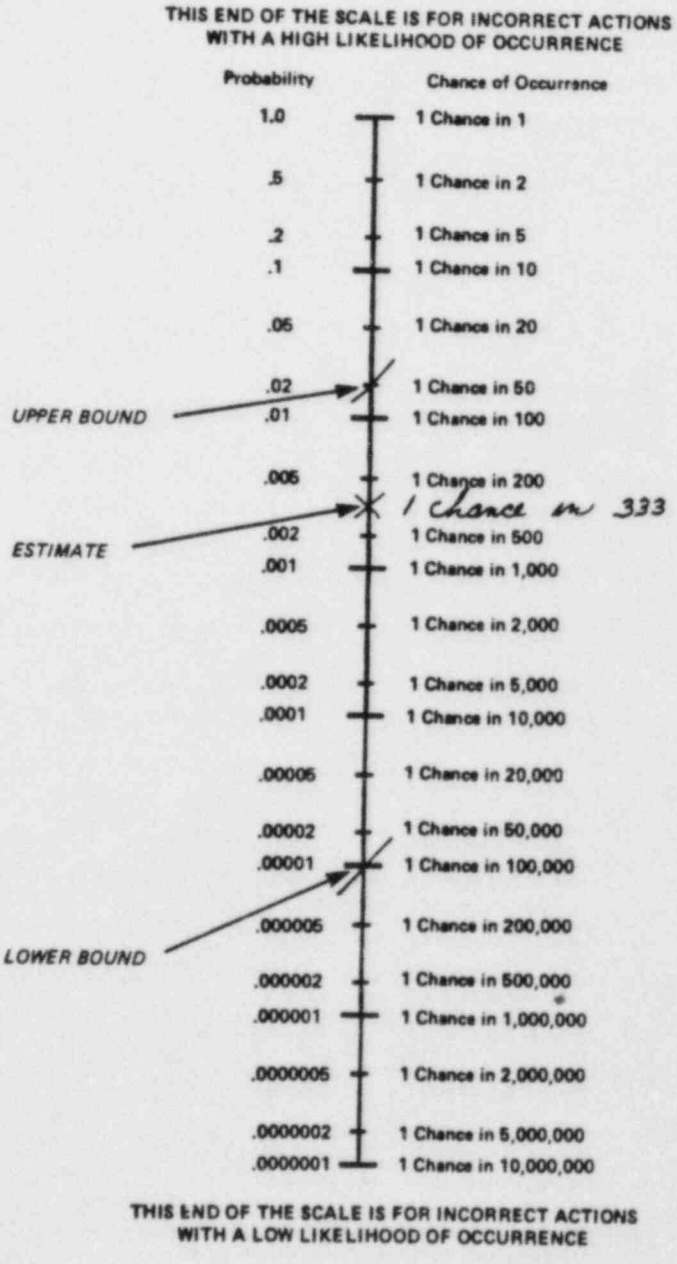
[Note: Applicable assumptions for Level A or Level B (Levels 2 and 3 of the Data Bank) preceded these instructions.]

EXAMPLE OF COMPLETED DIRECT ESTIMATE

Estimate the chances that the following will occur:

An operator is performing an initial control room evaluation. He fails to detect that an indicator light shows that a component is in an incorrect state. No written materials are used.

What assumptions did you make that affected your answer?



[Note: Applicable assumptions for Level A or Level B (Levels 2 and 3 of the Data Bank) preceded this example.]

BACKGROUND INFORMATION

1. Your name (optional): _____

2. Present educational level attained (circle one):

   a. High school degree or equivalent

   b. Trade school (1-2 years) or Associate's degree

   c. Bachelor's degree

   d. Master's degree

   e. Other (please explain)

   _____

3. Power plant experience: (years)

Operations	Training	Other	
Military:	_____	_____	_____
Fossil (commercial):	_____	_____	_____
Nuclear (commercial):	_____	_____	_____
Other:	_____	_____	_____
Total:	_____	_____	_____

4. Present type of license or certification (circle one):

   a. Former RO or SRO

   b. BWR-certified instructor

   c. Other (explain: _____)

CIRCLE THE APPROPRIATE RESPONSE

ANSWER KEY

SA:	Strongly agree
A:	Agree
AS:	Agree slightly
DS:	Disagree slightly
D:	Disagree
SD:	Strongly disagree

Questions on Paired Comparisons only:

1. I accurately judged which incorrect action was more likely.	SA	A	AS	DS	D	SD
2. I found it easy to make the comparisons.	SA	A	AS	DS	D	SD

Questions on Direct Estimates only:

3. My judgments are accurate estimates of the true chances of incorrect action.	SA	A	AS	DS	D	SD
4. I found it easy to make the estimates of incorrect actions.	SA	A	AS	DS	D	SD
5. The judgments of uncertainty bounds are accurate.	SA	A	AS	DS	D	SD
6. I found the uncertainty bounds easy to estimate.	SA	A	AS	DS	D	SD
7. The scale on which I made the estimates was easy to use.	SA	A	AS	DS	D	SD

In General:

8. The task descriptions in Period 1 were easy to understand.	SA	A	AS	DS	D	SD
9. The task descriptions in Period 2 were easy to understand.	SA	A	AS	DS	D	SD

Did you assume, when making your estimates, that written procedures were being used in those cases where we did not specify whether procedures were to be used or not?

Is there anything else that you would like to tell us about the assumptions you used to make your judgments?

COMMENTS

Thank you for your participation.

ATTACHMENT 4 TO APPENDIX B

DATA COLLECTION SESSION INSTRUCTIONS

ATTACHMENT 4 TO APPENDIX B

INSTRUCTIONS FOR THE DATA COLLECTION SESSIONS

1. INSTRUCTIONS FOR SESSION ADMINISTRATOR

The purpose of this study is to test several procedures by which experts can judge the likelihood that certain incorrect actions will occur during nuclear power plant operation. The data collected during these experimental sessions will be used to determine the quality of estimates of human error probability produced by expert judgment.

Quality estimates of these likelihoods will be used in probabilistic risk assessments (PRAs) of nuclear power plants. Several research efforts are currently under way to develop the quantitative data needed for this purpose. Our results will be used to determine how best to obtain these data in terms of both cost and quality.

In this study, the experts will be asked to make judgments about two sets of incorrect actions. One set is made up of incorrect actions that are very specific. Each of these actions is part of the more complex behavioral sequences undertaken by an operator in a nuclear power plant. An example of a specific incorrect action is "read the wrong meter in a group of meters that all look very similar and are identified only by labels." This simple action is part of many behavioral sequences that the operator performs.

The other set of actions consists of complex behavioral sequences. Each of these sequences requires that several individual actions be correctly performed for the entire sequence to be successful. For example, the experts will be asked to make judgments about the likelihood of the following situation:

> During a loss-of-off-site-power transient, several failures have rendered the high pressure coolant injection (HPCI) and the reactor core isolation cooling (RCIC) systems inoperable. Core cooling can be established with either low pressure coolant injection or low pressure core spray, but pressure must be reduced first. Procedural guidelines specify manual actuation of the automatic depressurization system (ADS) to reduce pressure. What is the likelihood that the operator will fail to actuate the ADS manually within 10 minutes?

The experts will be asked to provide three kinds of judgments. In the first type of judgment they will be asked to determine which of a pair of incorrect actions is more likely to occur. The experts will be asked to make judgments about all possible pairs from within each set. The second type of judgment will be a direct estimate of the likelihood of the incorrect action. The expert will be asked to express an estimate of the chances that the incorrect action will occur, out of some number of opportunities. For example, the expert will be asked, "What do you think the chances are that an operator will choose the wrong switch from a set of switches that all look similar and are identified only by labels?" The experts will be provided a scale that shows

successively lower chances of occurrence of the event, from 1 chance in 1 to 1 chance in 10,000,000, and will be instructed to place a mark on the scale that corresponds to their estimate of the chances that the incorrect action will occur.

The third type of judgment will be an estimate of the uncertainty about a direct estimate of the likelihood of an action. The experts will be asked to place boundaries around their estimates of the chances of an action's occurrence so that they are certain that 90 percent of the time the actual chances of an incorrect action's occurrence will be within those boundaries. For each expert's judgment on each scale, these boundaries should surround the mark placed for their exact estimate. These boundaries will provide information about the expert's uncertainty about their judgments.

In the final portion of the experiment, the experts will be asked for information about their experience and training, the usability of the procedures, and the quality of their own judgments. We will ask for the highest educational level attained, total years of power plant experience (including commercial and military, fossil and nuclear), and the present type of operator's license that each expert has. The "quality of judgments" questions will express their level of agreement with a statement like "I accurately judged which incorrect action was the more likely," while the usability questions will ask, for example, that the experts express their relative agreement with the statement "I found it easy to make the paired comparison judgments."

Sample questions are provided for the experts in the response booklets. You can use these to ensure that the procedures are correctly understood. With these questions, you are only seeking to determine whether the experts understand the use of the judgmental procedures, not whether they agree with what you think is the "correct" probability. Make no attempt to change their answers except to explain further the type of judgment being asked for if their judgments are inconsistent with what is required by a procedure. An example of inconsistent judgments that should be pointed out to the expert is a case where the mark for the upper uncertainty boundary is put below the mark for the error probability. By definition the boundaries should surround the mark for the error probability with the upper boundary always above and the lower boundary below. This sort of inconsistency should be pointed out to the expert and an attempt made to reexplain the judgment required.

During the session, if the experts ask questions, you are not allowed to provide impromptu answers. Refrain from answering any technical questions. If additional guidance is needed to clarify the instructions, please provide it. If you are not sure what the appropriate action is, consult the psychological scaling expert who will be present during the sessions.

2. GENERAL INSTRUCTIONS TO BE READ TO EXPERTS

The purpose of this study is to test several procedures for eliciting judgments of the likelihood of certain events. The events with which the study is concerned are various incorrect actions performed in the process of operating a nuclear power plant. During any specific action or operation, for example, closing a valve, there will be a chance that the operator will make an error, that is, fail to close the valve correctly. As experienced trainers, you have as much or more first-hand knowledge about the chances of incorrect actions than anyone else does. For this reason, we have asked you to participate in the study.

We will be asking you to make judgments about the likelihood of various incorrect actions that might occur during the operation of a nuclear power plant. You should try to incorporate all your knowledge of power plant operations and the likelihood of the various actions into these judgments. As an example, you may know that some of these actions are more difficult or complex. Thus, one might expect the chance of incorrectly performing that action to be higher. Some actions may occur during more stressful situations, so those actions might have a higher likelihood of being performed incorrectly. As you make each judgment, try to think of all information that is relevant to the chances of performing that action incorrectly. You are to assume that the operator does not have an unlimited amount of time in which to take action. He must respond to the system demands prior to the onset of consequences that would result from his inaction. In other words, he must respond within the period of time required by the situation and his specific plant design.

We will ask you to make several types of judgments. Instructions for each type of judgment will be given as needed, along with examples. There will be two sets of tasks, Level A tasks and Level B tasks. Each type of task will be associated with a different set of assumptions. The assumptions for Level A tasks are:

- A senior reactor operator and a reactor operator are in the control room at all times.
- When reading the Level A tasks, assume that everything that is not underlined is "given" and sets the stage for the underlined question.
- The person(s) performing the action in each task has been in his current job position for at least six months.
- No one involved in performing these tasks is wearing any type of protective clothing.
- The operator(s) does not have an unlimited amount of time in which to take action.

The assumptions for Level B tasks are:

- There is a one-man team in the control room during the performance of these tasks.

- These tasks take place during routine operations.
- The person performing the action in each task has been in his current job position for at least six months.
- No one involved in performing these tasks is wearing any type of protective clothing.

The assumptions associated with each set of tasks are clearly labeled in the response booklets.

This data collection session will include two short breaks. It is important to have independent judgments from each of you, so please do not discuss your judgments with each other. If you have any questions, please let me know. I will try to answer your questions in a way that does not lead to differences between your judgments and those of others who have not heard your questions and my responses.

3. INSTRUCTIONS TO BE READ TO EXPERTS FOR PERIOD 1

(ADMINISTRATOR: Pass out the set of task definitions.)

You have been given a set of task definitions. Please read through them to be sure you understand each task. If you have questions, please ask. We may be able to provide you with some guidance. However, I will not be able to provide extensive explanation of the tasks because we want all experts in these sessions to be given the same amount of information.

Level A tasks are defined on pages 1 to 3. Assume the following conditions for these tasks:

- A senior reactor operator and a reactor operator are in the control room at all times.
- When reading the Level A tasks, assume that everything that is not underlined is "given" and sets the stage for the underlined question.
- The person(s) performing the action in each task has been in his current job position for at least six months.
- No one involved in performing these tasks is wearing any type of protective clothing.
- The operator(s) does not have an unlimited amount of time in which to take action.

Level B tasks are listed on pages 4 and 5. Assume the following conditions for these tasks:

- There is a one-man team in the control room during the performance of these tasks.
- These tasks take place during routine operations.
- The person performing the action in each task has been in his current job position for at least six months.
- No one involved in performing these tasks is wearing any type of protective clothing.

In addition, assume that typical control room conditions exist for both Level A and Level B tasks. Also, when making the judgments, remember that we are only interested in operator errors, not in any additional equipment failures.

(ADMINISTRATOR: Wait for experts to review the task definitions. Ask if they are ready to proceed. Then pass out response booklets.)

Review the assumptions on the first page of the response booklet. (Pause.) You will be shown tasks in pairs. Each task involves an incorrect action that an operator could take. For each pair, decide which of the two incorrect

actions is more likely to occur. Thus, a very difficult action, even though the operator might not perform it often, should have a higher relative chance of being performed incorrectly than an easier action. Remember that you are not trying to determine which task describes a better or worse operating situation or control design. Rather, you are simply judging which task an operator is more likely to perform incorrectly. Mark your choice with a checkmark in the space provided.

Examine the completed example in your response booklet. The first incorrect action was checked. For our hypothetical respondent, this reflects the belief that action no. 1 is more likely to occur out of the chances it has to occur than action no. 2. The second example shows that our hypothetical respondent believes that action no. 2 is more likely to occur than action no. 1.

We would like you to make a choice for each pair of events. Do not leave any pair of actions unchecked, and do not check both actions of any one pair. If you are unsure of the relative likelihood of the two actions, make your best guess as to which of the two is more likely.

At this time, please turn to the next page in your response booklet. You should find two uncompleted examples. Mark these examples as you have been instructed. Are there any questions about the procedure?

After you have completed all responses, please give me your response booklet. Then you may take a short break. If you have any questions while you are making the judgments, please let me know.

4. INSTRUCTIONS TO BE READ TO EXPERTS FOR PERIOD 2

(ADMINISTRATOR: Pass out response booklets.)

The next set of judgments will be similar to the first but with different tasks. The procedure will be the same; that is, you will judge which of a pair of incorrect actions is more likely to occur. The only change will be in the actions themselves and in the assumptions you are to use when making the judgments. Once again, questions may be asked anytime during the session. Remember to read the list of assumptions. They are different than the assumptions used in Period 1. After completing your responses, please give me your booklet and you may again take a short break.

5. INSTRUCTIONS TO BE READ TO EXPERTS FOR PERIOD 3

(ADMINISTRATOR: Pass out response booklets.)

Now, rather than comparing pairs of tasks, you will be giving numerical estimates of the chances that an operator will perform a single action incorrectly. The response booklet shows an example of this type of judgment. The action in this case is

> "An operator will incorrectly read information from a graph that is a procedure."

Your first judgment will be an estimate of the chances that such an error will be made. In making this estimate, you should consider all possible operators and all circumstances that fit the task description. Taking these possibilities into account, we want your best estimation of how likely this incorrect action is. Would you expect such an incorrect action to occur once out of every ten times these circumstances occur? once out of a thousand? once out of a million? or something in between?

The scale on the right side of the page has been provided for you to mark your estimate. The scale is marked with both the chance of occurrence and the corresponding probability. For example, one chance in 100 corresponds to a probability of point zero one. A probability of point zero five is the same as five chances in 100 or one chance in 20. You should put an X on the scale at the point that corresponds to your estimate of the chances or the probability that the given incorrect action will occur.

If the scale does not include the exact chances or probability that you estimate, mark the scale with an X in approximately the correct position and write your estimate to the right of the scale. For example, if you think the given incorrect action would occur about three times in a thousand, you should put an X between one chance in 200 and one chance in 500. This estimate corresponds to one chance in 333 or point zero zero three. In addition to your X, you should write either "1 in 333" or ".003" to the right of the scale. The X labeled "estimate" on the example corresponds to this judgment.

We recognize that you cannot know for sure exactly what the chances of these incorrect actions are. Your response is simply your best estimate. Therefore, we also want to get estimates from you about what you think the range of chances for this incorrect action is. You might think, for example, that while your best estimate is one chance in 333, the actual chances may be quite a bit higher or lower than this estimate, depending on circumstances. Therefore, we will also ask you for an upper and lower estimate or bound that represents the range over which this estimate may vary. Specifically, you should indicate an upper and lower bound so that you think there is a 90 percent chance that in any circumstances the probability of error is between these bounds. In determining these bounds, you should consider the range of circumstances in which this task is performed. This includes different operators (e.g., with different capabilities or training), the physical and mental condition of operators (e.g., tired versus rested, under stress), the quality of instructions, and the physical conditions of the plant (e.g., temperature, layout of controls).

Suppose the upper bound is one chance in 50 and the lower bound is one chance in 100,000. This would indicate that you are quite certain--90 percent sure--that the actual chance of this incorrect action occurring is between these bounds. These bounds would also be marked on the scale as indicated in the example.

The scale provided goes as low as one chance in 10 million. You do not need to use the entire scale unless you think the chances of error are really that low. The scale is provided only so that you may respond as you think appropriate, and not as any guide to what we consider appropriate responses. However, if you use the very top of the scale, where the probabilities are between .5 and 1.0, it is particularly important that you write in the actual probability.

Each page has a place for you to list any assumptions that you might have made when making your estimate. You are <u>not required</u> to fill in this information for each task. Factors that might be listed include such things as time of day, environmental conditions in the control room, and quality of procedures. Indicate for each assumption whether it applies to the best estimate or to the uncertainty bounds or both.

Now, if you have any questions about how you are to give these estimates, I will try to answer them.

(ADMINISTRATOR: Answer questions.)

If there are no further questions, on the next two pages of the booklet are examples for which you should mark your best estimate and your uncertainty bounds. After you have completed these examples, I will check your responses to be sure they are consistent with the kinds of responses we are looking for.

After I have examined your responses to the sample questions, you will be free to proceed through the booklet. The tasks appear, one per page, with a scale to mark your responses. Estimate the chances of occurrence and upper and lower uncertainty bounds for each. You are free to turn back to previous pages once you have completed them. Tasks from Level A and Level B are grouped separately in this booklet. The assumptions for each level of tasks are presented prior to the questions on those tasks. Remember: each set of assumptions for the two sets of tasks is different. If you do not have any questions, proceed with the judgments.

6. INSTRUCTIONS TO BE READ TO EXPERTS FOR PERIOD 4

(ADMINISTRATOR: Pass out Section 4 of the response booklet.)

You have now finished all the judgments on the incorrect actions. We have one final request, which is that you answer the questions provided. You do not have to give your name, but it would be helpful to us so that we can follow up on any of your comments and ask questions if we need to. If you do give your name, it will be kept confidential.

The questions about your past experience are for research purposes only. For example, we want to determine whether the number of years of experience of an individual makes a difference in the responses that individual gave. The additional questions ask your opinion about the responses you provided. Any additional comments you have are welcome and can be entered in the space provided. If you have any questions, feel free to ask them. Otherwise, proceed with the questions.

APPENDIX C

HUMAN ERROR PROBABILITY ESTIMATES

1. INTRODUCTION

Appendix C contains the human reliability estimates that were collected as part of this project. It is written for those who have an interest in or a need for estimates of human error probabilities (HEPs) and the associated uncertainty bounds. An overview of the project is presented in the main report, the details of how the estimates can be generated are presented in Appendix A, and a description of how the estimates were generated and analyzed for this evaluation is contained in Appendix B.

1.1 Description of the Estimates

The estimates that were collected correspond to two separate task lists. These task lists are identified as Level 1 and Levels 2 and 3 to correspond to the Human Reliability Data Bank as described in NUREG/CR-2744 Volume 2 (Comer, et al., 1983). The estimates are presented according to these categories in Tables C.1, C.2, C.3, and C.4.

The estimates displayed in the tables were generated using two different psychological scaling techniques: paired comparison and direct numerical estimation. The paired comparison technique resulted in HEP estimates but did not yield uncertainty bounds. Because the technique yields ratios and not probabilities (see Appendix A), the paired comparisons were converted to probabilities using "anchors." These anchors were probabilities taken from a source other than paired comparisons. Therefore, the tables display paired comparison HEP estimates that were derived using two anchors and those that were derived using four anchors. There are three different sources of anchors; direct estimates that were generated during the same data collection sessions as the paired comparisons, results of simulator experiments as reported in NUREG/CR-3309 (Beare et al., 1984), and Handbook data from NUREG/CR-1278 (Swain and Guttmann, 1983). Direct numerical estimates resulted in both HEP estimates and uncertainty bound estimates.

Simulator and Handbook estimates correspond to Level 2 and 3 tasks and not to Level 1 tasks. Of the 20 Level 2 and 3 tasks, four correspond to simulator data. Therefore, paired comparison estimates for Level 1 are anchored with Level 1 direct estimates and not with either of the other two sources.

1.2 Assumptions

During the data collection sessions, the experts were asked to make judgments about the likelihood of occurrence of the task statements. In

addition to the information provided in each task, the following were assumed:

Level 1 Tasks

- A senior reactor operator and a reactor operator are in the control room at all times.
- When reading the Level 1 tasks, assume that everything that is not underlined is "given" and sets the stage for the underlined question.
- The person(s) performing the action in each task has been in his current job position for at least six months.
- No one involved in performing these tasks is wearing any type of protective clothing.
- The operator(s) does not have an unlimited amount of time in which to take action.

Level 2 and 3 tasks

- There is a one-man team in the control room during the performance of these tasks.
- These tasks take place during routine operations.
- The person performing the action in each task has been in his current job position for at least six months.
- No one involved in performing these tasks is wearing any type of protective clothing.

In addition to these specific assumptions, the experts were asked to assume that typical control room conditions existed at the time the task was performed.

1.3 Cautions To Be Considered When Using the Estimates

Convergent validity, and across-expert and within-expert (where relevant) reliablility were established for these estimates as part of this study. However, no means for establishng predictive validity were available.

The tasks for which data were collected were tailored specifically for boiling water reactors (BWRs). The experts who judged the tasks were BWR-certified instructors. Also the simulator data which were used as anchors for the Level 2 and 3 paired comparisons were gathered from BWR simulators. Therefore, it is suggested that the HEPs and associated uncertainty bounds displayed in the tables only be applied to BWRs.

2. TABLES

Each of the tables is described separately below.

Table C.1 contains the task descriptions, direct estimate HEPs, and direct estimate uncertainty bounds for Level 1 tasks.

Table C.2 contains the task descriptions, direct estimate HEPs, and direct estimate uncertainty bounds for Level 2 and 3 tasks.

Table C.3 contains the HEP estimates for Level 1 tasks from direct numerical estimation and paired comparisons with two anchors and four anchors.

Table C.4 contains the HEP estimates for Level 2 and 3 tasks from direct estimates and from paired comparisons. The sources of anchors for the paired comparison estimates are also displayed. They are direct estimates, Handbook estimates and simulator estimates.

Any of the estimates displayed in Appendix C may be used in probabilistic risk assessments (PRAs); however, if a decision must be made as to which source of estimates to use, the authors recommend that either the estimates from direct estimates or the paired comparisons derived using four anchors be used. Also, we recommend that if paired comparison estimates are used, they be from estimates with direct estimates or Handbook anchors.

Table C.1 Level 1 tasks and direct estimate HEPs and uncertainty bounds

Task Descriptions	
(1) During a loss-of-off-site-power transient, several failures have rendered the high pressure coolant injection (HPCI) and the reactor core isolation cooling (RCIC) systems inoperable. Core cooling can be established with either low pressure coolant injection or low pressure core spray, but pressure must be reduced first. Procedural guidelines specify manual actuation of the automatic depressurization system (ADS) to reduce pressure. What is the likelihood that the operator will fail to actuate the ADS manually within 10 minutes?	HEP* = 0.0007 LB** = 0.00006 UB*** = 0.008

* HEP = Human Error Probability
** LB = Direct Estimate Lower Uncertainty Bounds
*** UB = Direct Estimate Upper Uncertainty Bounds

Table C.1 Continued

Task Descriptions	
(2) During a loss-of-off-site-power transient, the generator has tripped, the reactor has scrammed, and the normal feedwater system is inoperable. According to the procedures, the reactor water level should be recovered and maintained by manually operating the reactor core isolation cooling (RCIC) system. What is the likelihood that the operator will fail to operate the RCIC system correctly?	HEP = 0.001 LB = 0.0002 UB = 0.006
(3) During a loss-of-off-site-power transient, the generator has tripped, the reactor has scrammed, and the normal feedwater system is inoperable. According to the emergency procedures, the operator must operate the nuclear instrumentation system by inserting the source and intermediate range monitors to verify that reactor power is decreasing following the scram. What is the likelihood that the operator will fail to operate the nuclear instrumentation system correctly?	HEP = 0.0008 LB = 0.00007 UB = 0.009
(4) One of the main steam relief valves inadvertently opens. The operator, after successfully closing the valve, is monitoring the suppression pool temperature. The indicated temperature of the suppression pool is 95°F. According to procedures, this requires that the residual heat removal (RHR) system be manually placed in the suppression pool cooling mode. What is the likelihood that the operator will fail to actuate the suppression pool cooling mode of RHR?	HEP = 0.0002 LB = 0.00002 UB = 0.003
(5) One of the main steam relief valves inadvertently opens. The operator mistakenly thinks he has reclosed the valve; however, the valve is still open. The operator properly places the RHR system in the suppression pool cooling mode when the temperature reaches 95°F. The temperature eventually reaches 110°F. The procedure then specifies that the operator must scram the reactor manually. What is the likelihood that the operator will fail to scram the reactor?	HEP = 0.0002 LB = .00003 UB = 0.001
(6) A transient has occurred, the high pressure coolant injection (HPCI) system is operating, and the suppression pool cooling is inoperable. The operator notices that the HPCI system has inadvertently switc to suppression pool suction. The condensate storage tank (CST) level and the suppression pool level are both normal. The operator checks and finds that the CST water is still plentiful. What is the likelihood that the operator will not realize that high suppression pool temperature could ultimately fail HPCI due to loss of net positive suction head?	HEP = 0.07 LB = 0.007 UB = 0.31

Table C.1 Continued

Task Descriptions	
(7) A transient has occurred, the high pressure coolant injection (HPCI) system is operating, and the suppression pool cooling system is inoperable. The operator notices that the HPCI system has automatically switch to suppression pool suction. He checks and finds that the condensate storage tank (CST) water is still plentiful. The operator realizes that high suppression pool temperature could ultimately fail HPCI. What is the likelihood that he will fail to take the appropriate action to return the system manually so that the CST is the water supply?	HEP = 0.006 LB = 0.0002 UB = 0.03
(8) The plant is experiencing an extended station blackout (loss of on-site and off-site power) greater than 5 hours. Continued operation of the reactor core isolation cooling (RCIC) and high pressure coolant injection (HPCI) systems depends on sufficient room cooling for the equipment. What is the likelihood that the operator will fail to take precautions such as opening doors or providing other ventilation to ensure that the vital system equipment is being properly cooled?	HEP = 0.04 LB = 0.005 UB = 0.30
(9) A transient has occurred, and the reactor has failed to scram. The operator, realizing what has happened, consults the emergency procedure for dealing with an anticipated transient without scram. The procedure states that he should attempt to trip the reactor manually. The operator attempts this but is unsuccessful. The procedure then calls for him to use the standby liquid control (SLC) system. What is the likelihood that the operator will fail to initiate SLC within 5-10 minutes after he reads the procedural step telling him to do so?	HEP = 0.0001 LB = 0.00002 UB = 0.002
(10) A station blackout including total failure of the diesel generator system has just occurred. After the first immediate steps have been taken, the emergency procedures are referenced. What is the likelihood that the operator will attempt to restore off-site power before he attempts to restore power using the diesel generators?	HEP = 0.01 LB = 0.001 UB = 0.20
(11) A transient has occurred, and the reactor protection system has failed to insert the rods. All attempts to manually scram the reactor have failed. According to the procedures, the operator is now required to manually insert the rods. What is the likelihood that the operator will fail to attempt to manually insert the rods using reactor manual control?	HEP = 0.0003 LB = 0.00002 UB = 0.002

Table C.1 Continued

Task Descriptions	
(12) A loss-of-coolant accident (LOCA) has occurred. The residual heat removal service water (RHRSW) system must be manually initiated within the first 30 minutes after the transient to obtain successful long-term decay heat removal. The emergency operating procedures contain detailed instructions on operating the RHRSW. What is the likelihood that the operator will fail to recognize that he should initiate RHRSW within 30 minutes?	HEP = 0.001 LB = 0.00009 UB = 0.03
(13) A loss-of-coolant accident (LOCA) has occurred. The residual heat removal service water (RHRSW) system must be manually initiated to obtain successful long-term decay heat removal. The emergency operating procedures contain detailed instructions on operating the RHRSW, but the operator has so much to do he fails to operate the RHRSW. After 40 minutes the operator gets a high suppression pool temperature alarm. What is the likelihood that he will then fail to diagnose the problem correctly and take steps to initiate RHRSW?	HEP = 0.002 LB = 0.0001 UB = 0.02
(14) The residual heat removal (RHR) system is providing shutdown cooling when the running RHR pump trips because of an electrical fault. The operator acknowledges that the pump tripped. Procedures state that the operator is to restore shutdown cooling. What is the likelihood that the operator will fail to attempt to restore RHR cooling within 10 minutes?	HEP = 0.0005 LB = 0.00004 UB = 0.003
(15) The high pressure coolant injection (HPCI) system and the reactor core isolation cooling (RCIC) system have automatically initiated. The plant has experienced a total loss of instrument air. The pneumatic valves that control the cooling water to HPCI and RCIC room coolers do not open on demand because of the loss of instrument air. Opening these valves requires local operation. What is the likelihood that the operator will fail to open these valves within 1 hour?	HEP = 0.03 LB = 0.005 UB = 0.39

Table C.2 Level 2 and 3 tasks and direct estimate HEPs and uncertainty bounds

Task Descriptions	
(1) An operator chooses the wrong switch from a set of switches that all look similar and are identified only by labels.	HEP* = 0.004 LB** = 0.0006 UB*** = 0.03
(2) An operator chooses the wrong switch from a set of switches that all look similar and are grouped according to their functions.	HEP = 0.002 LB = 0.0003 UB = 0.01
(3) An operator chooses the wrong switch from a set of switches that all look similar and are arranged with clearly drawn mimic lines.	HEP = 0.0005 LB = 0.0001 UB = 0.003
(4) The controls in a control room are all designed so that they are moved to the <u>right</u> if the operator wants to turn <u>on</u> a component. The operator makes an error and turns a rotary control that has three or more positions to the left when he intends to turn the component on.	HEP = 0.0005 LB = 0.00008 UB = 0.004
(5) Two or more locally operated valves are not clearly labeled. In addition, they are very similar in size and shape, they are in the same state (either open or closed), and they all have been tagged in a similar fashion. (The tags are all the same color, etc.) The operator attempts to place one of these valves back in service, but he mistakenly chooses the wrong one.	HEP = 0.02 LB = 0.002 UB = 0.26
(6) A locally-operated valve is clearly and unambiguously labeled and is not located near any similar-appearing valves. The operator intends to place the valve back in service, but he mistakenly chooses the wrong one.	HEP = 0.0004 LB = 0.00004 UB = 0.003
(7) An operator reads the wrong meter in a group of meters that all look similar. They are arranged with clearly drawn mimic lines.	HEP = 0.001 LB = 0.00009 UB = 0.01
(8) An operator reads the wrong meter in a group of meters that all look similar. The meters are grouped according to their functions.	HEP = 0.006 LB = 0.0006 UB = 0.03
(9) An operator reads the wrong meter in a group of meters that all look similar and are identified only by labels.	HEP = 0.01 LB = 0.001 UB = 0.05

* HEP = Human Error Probability
** LB = Direct Estimate Lower Uncertainty Bounds
*** UB = Direct Estimate Upper Uncertainty Bounds

Table C.2 Continued

Task Descriptions	
(10) An equipment or auxiliary operator selects the wrong circuit breaker from a group of circuit breakers that are located outside the control room. The circuit breakers are densely grouped and identified only by labels.	HEP = 0.003 LB = 0.0002 UB = 0.02
(11) A locally-operated valve has a rising stem and a position indicator. An auxiliary operator, while using written procedures to check a valve lineup, fails to realize that the valve is not in its proper position after a maintenance person has performed a procedure intended to restore it to its proper position after maintenance.	HEP = 0.003 LB = 0.0002 UB = 0.04
(12) A meter has jammed so that the pointer is stuck on the scale. When an operator reads the meter, he fails to realize that it is jammed even though the value displayed is erroneous.	HEP = 0.02 LB = 0.002 UB = 0.10
(13) An operator incorrectly reads information from a graph that is in a procedure.	HEP = 0.007 LB = 0.0005 UB = 0.03
(14) Assume that five annunciators are alarming. An operator fails to act on any of them.	HEP = 0.000002 LB = 0.0000004 UB = 0.000009
(15) Assume that ten annunciators have alarmed and an operator has responded to nine of them. The operator fails to act on the one remaining annunciator.	HEP = 0.04 LB = 0.003 UB = 0.29
(16) An operator reads a digital indicator incorrectly.	HEP = 0.00005 LB = 0.000009 UB = 0.0003
(17) A chart recorder has normal bands indicated on the scale. An operator incorrectly interprets the value shown when he scans the recorder.	HEP = 0.001 LB = 0.0001 UB = 0.008
(18) A chart recorder does not have normal bands indicated on the scale. An operator incorrectly interprets the value shown when he scans the recorder.	HEP = 0.01 LB = 0.001 UB = 0.04
(19) A meter has normal bands indicated on the scale. An operator does not notice that the meter is out of range after he performs an initial control room evaluation. No written materials are used.	HEP = 0.003 LB = 0.001 UB = 0.08
(20) An operator intends to operate a 10-position rotary selector switch. He sets it to the wrong position.	HEP = 0.003 LB = 0.0005 UB = 0.02

Table C.3 HEP estimates for Level 1 tasks

Task	Direct Numerical Estimation	Paired Comparisons	
		2 Anchors	4 Anchors
1	0.0007	0.0003	0.0003
2	0.001	0.0006	0.0007
3	0.0008	0.0004	0.0004
4	0.0002	0.0002	0.0002
5	0.0002	0.0002	0.0003
6	0.07	0.07	0.06
7	0.006	0.003	0.003
8	0.04	0.05	0.04
9	0.0001	0.0001	0.0002
10	0.01	0.001	0.001
11	0.0003	0.0002	0.0003
12	0.001	0.002	0.002
13	0.002	0.001	0.001
14	0.0005	0.004	0.0005
15	0.03	0.04	0.03

Table C.4 HEP estimates for Level 2 and 3 tasks

HEP ESTIMATES							
		Paired Comparisons					
	Direct Numerical Estimation	Direct Numerical Estimation		Handbook		Simulator	
Task		2 Anchors	4 Anchors	2 Anchors	4 Anchors	2 Anchors	4 Anchors
1	0.004	0.003	0.006	0.03	0.007	0.007	0.003
2	0.002	0.001	0.002	0.01	0.004	0.004	0.003
3	0.0005	0.00002	0.00005	0.0006	0.0002	0.0005	0.001
4	0.0005	0.00005	0.0001	0.001	0.0005	0.0009	0.001
5	0.02	0.02	0.02	0.12	0.02	0.02	0.004
6	0.0004	0.000007	0.00002	0.0003	0.0001	0.0003	0.001
7	0.001	0.00006	0.0002	0.002	0.0006	0.001	0.002
8	0.006	0.006	0.01	0.05	0.01	0.01	0.004
9	0.01	0.01	0.02	0.11	0.02	0.02	0.004
10	0.003	0.003	0.005	0.03	0.007	0.007	0.003
11	0.003	0.0001	0.0003	0.003	0.0009	0.001	0.002
12	0.02	0.0[illegible]	0.01	0.06	0.01	0.01	0.004
13	0.007	0.003	0.005	0.03	0.006	0.006	0.003
14	0.000002	0.000002	0.000006	0.0001	0.00006	0.0002	0.0008
15	0.04	0.04	0.06	0.25	0.04	0.03	0.005
16	0.00005	0.00001	0.00004	0.0005	0.0002	0.0005	0.001
17	0.001	0.0001	0.0003	0.003	0.0008	0.001	0.002
18	0.01	0.03	0.04	0.20	0.03	0.02	0.005
19	0.003	0.0008	0.002	0.01	0.003	0.004	0.002
20	0.003	0.001	0.002	0.02	0.004	0.005	0.003

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 TW11OLN
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 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. 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: October YEAR: 1984

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

DATE REPORT ISSUED
MONTH: YEAR:

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

21. NO. OF PAGES
233

20. SECURITY CLASS *(This page)*
Unclassified

22. PRICE
$

*U.S. GOVERNMENT PRINTING OFFICE: 1984-0-576-035/4399