NUREG/CR-3518
BNL-NUREG-51716
VOL. 1

SLIM-MAUD: AN APPROACH TO ASSESSING HUMAN ERROR PROBABILITIES USING STRUCTURED EXPERT JUDGMENT

VOLUME I: OVERVIEW OF SLIM-MAUD

D.E. Embrey, P. Humphreys, E.A. Rosa,
B. Kirwan, and K. Rea

Date Published - March 1984

DEPARTMENT OF NUCLEAR ENERGY, BROOKHAVEN NATIONAL LABORATORY
UPTON, LONG ISLAND, NEW YORK 11973

Prepared for
United States Nuclear Regulatory Commission
Office of Nuclear Regulatory Research
Contract No. DE-AC02 76CH00016

NUREG/CR-3518
BNL-NUREG-51716
VOL. 1
AN, RX

SLIM-MAUD: AN APPROACH TO ASSESSING HUMAN ERROR PROBABILITIES USING STRUCTURED EXPERT JUDGMENT

VOLUME I: OVERVIEW OF SLIM-MAUD

D.E. Embrey,* P. Humphreys,** E.A. Rosa,†
B. Kirwan,* and K. Rea*

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

**London School of Economics and Political Science
Houghton Street
London WC2A 24E, England

Manuscript Completed - February 1984
Date Published - March 1984

Prepared under Contract to
†DEPARTMENT OF NUCLEAR ENERGY
BROOKHAVEN NATIONAL LABORATORY
ASSOCIATED UNIVERSITIES, INC.
UPTON, LONG ISLAND, NEW YORK 11973

Prepared for
UNITED STATES NUCLEAR REGULATORY COMMISSION
HUMAN FACTORS AND SAFEGUARDS BRANCH
OFFICE OF NUCLEAR REGULATORY RESEARCH
FIN. NO. A-3219

NOTICE

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

The views expressed in this report are not necessarily those of the 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

ABSTRACT

This two-volume report presents the procedures and analyses performed in developing an approach for structuring expert judgments to estimate human error probabilities. Volume I presents an overview of work performed in developing the approach: SLIM-MAUD (Success Likelihood Index Methodology, implemented through the use of an interactive computer program called MAUD--Multi-Attribute Utility Decomposition). Volume II provides a more detailed analysis of the technical issues underlying the approach.

CONTENTS

CONTENTS (CONTINUED)

FIGURES

ACKNOWLEDGMENTS

We would like to acknowledge the contributions of a number of individuals and organizations* to this report. W. J. Luckas, Jr. of Brookhaven National Laboratory (BNL) was contract monitor and T. G. Ryan of the U.S. Nuclear Regulatory Commission (NRC) was program manager for the project. We thank them for their support and encouragement. The United Kingdom Atomic Energy Authority provided facilities for the experimental study described in Section 4; the original data of the Phase I experiment formed part of a Masters dissertation submitted to Birmingham University by B. Kirwan. The Decision Analysis Unit of the London School of Economics made available, free of charge, a non-exclusive end user license for the MAUD program used in this study. We would also like to thank J. C. Williams of the Safety and Reliability Directorate of the United Kingdom Atomic Energy Authority for his assistance in this work. Finally, we would like to acknowledge the contribution of our subjects in both the experimental and field studies.

*Work was carried out under the auspices of the U.S. Nuclear Regulatory Commission.

1. PURPOSE OF THE WORK

This two-vclume report presents the results of a research program devoted to the refinement and further development of the Success Likelihood Index Methodology (SLIM). SLIM comprises a set of procedures for eliciting and organizing the estimates of experts concerning the probability of success or failure of specified human actions in nuclear power plants. The goal is to produce human error probability (HEP) estimates in support of human reliability analysis (HRA) segments of probabilistic risk assessments (PRA) of nuclear power plants.

The SLIM research program consisted of three phases of investigation: phase I involved an experimental evaluation of SLIM; in phase II a field test of SLIM was conducted; and in phase III SLIM was linked to a computer based elicitation procedure based upon Multi-Attribute Utility Decomposition (MAUD). This report discusses the results obtained in each of the separate phases of investigation, together with a detailed plan for the next phase of research, the assessment of the utility of the MAUD-based implementation of SLIM (SLIM-MAUD).

Volume I of this report presents an overview of SLIM, a discussion of the results of the experiment and field test, a discussion of the linking of SLIM to MAUD, and an outline of a Test Plan for the next phase of research.

Volume II discusses criteria for evaluating subjective techniques for estimating human reliability, presents an in-depth, theoretical and technical discussion of SLIM and the SLIM-MAUD implementation, and provides a detailed description of the Test Plan for the next research phase. In addition, task descriptions used by subjects in the SLIM experiment and definitions of performance shaping factors (PSFs) used in the field test are presented, together with an example of a frame-by-frame computer interaction from a SLIM-MAUD session, along with the results produced.

2. BACKGROUND

PRA is an approach which has been extensively applied in recent years to the nuclear, chemical, offshore oil drilling, and other industries in order to identify the potential risks in a system and to evaluate their probability of occurrence and the expected consequences. The PRA process involves first modeling the system to evaluate the various ways in which subsystem failures can occur, and then assigning probabilities to these failures. These are subsequently combined together to give the overall probability of failure for the system as a whole. Originally, PRA was primarily concerned with failures of hardware components such as pumps and valves, particularly where these were part of safety related systems. In recent years there has been a growing realization that human actions can have a significant effect on the likelihood of failure of a system. This was reinforced by incidents such as Three Mile Island, where human errors, exacerbated by design deficiencies, led to the most serious incident yet experienced by the nuclear industry in the U.S.

One of the major problems encountered in human reliability assessment is the special difficulty of obtaining data on human errors for use in PRA. In the case of hardware components, such as valves, it is relatively easy to observe how many mechanical failures occur compared with the number of successful operations. The frequency of failures divided by the total number of operations can then be used to estimate the probability of failure.

In the case of human actions, the situation is considerably more complicated. Blame and guilt tend to be associated with errors, and therefore many erroneous actions are not reported because of the likelihood of punitive actions against the individual operator. In addition, many errors are due to "cognitive malfunctions" such as inappropriate decision making or a misperception of the nature of a situation. Thus, a failure to operate a valve is an external "error mode" which could have arisen from a variety of cognitive malfunctions such as a failure to understand which valve was to be operated, or confusing the situation with another similar situation. Clearly, it is not possible to directly observe the number of times such internal "cognitive malfunctions" occur, and therefore it is almost impossible to collect numerical data on these events. Although some numerical data on the probability of human errors is available, these tend to be confined to fairly simple, easily observable actions obtained either from production line situations or laboratory experiments. There are considerable problems associated with extrapolating such data to the very different environment of a nuclear power plant.

These problems can be overcome to a large extent by the use of techniques which utilize expert judgment. The rationale underlying such approaches is that, using experienced judges, it is possible to elicit estimates of the ways in which the probability of error is likely to be affected by factors such as the operators level of training, time available to carry out the required action, the existence of good quality procedures, etc. If this information can be used to derive failure probabilities for individual human actions, then the data problem is considerably reduced. The question of the validity of this approach, in terms of the degree to which it generates similar probability estimates as field data on the same human errors collected from a real plant, is discussed in detail in Volume II, Section 1.10 of this report.

SLIM, the expert judgment methodology which was refined and developed further in this study, is an extension of previous work on the problem (Embrey, 1983a) NUREG/CR-2986. During that earlier work, the basic form of SLIM was developed and a limited pilot experiment carried out to test the approach. Phase I of the SLIM research program concerned itself with testing some of the underlying assumptions of the method. Phase II was devoted to carrying out a field study to determine the applicability of SLIM to real and representative nuclear power plant critical scenarios and to evaluate the reactions of potential users of the technique. Phase III, representing a major proportion of the work reported here, was the development of a computer based implementation of the original SLIM methodology using an interactive program called MAUD*. SLIM will be described in detail in subsequent sections, together with a plan for the testing and validation of the approach by potential users.

3. THE SUCCESS LIKELIHOOD INDEX METHODOLOGY (SLIM)

A detailed technical description of SLIM is available in a number of publications, e.g., Embrey (1983a,b,c), and an explanation of the theory underlying SLIM is presented in Volume II, Section 1.6 of this report. In this section, the original form of the approach will be described, together with the procedures for carrying out a SLIM assessment. The procedure is generally carried out using multiple judges (either working alone or together in a group), in order to take into account a range of experience and to reduce biases which may be present within individual judgments.

The basic rationale underlying SLIM is that the likelihood of an error occuring in a particular situation depends on the combined effects of a relatively small set of performance shaping factors (PSFs). In brief, PSFs include both human traits and conditions of the work setting that likely influence an individual's performance. Examples of human traits that "shape" performance might include the competence of an operator (as determined by training and experience), his/her morale and motivation, etc. Conditions of the work setting affecting performance might include the time available to complete a task, task performance aids, etc. It is assumed that an expert judge (or judges) is able to assess the relative importance (or weight) of each PSF with regard to its effect on reliability in the task being evaluated. It is also assumed that, independent of the assessment of relative importance, the judge(s) can make a numerical rating of how good or how bad the PSFs are in the task under consideration, (e.g., achieving recirculation in a pressurized water reactor [PWR] loss-of-coolant accident [LOCA]) where "good" or "bad" mean that the PSFs will either enhance or degrade reliability.

Having obtained the relative importance weights and ratings, these are multiplied together for each PSF and the resulting products are then summed to give the Success Likelihood Index (SLI). The SLI is a quantity which represents the overall belief of the judge(s) regarding the positive or negative effects of the PSFs on the likelihood of success for the task under consideration. If we can assume that as a result of their knowledge and experience the judge(s) have a correct idea of the effects of the PSFs on the likelihood of success, then we would expect the SLI to be related to the probability of success that would be observed in the long run in the situation of interest (i.e., the actuarially determined probability).

A major assumption of the SLIM approach is that a SLI generated by this process bears a consistent relationship to the expected long-term probability

*MAUD (Multi-Attribute Utility Decomposition) is a stand alone interactive software package running under the CP/M operating system, which aids the user in assessing alternatives. MAUD is proprietary to the Decision Analysis Unit, London School of Economics, and was made available for this study through a non-exclusive end user license.

of success and can be converted to it in a simple manner. Experimental evidence suggests that the SLI is related to the logarithm of the probability of success for a task. Pontecorvo (1965) showed a logarithmic relationship between an index similar to the SLI and the log probability of success for maintenance tasks. Hunns (1982) provides an intuitive argument giving support to the notion of a logarithmic relationship in this context. The main justification for the use of a logarithmic relationship is, however, empirical rather than theoretical. Thus, support for a logarithmic relationship, or any other consistent relationship assumed within SLIM, must come from actual data. There are also practical advantages in using a logarithmic relationship because of the wide range of magnitudes of human error probabilities (HEPs) (1 to 10^{-5}) which need to be considered.

The logarithmic relationship between expert judgments and success probabilities can be expressed with the following calibration equation:

log of the success probability = a SLI + b

where:

a and b are empirically derived constants.

In order to produce an empirical calibration relationship between the SLI and the log of the success probability, at least two tasks must be available for which the probability of success is known, in the task set being evaluated. If this is the case, the constants a and b in the above equation can be evaluated and the calibration equation can then be used to transform any SLI value produced by the judge(s) into a log probability of success for the task. The log probability of success is readily convertable into the probability of success. An estimate of the HEP or likelihood of task failure, the ultimate goal of SLIM, is found by simply subtracting the success probability from one.

3.1 An Example of the SLI Procedure

The concepts described above can best be illustrated by a simple worked example. This section will also provide a detailed description of the practical application of SLIM. Suppose one desired to evaluate the probability that an operator will correctly diagnose the state of a nuclear power plant, and initiate manual intervention when a failure occurs in an emergency feedwater pump during a transient. The following steps would be carried out.

3.1.1 Step 1: Modeling and Specification of PSFs

During this first step, the judges thoroughly discuss the task to be evaluated, with particular attention being paid to identifying the various ways in which errors of omission and commission could occur (error modes) and the PSFs which could impact on these error modes. The various forms of task analysis which are available may be employed here, together with documentation of emergency operating procedures, photographs of the control room, etc. Operator input is very important at this stage. The modeling should be as exhaustive as possible and the results documented to indicate the error modes tnat the judges have in mind when making their assessment. This is necessary so that the procedure can be subsequently audited if required. The documentation could be in the form of a fault tree, the technique used to represent failure modes in the hardware assessment aspects of PRA, or some other form of representation. At the end of the modeling phase, all credible error mode: will have been considered and the PSFs which have a significant effect on these errors will have been identified.

In our example, we will assume that the judges have decided that the following PSFs are the major factors influencing success in the task being evaluated:

- Quality of the information available to the operator from the control panel.

- Quality of the procedures.

- Time available to diagnose the situation and to carry out the appropriate actions.

- Degree of operator training.

The documentation for this phase should include some description of exactly what is meant by each of these PSFs as used in the modeling session. The process of documenting the sessions will be monitored by the facilitator, the individual who leads the exercise.

3.1.2 Step 2: Weighting the PSFs

The determination of the relative importance of the PSFs can be accomplished by several procedures. In the initial feasibility study, Embrey (1983a), the simple multiattribute rating technique (SMART) (Edwards, 1977) was used to estimate weights. A variant of this technique was used in the phase I evaluation experiment to be described shortly. In this particular variant of SMART, judges are first asked to consider the task being assessed and to visualize a situation where all the PSFs are as bad as they could credibly be in a real plant. They are then asked to decide which single PSF would have the most significant effect on enhancing the probability of success if it were improved. This is assigned a weight of 100. The PSF which would have the next most significant effect on succcss is then chosen and a weight is assigned to it relative to the most significant PSF. Thus, if the second PSF were judged to be half as important as the first in terms of its effect on success likelihood, it would be given a weight of 50. This process is then repeated for all the PSFs. The results for our example might be as shown in Table 3.1.

The normalized weights are obtained by dividing each individual weight by the sum of the weights. The normalized weights sum to one and represent the relative importance of each PSF in terms of how strongly it influences the likelihood of success.

Table 3.1 PSF Weights.

PSF	Assigned Weight	Normalized Weight
Quality of information	100	100/200 = 0.50
Training	50	50/200 = 0.25
Time available	30	30/200 = 0.15
Procedures	20	20/200 = 0.10
	$\Sigma = 200$	$\Sigma = 1.00$

3.1.3 Step 3: Rating the Task

The rating procedure is carried out next with the judge(s) directly assigning a numerical value to each PSF on a scale of 0-100,* where zero indicates that the PSF is as poor as is credibly likely, and where 100 indicates that it is as good as is credibly likely in a real plant, in terms of its effect on the likelihood of success. At this point, it is important to differentiate between importance weights and ratings. The ratings are independent of the weights assigned to the PSFs. The weights indicate the relative importance of the PSFs in terms of their overall effect on the success likelihood, and are, therefore, not independent of one another. The ratings essentially represent the experts opinions' regarding the actual situation in the nuclear power plant for the task being assessed. The rating assigned to each PSF is independent of all the others in the set of PSFs being assessed.

In our example, we will assume that the following ratings have been assigned: quality of information, 70; training, 20; time available, 10; procedures, 50. These ratings might arise from the following situation: The operator has a wide variety of information available which is much better than average, but not as good as in the best plants. The operator's training for this particular situation is inadequate and the time available to perform the action is so short that it will negatively impact on the likelihood of success. The procedures are about average for the nuclear industry.

3.1.4 Step 4: Calculation of SLIs

The calculation procedure for each SLI is shown in Table 3.2 below. As can be seen from Table 3.2 the process of calculating each SLI involves simply forming the products of the normalized weights and the ratings for each PSF and then summing the results. The SLI can range from 0 to 100, where 0 indicates that the task has a high probability of failing, and 100 where it has a high probability of success. The SLI of 46.5 in this example indicates that the task has a slightly less than average likelihood of success.

*Other rating scale ranges can be used, depending upon the assumptions of particular applications. For example, the MAUD-based implementation of SLIM uses a scale of 0-1. The range of possible SLI values is the same as the range of the rating scale.

Table 3.2 Calculation of the SLI.

PSF	Normalized Weight (From) Table 3.1)	Rating	Product Weight x Rating
Quality of information	0.50	70	35.0
Training	0.25	20	5.0
Time available	0.15	10	1.5
Procedures	0.10	50	5.0
	Σ = 1.00		SLI = Σ = 46.5

3.1.5 Step 5: Conversion of the SLI to Probabilities

Transforming the SLI to a probability estimate can be achieved by several procedures. These various procedures are discussed in detail in Volume II, Section 1.10 of this report. For the purpose of this example, the approach employed requires the availability of at least two tasks for which the probabilities of success (or failure) are known. In this case, let the tasks be Task A, with a known failure probability of 10^{-3} (0.001) and Task B, with a failure probability of 10^{-2} (0.01). The success probability is 1 minus the failure probability. This means the success probabilities for Tasks A and B are 0.999 and 0.99, respectively. Assume that the judges assigned SLI values of 80 to Task A and 20 to Task B, using the same procedure as has been outlined for the original task.

These values are substituted into the calibration equation given earlier, i.e.:

log of the Probability of Success = a SLI + b

This produces two simultaneous equations which can be solved as follows:

$$
\begin{array}{lll}
\text{Solution for a:} & \text{Task A} & \log(.999) = a80 + b \\
 & \text{Task B} & \underline{\log\ (.99) = a20 + b} \\
 & & -.000434 = a80 + b \\
 & & \underline{-.004365 = a20 + b} \\
 & & .00393 = a60 \\
 & & a = .0000655 \\
 & & \\
\text{Solution for b:} & \text{Substituting} & -.000434 = (.0000655)\ 80 + b \\
 & & -.000434 = .0054 + b \\
 & & = .00567
\end{array}
$$

Substituting the values obtained for the constants a and b back into the original equation gives:

log of the Probability of Success = 0.000065 SLI - 0.0057.

This is a general calibration equation for the group of tasks evaluated by this particular set of judges. We can therefore substitute the SLI value of 46.5 into this equation to obtain the probability of success for the specific task in our example. If we do this, we obtain a log probability of success of -0.002654. This is equivalent to a success probability of about 0.994, i.e., a failure probability of 0.006. In other words, the operator might be expected to fail to correctly diagnose the situation and perform the appropriate actions on about six occasions out of every thousand times that this action is required. As a rough check on this result, one would expect an SLI of 50 which lies half way between the reference task SLIs of 20 and 80, to correspond to a failure probability of 0.005, which is half way between the reference event failure probabilities of 0.01 and 0.001. The calculated failure probability for an SLI of 46.5 is 0.006. In other words, the failure probability for an SLI of 46.5 is slightly worse (6 per thousand attempts) as compared to that for an SLI of 50 (5 per thousand attempts).

In this example, the SLIs were converted to probabilities by using two tasks (A and 3) for which the probabilities were assumed to be known. However, reference tasks with known probabilities may not always be available. In such instances, it may be necessary for judges to make absolute probability estimates for two tasks which would then serve as reference tasks for converting the SLIs to probabilities. One procedure would have judges make absolute probability estimates of two of the original tasks in the set being assessed. Another procedure would have judges consider two hypothetical situations where all PSFs are first as good as they could credibly be and second where they are as bad as they could credibly be in a real plant. The first situation would be assigned an SLI of 100 and the second situation and SLI of zero. The latter procedure was the one taken in the Phase II research, the field study which is described in Section 5.

3.1.6 Step 6: Calculation of Uncertainty Bounds

The measurement of any phenomenon always involves a certain amount of error, or degree of uncertainty. In a PRA context, it is necessary to have a measure of the uncertainty of the failure probabilities included in fault-trees so that upper and lower bounds can be calculated for the overall system reliability. This requirement also applies to HEP estimates; therefore, a method for generating uncertainty bounds around the point estimates produced by SLIM is needed.

There are several approaches to generating uncertainty bounds around SLIM produced HEPs. In the first of these procedures, judges are asked to make a direct estimate of the upper and lower bounds for each HEP estimate produced by SLIM. Seaver and Stillwell (1983) suggest the use of a logarithmic probability/odds scale (see Figure 3.1), together with a question such as:

> "For this event, what are the upper and lower bounds of the HEP that make you 95% certain that the true HEP falls between these bounds?"

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:

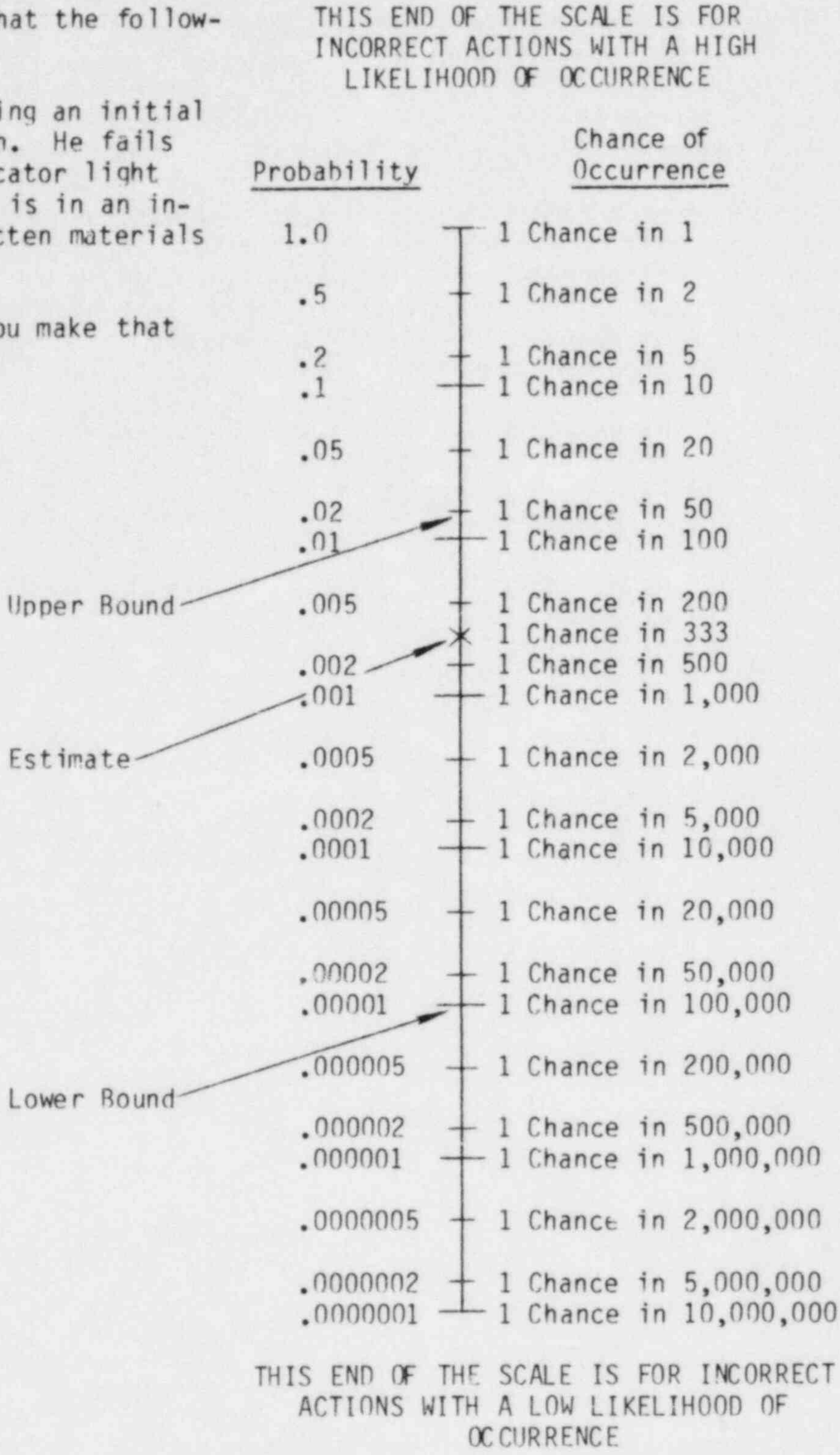

Figure 3.1 Logarithmic probability odds scale for obtaining direct estimates of upper and lower bounds of SLIM produced HEP estimates.

When SLIM is conducted as a consensus process, the uncertainty bounds should be arrived at consensually. When judges independently estimate HEPs with SLIM, uncertainty bounds can also be estimated on an individual basis. Aggregating both the HEPs and the uncertainty bounds is accomplished by taking the geometric mean of the estimated values.

Statistical estimation of uncertainty bounds is a straightforward application of statistical theory to the problem of estimating probabilities. Confidence limits, or error bounds in this application, are placed around HEP estimates on the basis of the standard deviation computed from the variability in HEP estimates by the individual judges. Specific procedures for accomplishing this are discussed in detail in Seaver and Stillwell (1983).

In many instances, upper and lower uncertainty bounds will be available for the calibration tasks used to solve the logarithmic calibration equation presumed to underlie SLIM. If these bounds are available, they can be used to derive calibration equations for generating uncertainty bounds for all tasks being assessed.

3.1.7 Summary

It can be seen that SLIM is a systematic method for positioning the likelihood of success of a task on a scale as a function of the differing conditions influencing the successful completion of the tasks. The absolute probability of success for tasks placed on this scale can be determined by calibrating the scale with reference tasks.

4. PHASE I RESEARCH - EXPERIMENTAL EVALUATION OF SLIM

In a previous investigation a preliminary pilot experiment to evaluate the feasibility of SLIM was carried out (Embrey, 1983a). In that experiment, the SLI methodology was applied to the evaluation of six human factors experimental tasks (not directly related to nuclear power plants) for which known failure probabilities were available. The results indicated a significant degree of correlation between the log probability of success and the SLI (r = .98), suggesting that the assumed calibration equation linking these quantities (i.e., log of the Probability of Success = a SLI + b) was supported.

The first objective of Phase I research was to extend the earlier pilot experiment to provide a more realistic evaluation of SLIM with a wider range of task types and expert judges. Detailed descriptions of this experiment are provided in Volume II, Section 2.1. Twenty-one tasks were utilized for which probabilities of failure were known. The 21 tasks, presented in detail in Volume II Appendix A, were chosen such that they formed three groups of seven tasks each; the three groups broadly correspond to the three categories of tasks described in a classification scheme developed by Jens Rasmussen of Risø National Laboratory in Denmark (Rasmussen, 1981). This classification scheme comprises three general types of categories--skill, rule, and knowledge based behavior.

- Skill based behavior - occurs when an individual is responding directly to some initiating event without having to think about his or her response in detail, or refer to a set of procedures. Examples of such behavior would be a driver braking heavily to avoid a sudden collision or an operator immediately silencing a "nuisance alarm" which was constantly sounding in a control room.

- Rule based behavior - involves the individual following a set of rules or procedures to achieve a goal. In a nuclear power plant, an example would be the calibration of the Nuclear Instrumentation System, or following an Emergency Operating Procedure after a particular incident had been diagnosed.

- Knowledge based behavior - is required when the operator is in an unfamiliar situation for which no defined procedures exist and therefore diagnosis, problem solving, and the formulation of a strategy may be necessary.

The reason for applying this classification scheme was to group together tasks which could be expected to be influenced in similar ways by the PSFs being considered. In other words, all tasks within a category were expected to have common relative weights associated with the PSFs.

In addition to the requirement that the 21 tasks could be classified into the above three groups, two other criteria were applied. The first of these was realism, in the sense that the tasks should be either collected from field situations in the process and power industries or should be realistic simulations of these situations. This criterion proved extremely difficult to meet, especially with regard to the knowledge based category of tasks. Data from laboratory experiments utilizing problem solving tasks similar to those in the process industries were therefore used. As is usual in the human reliability field, data from real plant situations were virtually unobtainable from the open literature and the only human error data available from a chemical process industry was confidential information. This information was used in the experiment although its origins were concealed.

The other criterion applied was that the tasks should encompass as broad a range of probabilities as possible. This also proved a difficult criterion to meet. As might be expected, virtually all of the probabilities of failure were in the high to medium (10^{-1} to 10^{-3}) range because of difficulties of collecting data on rare errors. In addition, most laboratory tasks are designed to produce fairly high failure rates in order to obtain sufficient data from reasonably sized experiments. The lowest failure probability employed (5×10^{-5}) came from an industrial assembly operation (omission of a soldered joint) and this was two orders of magnitude lower than the next group of probabilities. Although this task was hardly typical of those encountered in PRAs of nuclear power plants, it was included to provide an "extreme" probability.

The judges who participated in the exercise were four reliability analysts, two individuals with nuclear operating experience, and two human factors specialists--eight judges in all. They were provided with comprehensive descriptions of six PSFs to be used in the assessment. The PSFs were as follows (detailed definitions are available in Volume II, Appendix B):

1. Relevance and comprehensiveness of training
2. Time available to perform task
3. Motivation
4. Information available
5. Quality of procedures
6. Degree of checking and supervision.

As described in Section 3, the typical procedure for obtaining the SLIs first involves obtaining importance weights for each PSF for each of the 21 tasks being evaluated, i.e., six PSF weighcs for each task. The importance weights are then normalized by dividing each weight by the sum of the weights on a task-by-task basis. Next, each PSF is rated for quality on a task-by-task basis on a scale ranging from 0 to 100. The SLI is then obtained for each of the 21 tasks by summing the products of the normalized importance weights and the quality ratings of each PSF on a task-by-task basis. An overall SLI for each task can then be obtained by taking the mean of the individual judge's SLIs.

Because of time constraints, it was not possible to calculate the PSF importance weights for each of the 21 tasks on a task-by-task basis. Instead, the judges were asked to consider each of the three broad categories of tasks (skill, rule, and knowledge based) and to assume that it was reasonable to assign one set of PSF importance weights to each category. In effect, this meant that all seven tasks falling within a given category received the same PSF importance weights. The judges were then asked to give PSF importance weights for each task category--skill, rule, or knowledge-based. This resulted in three sets of normalized PSF importance weights (i.e., six weights for each of the three categories). For discussion purposes, it is convenient to refer to these category specific importance weights as "generic weights."

To compute the SLIs, the generic weights were multiplied by the PSF quality ratings for each task within the particular category to which the tasks belonged. The SLI for each task was calculated by summing together these products, the typical procedure for combining weights ana ratings into SLI.

4.1 Results

The first part of the analysis involved a comparison between the logs of the HEPs for the 21 tasks and the corresponding median (a form of average) of the SLI values for each task, averaged across the judges. If these two quantities are plotted against one another, the extent to which the resulting points fall along a straight line (when plotted on a log scale) indicates the degree to which the underlying logarithmic assumption of SLIM is supported.

The goodness of fit between the plotted data and the straight line, indicating the strength of association between the two sets of data, can be measured with the correlation coefficient r. If all the data points fall exactly on the straight line, r will be either 1.0 or -1.0 depending on whether the relationship is positive or negative. If the points are randomly scattered, indicating no relationship whatsoever between the two data sets, r will be zero. The higher the r value, the greater the certainty in predicting the data points of one set from the second set.

Figure 4.1 shows the plot of the log HEPs and the median SLI values, with a superimposed best fitting straight line* drawn through the data. As can be seen from the figure, the points are generally scattered, tending not to cluster near the straight line. The correlation coefficient which measures the degree of association between the log HEPs and the SLIs was -.47 (the negative sign occurs because as the SLI increases the log failure probability decreases). This correlation is shown by statistical tests to be no stronger than would be expected to occur by chance. Thus, at first, the results seem to give no strong support to the assumed logarithmic relationship between the probability of success and the SLI.

There are several plausible explanations for this result. The assumed logarithmic calibration equation for converting SLIs to HEPs may be incorrect. Or, there may be sources of variation in the weights or ratings which have not been explicitly considered, and these may have attenuated the results.

Because of the positive support for SLIM in the earlier pilot experiment (Embrey, 1983a), it seemed premature to dismiss the methodology on the basis of the weaker results reported above. Instead, other reasons for these results were investigated. A likely explanation lies with the classification system that was used. If, as is quite feasible, the tasks were assigned to incorrect categories, then the generic weights applied to all tasks within a category for calculating the SLIs would also be incorrect. This would have the effect of adding random error to the SLIs with the consequence that any consistent relationship between the SLIs and the log HEPs would be attenuated. In fact, much difficulty was encountered in assigning tasks to appropriate categories during the design phase of the experiment. The 21 tasks assessed came from a very wide variety of laboratory and production situations. In contrast, the Rasmussen classification scheme had been developed for a far more limited range of situations--primarily for nuclear or chemical process control settings.

It was therefore decided not to utilize the PSF weights and instead to calculate the SLI values assuming that equal weights applied to all the PSFs, i.e., all the PSFs were equally important in affecting the success likelihood. The SLIs were recalculated on this basis. The correlation coefficient between the log HEPs and the new SLI values, obtained using the equal weights assumption, was calculated. The correlation coefficient now became -0.60, a highly significant result which has a probability of less than 5 in 1,000 of

*Determined by the method of ordinary least squares (OLS).

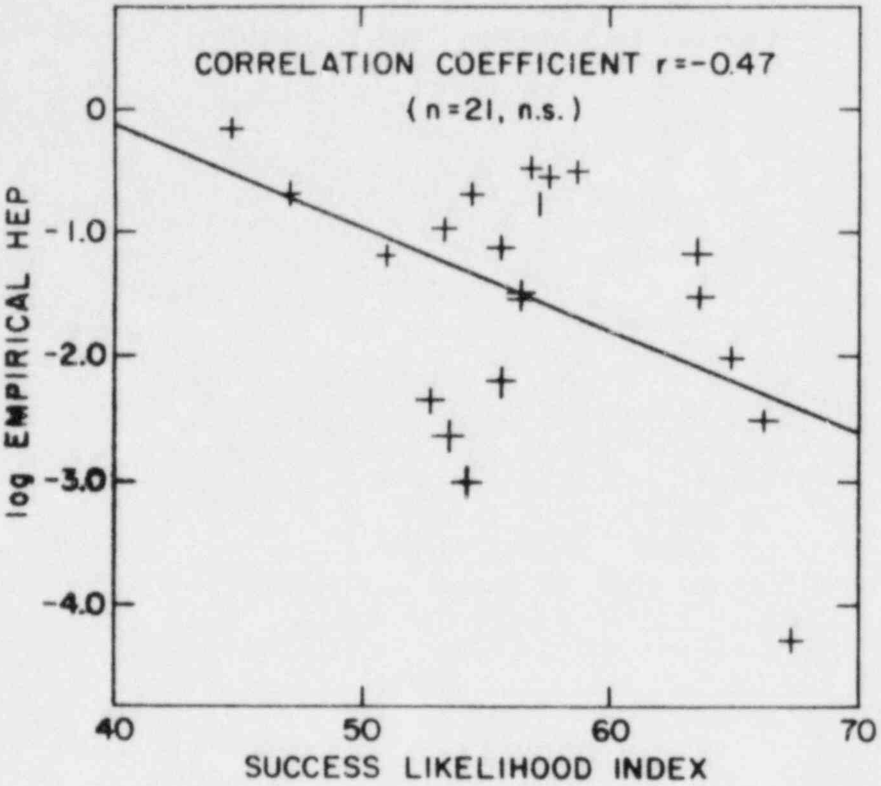

Figure 4.1 Success likelihood index (calculated using generic weights) vs. log empirical human error probabilities.

occurring by chance. The SLIs were then plotted against the log HEPs as shown in Figure 4.2 and, it can be seen, the points now fall much closer to the straight line than in Figure 4.1.

Three of the tasks were then removed from the analysis on the basis of a content analysis (see Volume II, Appendix A) which indicated that they contained insufficient information to allow a proper evaluation of the SLIs. The remaining 18 SLIs were plotted against the corresponding log failure probabilities as shown in Figure 4.3. The correlation coefficient increased to -0.71. Statistical tests again indicate that this has a very low probability of being a chance relationship (less than 2 in 1,000). These results lend support to the log relationship assumed to underlie SLIM. The fact that importance weights were not used in this experiment does not mean that they would not be employed in other applications of the SLIM technique. Whether the use of individual importance weights derived for each task would have produced a higher correlation coefficient than the -0.71 obtained in the present experiment remains an open question, since this information was not collected. However, if the judges actually possess some prior knowledge regarding the relative impact of the PSFs on likelihood of success, combining this information with the ratings should always produce better results than if equal weights are assumed. The results obtained in this experiment are therefore conservative, in that they assume that the judges possess no prior information on the relative importance of the PSF. As reported above, in the initial work on developing SLIM (Embrey, 1983a), a correlation coefficient of 0.98 was obtained when the judges derived both importance weights and ratings for each of the tasks they were assessing.

The support for the log relationship between the probability of error and the SLI is important because it provides the justification for converting the SLI values to probabilities via the calibration equation derived from reference tasks. It should be emphasized, however, that other relationships between HEPs and SLIs are possible. The goal of the research reported here is not necessarily to establish the logarithmic relationship as being superior to any other relationship on theoretical grounds. Rather, the intention is to provide empirical support for a calibration equation which can be used pragmatically to derive HEPs from SLIs in PRA work. The validity of SLIM does not stand or fall on the basis of whether a particular calibration relationship is the "correct" one, but on the basis of the consistency in the relationship between the SLIs and HEPs. The generality of the logarithmic relationship can only be established by further research.

4.2 Discussion

A number of useful findings emerged from the study. Perhaps the most important of these is that the assignment of generic weights to broad groups of tasks is only appropriate when an adequate task classification scheme is available. Otherwise, task specific weights should be derived when using SLIM. Problems also arose when the judges attempted to use the predefined PSFs for all the tasks in the study. Although the PSFs were very applicable

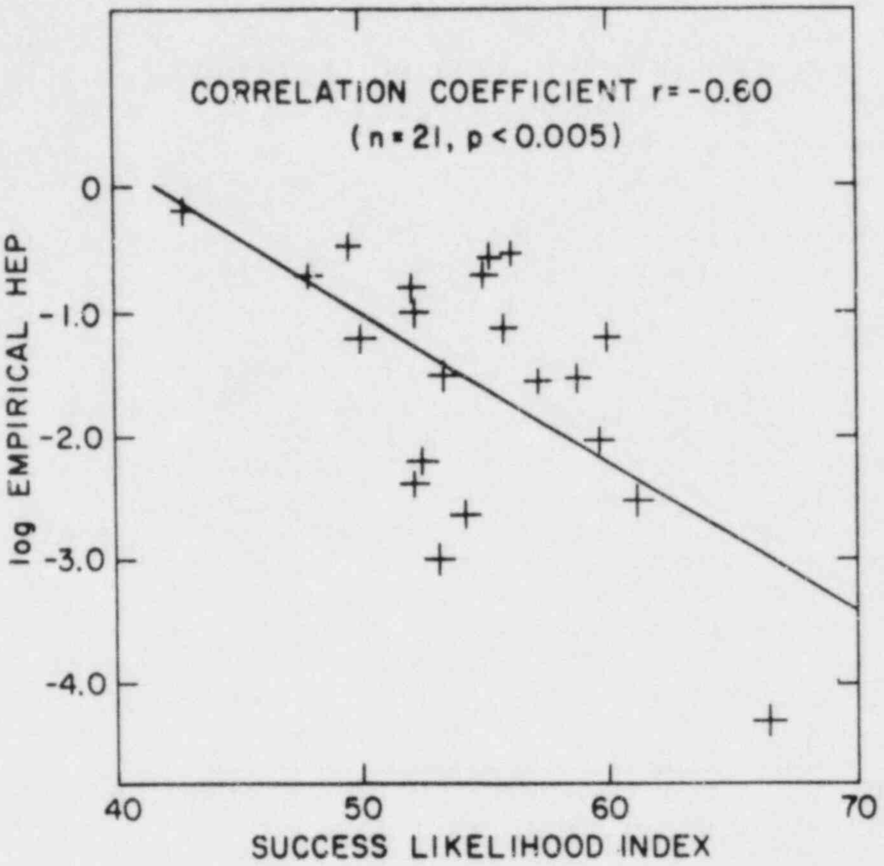

Figure 4.2 Success likelihood index (calculated using equal weights) vs. log empirical human error probabilities.

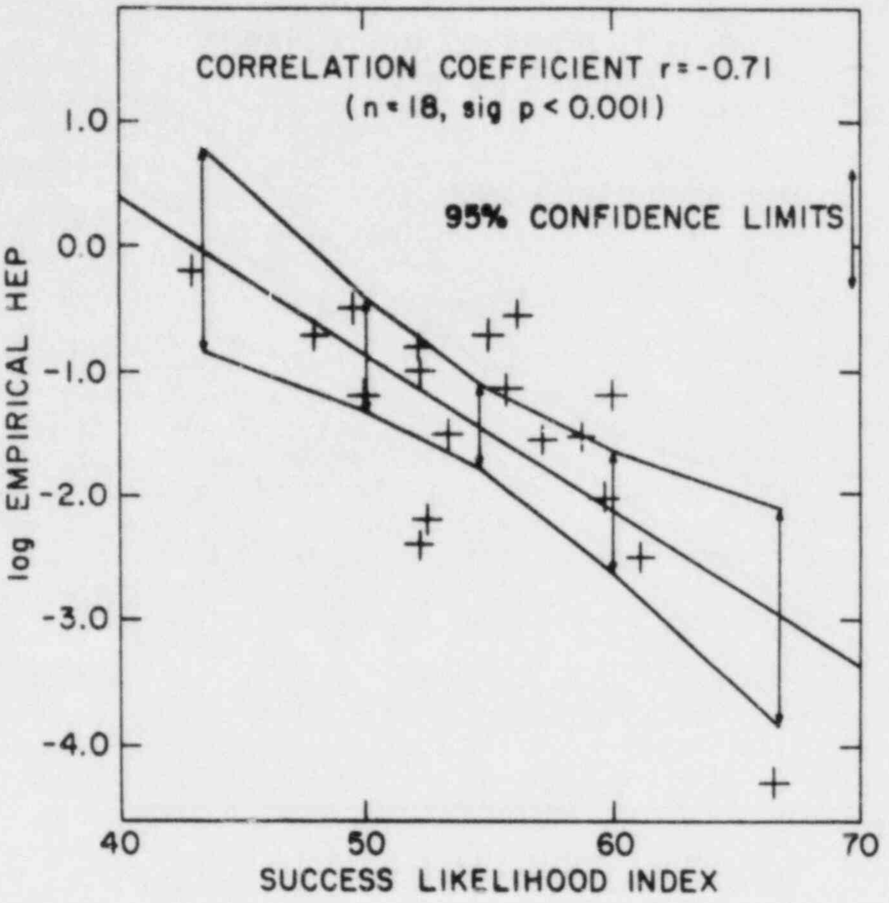

Figure 4.3 Success likelihood index (calculated using equal weights) vs. log empirical human error probabilities, based upon the removal of three tasks from the analysis.

to the industrial (process control and nuclear) tasks, they were less appropriate for the laboratory tasks. These results underscore the importance in developing an adequate task classification scheme or taxonomy.

Another finding was that in many cases it was impossible to provide much of the information required by the judges to carry out their assessments. Even published studies were remarkably lacking in the detailed information required to adequately weight and rate the PSF. This finding emphasized the need for specific information on the situation being assessed. This suggests that when SLIM is being used to assess nuclear power plants it will be important to include an individual on the assessment team with plant and preferably site specific knowledge to provide the necessary detailed information.

The difficulties experienced in collecting the original data for this experiment re-emphasize problems endemic to work in the human reliability area. The apparently modest requirement to collect human error data on 21 skill, rule, and knowledge based tasks from the nuclear and process industries was impossible to achieve, particularly for the knowledge based category. Even published experimental work using nuclear training simulators yielded very little data which were usable for human reliability purposes. This almost total absence of first hand data emphasizes the difficulty of any human reliability evaluation approach based on a data bank concept.

The final conclusion to emerge from the study is the need to train assessors. The robustness of SLIM was demonstrated in that it was able to produce reasonably coherent results despite the fact that it was the judges' first encounter with the methodology and they had a relatively short time with which to familiarize themselves with the tasks to be assessed. It seems reasonable to assume that further training would have contributed to improved performance.

5. PHASE II RESEARCH - A FIELD STUDY OF SLIM

In addition to the experimental evaluation described in the preceding section, SLIM was also applied in a field setting to evaluate human reliability in degraded core scenarios of nuclear power plants. A quantification workshop was held during which time expert judges assessed five such scenarios. Using SLIM, the judges produced quantitative estimates of the probability of an operator failing to carry out eight critical actions for the five scenarios.

5.1 Judges Employed in the Study

The pool of 12 judges available for the study included PRA analysts, a human factors specialist, simulator instructors who had operational experience in some of the plants being evaluated, and a thermohydraulics expert. Either seven or eight of these judges were used to evaluate each of the eight human actions assessed. Because of scheduling difficulties, only three of the judges were common to all eight assessments, the remainder being drawn from the pool as available.

5.2 Procedure

During the first session, the set of PSFs that were to be used for the purposes of quantification, were distributed and discussed. These PSFs had been developed as a result of interviews with former plant operators, supervisors, and simulator instructors during earlier phases of the study. From the session discussions, definitions of the following seven PSFs were developed. (See Volume II, Appendix A for a detailed description of the PSFs.) The PSF definitions established the end points for each PSF scale in terms of the features of the worst licensable plant and of the best feasible plant.

1. Quality of design
2. Meaningfulness of procedures
3. Role of operations
4. Teams
5. Stress
6. Morale/motivation
7. Competence.

There was general agreement that the set of PSFs provided were comprehensive, and accurately represented the major influences on operator performance in the sequences to be considered. After the PSFs had been defined and discussed, the next step was to consider the sequences in detail. These discussions occupied a considerable proportion of the available time, but were necessary in order that all of the experts had a shared perception in terms of their understanding of the required operator actions, and the factors which impinged on the likelihood of these actions being achieved. SLIM was then exercised as described below.

The weighting and rating assessments were carried out independently by the judge(s) as described in Section 3. These were then reviewed by the group as a whole and in some cases the judge(s) modified their individual assessments. No attempt, however, was made to force consensus.

As the final step in the procedure, the judges were asked to make absolute judgments of the probability of failure for two "boundary conditions." These were for the best credible situation, i.e., all the PSFs being as good as they could credibly be in a real plant, and the worst case situation where the PSFs were as bad as could credibly occur in a licensed power plant, for each scenario considered. These judgments were made independently, and then discussed in order to reach a consensus. In most cases it was possible to agree on absolute probability estimates for the two boundary conditions. With three of the eight critical operator actions, however, no consensus was obtained, and in these cases alternative boundary values were retained to calculate separate values for the failure probabilities, as described below.

5.3 Calculations and Results

A typical set of data obtained from a SLIM session for one of the eight operator actions assessed is shown in Tables 5.1 and 5.2. The data presented pertains to action 1 in Scenario 1.

Table 5.1 Weights.

	PSFs						
Judge	Design	Proc.	Role Oper.	Teams	Stress	Morale	Competence
A	90	90	80	100	50	50	95
B	70	55	60	100	85	20	50
C	70	75	0	100	70	0	90
D	70	70	70	100	50	20	95
E	70	80	40	90	60	50	100
F	50	70	25	40	60	20	100
G	50	80	10	100	40	10	95
n = 7							

Table 5.2 Ratings.

	PSFs							Estimated Boundary Conditions	
Judge	Design	Proc.	Role Oper.	Teams	Stress	Morale	Comp.	Best	Worst
A	70	90	75	90	60	70	90	$5x10^{-4}$	$5x10^{-2}$
B	65	85	80	85	85	80	70	10^{-4}	10^{-2}
C	75	60	75	75	50	75	80	10^{-4}	10^{-2}
D	70	60	75	85	60	75	85	10^{-4}	10^{-2}
E	75	75	90	85	70	85	85	10^{-5}	10^{-2}
F	70	70	80	70	50	60	75	10^{-5}	10^{-2}
G	80	80	80	90	50	70	90	10^{-4}	10^{-1}
n = 7				Consensus Values for bounds				10^{-4}	10^{-2}

As described in more detail in Volume II, Section 2.2, two separate procedures were followed in calculating the human error probability (HEP) for each scenario. The first of these procedures involved the following three steps. First, an SLI was calculated for each judge on the basis of the elicited PSF weights and ratings. Then, the resultant SLIs were used together with the individual judge's boundary condition (best or worst) probability estimates to arrive at a log HEP value for each judge for each scenario. Finally, the geometric mean of all judges' log HEP values was calculated to give an estimate of the overall HEP for each scenario. The second procedure used the consensus boundary condition HEPs and the post consensus weights and ratings, where these differed from the pre-consensus values. A statistical test indicated that the HEPs calculated using these two slightly different approaches to aggregation were not significantly different.

Using the individual log HEPs for the three judges who participated in all the assessments, it was possible to carry out a number of further analyses of the data. The first area investigated was the degree of inter-judge consistency. Results from an analysis of variance statistical procedure (ANOVA) indicated that the degree of agreement among the judges approached statistical significance, and that most of the variability in the log HEPs could be attributed to the differences between scenarios, not between judges.

Statistical uncertainty bounds on the log HEPs were calculated using the method described by Seaver and Stillwell (1983). The average uncertainty (95% confidence limits) about the log HEP estimates was 1.04 log units. Only one estimate had an uncertainty of greater than one log unit.

Sensitivity analyses were carried out to investigate which PSFs were judged to have the greatest effect on the HEPs evaluated. These analyses indicated the relative importance of the PSFs as shown in Table 5.3 below.

Table 5.3 Comparison of PSF Importance.

PSF	Mean Weight	Weight Relative to Least Important PSF
Competence	93.80	3.00
Teams	86.91	2.75
Procedures	85.71	2.70
Design	68.47	2.10
Stress	58.24	1.80
Morale	35.80	1.10
Role of Operations	31.60	1.00

Detailed analyses of the statistically significant differences between PSFs are given in Volume II, Section 2.2.4. However, it is apparent that the three highest ranked PSFs (competence, teams, and procedures) are perceived to be considerably more important than the three lowest ranked PSFs (stress, morale, and operations); the average of the mean weights for the highest three is more than twice the average for the lowest three. In addition, the statistical analyses of the PSF weights showed that their relative importance was not significantly different between the critical actions evaluated. Thus, in this test of SLIM, it would have been permissable to use a generic set of weights for all the scenarios.

The final analysis conducted was for the rating data. This indicated that between plants there were significant differences in the way in which the PSFs were rated, indicating that some PSFs were perceived to be significantly worse than others in the plants examined. Significant differences in PSF ratings were also found between scenarios. This means that the plants considered were not identical in terms of the overall ratings obtained when all the PSFs were aggregated together. However, the rank-ordering of the ratings did not differ between plants; i.e., the PSF rated as most important for one plant was similarly rated for the other plants.

5.4 Discussion

In general, the field evaluation of the basic SLI methodology was successful in achieving several objectives. Although it was not possible to verify the accuracy of the human error estimates produced by SLIM because of the absence of sufficient field data on the rare event scenarios being evaluated, the judges involved in the exercise had considerable confidence in the results. It also seemed apparent that SLIM provided a useful structure which assisted the judges in modeling the potential failure modes. There was general agreement among the judges that SLIM possessed a high degree of face validity.

The detailed sensitivity analysis showed the relative impacts of the different PSFs on the overall probability of error for the various human actions evaluated. This information is especially useful to management since it can be used by designers and managers to reduce error probability in a cost-effective way.

6. RECOMMENDED PROCEDURES FOR USING SLIM

Section 3.1 presented an outline of the procedures to be followed in using SLIM. The procedures recommended there, however, were preliminary to the additional experience gained in the experiment and field study implementation of SLIM. Taking into account that additional experience, the sections that follow present the current recommendations for implementing the basic form of SLIM. These recommendations, it should be noted, do not apply to the MAUD-based version of SLIM described in Section 7.

6.1 Step 1: Modeling and Specification of PSFs

The fundamental question to be addressed in this phase of SLIM implementation is whether judges should originate the set of PSFs for the scenarios being assessed, or whether judges should be provided with a pre-defined set of PSFs. The preferred procedure is to provide judges with a set of pre-defined

PSFs. Following this procedure not only facilitates the assessment process, but also assists in orienting the judges to the tasks to be assessed.

The preferred way of arriving at a pre-defined set of PSFs is to conduct an in-depth pre-analysis of the specific plant and scenarios to be assessed with individuals having operating experience with the plant. Extensive discussions should be conducted with the aim of obtaining a consensus on the set of PSFs that are relevant to the plant and scenarios to be assessed.

Should an in-depth pre-analysis prove infeasible, the set of PSFs used in the field test of SLIM described in Section 5 (presented in detail in Volume II, Appendix A) can be used as a starting point in the elicitation process. That set of PSFs seems to be sufficiently generic to be applicable across a range of plants and scenarios.

Thus, the recommendation that judges should be provided with a set of pre-defined PSFs can be accomplished through a plant specific pre-analysis or by the use of generic PSFs. Regardless of the source of the pre-defined PSFs, however, it must be emphasized and made clear to judges that the set of PSFs provided them are not the only possible ones affecting human performance. Judges should be encouraged to modify the set (i.e., by adding relevant PSFs and deleting irrelevant ones) in the light of their own knowledge and experience.

6.2 Step 2: Weighting the PSFs

The approach described in Section 3.1.2 where judges first assign a weight of 100 to the most important PSF and then weight the remaining PSFs as a ratio of the most important one is still recommended. The additional step of having the group of judges discuss their individual PSFs to arrive a consensus weights, as described in Section 5.2, is also recommended. The consensus step is recommended to avoid the loss of information that occurs when mathematical aggregation procedures are followed.

6.3 Step 3: Rating the Tasks

As with the recommended weighting procedure, it is recommended that ratings from the individual judges are first obtained, which are subsequently discussed to arrive at consensus ratings.

6.4 Step 4: Calculation of the SLIs

SLIs are calculated by forming the products of the normalized weights and ratings for each PSF and then summing the results as shown in Section 3.1.4.

6.5 Step 5: Conversion of the SLIs to Probabilities

The recommended calibration procedure depends upon the availability of calibration tasks with known HEPs. If a sufficiently large number of tasks with known HEPs are available a regression approach can be followed. This would involve calculating a regression equation between the log of the known HEPs and the corresponding SLIs produced by the judges. The equation can then be used to convert the SLIs into HEPs for the tasks being assessed.

If there are only a few, but at least two, tasks with known HEPs, the procedure described in Section 3.1.5 can be used to convert the SLIs to HEPs. In particular, the known HEPs can be substituted into the logarithmic calibration equation to solve for the equations two unknowns (parameters). The equation can then be used to convert the SLIs on the assessed tasks into HEPs. A simple computer code to complete this task is given in Volume II, Section 3.5.

In instances where there are no tasks available with known HEPs, the procedure described in Section 5.2 will need to be followed. That procedure involves having the judges make absolute judgments of the HEPs for two boundary conditions, using the log probability/odds scale presented in Figure 3.1.

6.6 Step 6: Calculation of Uncertainty Bounds

Judgmental and statistical uncertainty bounds can be estimated by following the procedures described in Section 3.1.6. In the case of judgmental estimation, it is recommended that the consensus procedure be followed since this is consistent with the emphasis placed on a consensual process for implementing SLIM. If uncertainty bounds are available for the calibration tasks, the recommended procedure is to derive calibration equations from the upper and lower bounds and to use these equations to generate uncertainty bounds for the tasks being assessed.

6.7 Background of Judges

An ideal team of judges should include experts with operational experience in the specific plant and with the types of scenarios being assessed. Otherwise, preference should be given to judges who have experience as similar as is feasibly possible to the specific plant and with the types of scenarios being assessed. Other acceptable members of a team of judges would include PRA experts, thermohydraulic experts, training supervisors, and human factors specialists.

7. PHASE III RESEARCH - SLIM-MAUD: AN IMPLEMENTATION OF SLIM THROUGH THE USE OF MAUD

A major aspect of the work carried out during this study has been the implementation of SLIM using MAUD.* MAUD is derived from multiattribute utility

*MAUD is available from the Decision Analysis Unit, London School of Economics, Houghton Street, London, WC24 2AE. The Decision Analysis Unit version of MAUD can be run on a wide variety of microcomputers. To use MAUD, the minimal requirements of the microcomputer are 64K bytes of memory, and 2 floppy disk drives. The computer must also be able to operate under the CP/M operating system or under IBM/PC DOS. Details of how to obtain MAUD and to configure it to implement SLIM are described in Volume II, Sections 3.1 and 3.2.

theory, and is a flexible, interactive computer based system which has the capability of implementing a methodology like SLIM. Implementing SLIM through the use of MAUD represents a more sophisticated way of eliciting from judges the rating and weighting information utilized in the basic SLIM approach.

Furthermore, the elicitation procedures of MAUD are in closer accord with the theoretical assumptions underlying the SLI methodology than is the case for SMART, the procedure used in the previous development of SLIM, including the experiment and field test described above. The MAUD-based implementation has the additional advantage of being able to deal with the evaluation of up to 10 tasks in the same session. It also employs MAUD's built-in checks to monitor any dependencies between PSFs which may be present. The MAUD system is fully interactive and is sufficiently "user friendly" such that it can be used unsupervised by individuals or groups of judges with minimal training in computer-based techniques. An example of the dialogue used in SLIM-MAUD and a detailed technical description of the technique is presented in Chapter 3 of Volume II.

In a typical MAUD session, the system first asks the judge(s) to name the various tasks for which HEPs are required. It is assumed that the SLIs for all the tasks being assessed in a particular session can be determined by the same PSFs with the same relative weights. At least two reference tasks for which HEPs are available need to be included in the session for calibration purposes. SLIM-MAUD then elicits interactively the PSFs which are relevant in determining probability of success. MAUD performs a comprehensive set of consistency checks on the judges' use of these PSFs in assessing the courses of action under consideration. This process is repeated with the various combinations of tasks to generate a series of factors which are equivalent to the PSFs that are elicited directly in the basic SLIM technique.

With SLIM-MAUD, judges first rate tasks and then weight them, reversing the order used in the other SLIM elicitation technique. Thus, judge(s) are first asked to rate each of the tasks on nine-point scales, and define their "ideal" point on each scale, i.e., the rating scale value which would be optimal in promoting success. MAUD uses this information to re-scale the PSFs so that increasing scale values always indicate increasing likelihood of success.

The next step of SLIM-MAUD develops the PSF weights by comparing pairs of tasks which have different values on two of the PSF scales. SLIM-MAUD asks the judges which of the two tasks would be most likely to succeed, and then iteratively "degrades" one of the PSF ratings of the task most likely to succeed and improves one of those of the task less likely to succeed. This process is repeated until the judges' opinions reverse themselves with respect to which of the two tasks is most likely to succeed. By repeating this process for a range of PSFs, SLIM-MAUD is able to determine the relative weights of the various PSFs for the task set under consideration, as perceived by the judge(s).

A separate computer program (described in Volume II, Section 3.5) may then be used to convert the SLI values into probabilities using the calibration equation derived from the two reference tasks.

8. OUTLINE OF TEST PLAN

This section presents an outline of the Test Plan for the next phase of research. It ennumerates both key requirements for conducting a valid test of SLIM-MAUD and the issues underlying the criteria for assessing the method's utility. Chapter 4 of Volume II is devoted to a full description of The Test Plan.

8.1 Risk Analysis Tasks

The first requirement of a valid test of SLIM-MAUD is to select a representative set of risk analysis tasks which will encompass the whole range of situations which the technique may be required to evaluate. The Test Plan will be implemented by using all 15 level A tasks and 12 of the 20 level B tasks taken from the list of risk analysis tasks (35 in all) developed by the U.S. Nuclear Regulatory Commission (NRC) and Sandia National Laboratories (SNL) (SNL, 1983). A detailed list of the 35 risk analysis tasks is presented in Volume II, Appendix B of this report.

8.2 Task Classification Scheme

A second requirement is the need to develop a task classification scheme for tasks to be assessed with SLIM to ensure that tasks within each category being assessed are homogeneous. The classification scheme should be checked to be certain it is applicable across the variety of tasks likely to be evaluated in nuclear power plant PRAs and across the whole range of potential users of the approach. The Test Plan describes an approach based upon multi-dimensional scaling (MDS) for identifying homogeneous subsets of tasks separately within the level A and level B tasks sets (see Volume II, Section 4.5.1).

8.3 Subject Matter Experts

A third requirement is the need to determine the types of judge expertise most appropriate for SLIM assessments. In order to assess the effects of different types of judges using the technique, three categories of expert groups should be investigated, i.e., PRA specialists, operators or trainers, and a group comprising engineers and plant designers. A comprehensive assessment of the effects of judge expertise on HEP estimates would require two panels of six judges for each of the three expert groups.

Each of the individual judges would take all possible pairs of tasks and assess the degree of "relatedness" of the tasks on a simple 10-point scale. Relatedness in this context refers to the degree to which pairs of tasks are perceived to be similar in terms of their likelihood of success; this being inferred for those PSFs with similar profiles and similar relative importance.

These data should be analyzed using clustering and Multi-Dimensional Scaling (MDS) techniques (Kruskal and Wish, 1978). This will enable a comparison to be made within and between each type of expert group of individual judge's perceptions of task similarity (with regard to likelihood of success). MDS enables groups of tasks which are perceived to be similar to be grouped into clusters or categories. Thus, MDS analyses permit an investigation of the extent to which different expert groups perceive tasks to be clustered in the same way.

The clusters for the individual expert groups can be compared with the clusters which emerge when the MDS analysis is applied to the whole of the data across all the groups. The degree to which the task clusters for the individual groups correspond to the global clusters for all the judges, will indicate the feasibility of a classification structure that can be used to elicit judgments from all types of experts.

Assuming that such a generic classification structure emerges from the analysis, SLIM can be used to identify the PSFs associated with each category of tasks. This can be done using new sets of experts from each of the three groups described earlier, i.e., PRA specialists, operators or trainers and designer/plant engineers. A composite group containing individuals from all of these categories could also be employed where feasible. Each of the expert groups would carry out a SLIM assessment for each of the task categories. The differences between the PSFs assigned by the different expert groups would then be investigated. If the PSFs are sufficiently similar across groups, a task classification scheme or taxonomy will have been developed, with a set of PSFs associated with each task category.

This taxonomy will be of considerable value in carrying out SLIM-MAUD analyses. In particular, the reference tasks required for calibration can be associated with each task category.

Should it prove impractical to assemble the number and types of judges recommended above, a modified initial test of SLIM could be conducted with a group of judges having common expertise. All features of the Test Plan, except for an examination of the effects of different types of expertise, can then be implemented.

8.4 Criteria for Assessing the Utility of SLIM-MAUD

The utility of the MAUD-based implementation of SLIM can be assessed on the basis of three key criteria: practicality, acceptability, and usefulness. Practicality emphasizes the pragmatic concerns associated with any methodology, such as the required time and resources, and the degree of flexibility in applying the methodology in a wide variety of settings. Acceptability refers to the actual adoption of the methodology by users who are responsible for producing HEP estimates. The usefulness of a methodology can be determined on the basis of prevailing conventions of scientific standards.

The three criteria comprise a number of specific issues which can be rigorously addressed within the Test Plan. These are described in detail in Chapter 4 of Volume II and are summarized in Table 8.1.

Table 8.1 SLIM-MAUD Test Plan: Issues and Procedures.

Issues	Methods/Data	Analysis
<u>Practicality</u>:		
Cost	Actual costs incurred for implementating Test Plan.	Costs summation plus discussions of potential cost additions or reductions.
Subject Matter Experts	If feasible, by examining three expert groups: PRA specialists, operators or trainers, and engineers.	Multi-dimensional Scaling (MDS) of user responses.
Support Requirements	Ennumeration of equipment and other materials needed to implement Test Plan.	Discussion of equipment used and other equipment capable of using MAUD.
Transportability	Test will likely be implented in more than one location.	Experience in setting up and running SLIM-MAUD in seperate locations.
Expandability	Development of categorization scheme.	Cluster analysis of user responses.
Time Requirements	Actual experience gained in implementing Test Plan.	Discussion of experienced time considerations and factors affecting time.
Interface With Reliability Data Bank	Ensured by tasks to be evaluated.	None needed.
Implementability of Procedure	Use of more than one session facilitator.	Comparison of the degree of difficulty experienced by different facilitators.
<u>Acceptability</u>:		
Scientific Community	Professional journal submission.	Reviewer comments and/or acceptance of articles.
Expert Participants	Debriefing interview and survey.	Evaluation of interviews and analysis of survey data.
Potential Users	Informal survey.	Evaluation of responses.
Nuclear Regulatory Commission (NRC)	None.	None.
Nuclear Utilities	None.	None.
<u>Usefulness</u>:		
Reliability	Inter-judge consistency.	Use of MDS to assess consistency between individual results.
Face Validity	Survey of expert participants, informal survey of potential users.	Evaluation of open-ended comments and analysis of survey data.
Convergent Validity	Comparison with HEP estimates provided by other subjective techniques.	Examination of magnitude of differences.

9. CONCLUSIONS

The aim of this volume has been to provide a nontechnical overview of a research program devoted to the refinement and further development of SLIM, including the implementation of SLIM using MAUD. More detailed discussions of nearly all the areas covered in this overview are provided in Volume II of this report.

The results which have been obtained in the present study indicate that the use of SLIM is a viable approach to the evaluation of human reliability. The MAUD-based implementation of SLIM has a number of built-in features which facilitate a comprehensive assessment of the strengths and weaknesses of the SLI Methodology in practical applications. A recommended plan for testing the MAUD-based implementation is described in detail in Chapter 4 of Volume II. The Test Plan is applicable to any implementation procedure associated with SLIM. Because of its several clear advantages, however, the MAUD-based approach is the recommended implementation procedure in the Test Plan.

REFERENCES

Edwards, W., (1977), "How to Use Multi-Attribute Utility Measurement for Social Decision Making," IEEE Transactions on Systems, Man and Cybernetics, SMC-7,5.

Embrey, D. E., (1983a), "The Use of Performance Shaping Factors and Quantified Expert Judgment in the Evaluation of Human Reliability: An Initial Appraisal," NUREG/CR-2986, Brookhaven National Laboratory.

Embrey, D. E., (1983b), "Modelling and Quantifying Human Reliability in Abnormal Conditions," paper presented at the Fourth National Reliability Conference, Birmingham, England.

Embrey, D. E., (1983c), "The Quantification of Human Reliability Using Expert Judgment: Current Findings and Future Developments," paper presented at the Institute of Chemical Engineers Symposium: Human Reliability in the Process Control Centre, Manchester, England.

Hunns, D. M., (1982), "Discussions Around a Human Factors Data Base," High Risk Safety Technology, E. A. Green, Ed., John Wiley Press, London.

Kruskal, J. B. and Wish, M., (1978), "Multidimensional Scaling," Sage University Paper series on Quantitative Applications in the Social Sciences, 17-011, Beverly Hills and London:Sage Publications.

Pontecorvo, L. B., (1965), "A Method of Predicting Human Reliability," Annals of Reliability and Maintenance, Vol. IV, pp. 337-342.

Rasmussen, J., (1981), "Models of Mental Strategies in Process Plant Diagnosis," Human Detection of System Failures, J. Rasmussen and W. B. Rouse, Eds., Plenum Press, New York.

Sandia National Laboratories (October, 1983), "Data Collection and Analysis Test Plan for the Psychological Scaling Techniques Implementation and Evaluation Study," General Physics Corp and Maxima Corp.

Seaver, D. and Stillwell, W. G., (1983), "Procedures for Using Expert Judgment to Estimate Human Error Probabilities in Nuclear Power Plant Operations," NUREG/CR-2743, Sandia National Laboratories.

NRC FORM 335
(11-81)

U.S. NUCLEAR REGULATORY COMMISSION

BIBLIOGRAPHIC DATA SHEET

1. REPORT NUMBER *(Assigned by DDC)*
NUREG/CR-3518
BNL-NUREG-51716

4. TITLE AND SUBTITLE *(Add Volume No., if appropriate)*
SLIM-MAUD: An Approach to Assessing Human Error Probabilities Using Structured Expert Judgment, Volume I: Overview of SLIM-MAUD

2. *(Leave blank)*

3. RECIPIENT'S ACCESSION NO.

7. AUTHOR(S) D. E. Embrey, P. Humphreys, E. A. Rosa, B. Kirwan, and K. Rea

5. DATE REPORT COMPLETED

MONTH	YEAR
February	1984

9. PERFORMING ORGANIZATION NAME AND MAILING ADDRESS *(Include Zip Code)*

Brookhaven National Laboratory
Upton, NY 11973

DATE REPORT ISSUED

MONTH	YEAR
March	1984

6. *(Leave blank)*

8. *(Leave blank)*

12. SPONSORING ORGANIZATION NAME AND MAILING ADDRESS *(Include Zip Code)*

U.S. Nuclear Regulatory Commission
Human Factors and Safeguards Branch
Washington, DC 20555

10. PROJECT/TASK/WORK UNIT NO.

11. FIN NO.
A-3219

13. TYPE OF REPORT
Formal

PERIOD COVERED *(Inclusive dates)*

15. SUPPLEMENTARY NOTES

14. *(Leave blank)*

16. ABSTRACT *(200 words or less)*

This two-volume report presents the procedures and analyses performed in developing an approach for structuring expert judgments to estimate human error probabilities. Volume I presents an overview of work performed in developing the approach: SLIM-MAUD (Success Likelihood Index Methodology, implemented through the use of an interactive computer program called MAUD--Multi-Attribute Utility Decomposition). Volume II provides a more detailed analysis of the technical issues underlying the approach.

17. KEY WORDS AND DOCUMENT ANALYSIS

17a. DESCRIPTORS

subjective expert judgment
performance shaping factors
human factors

17b. IDENTIFIERS/OPEN ENDED TERMS

18. AVAILABILITY STATEMENT
UNLIMITED

19. SECURITY CLASS *(This report)*
UNCLASSIFIED

20. SECURITY CLASS *(This page)*

21. NO. OF PAGES

22. PRICE
$