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

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

VOLUME II: DETAILED ANALYSIS OF THE TECHNICAL ISSUES

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

Date Published — July 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. II
AN, RX

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

VOLUME II: DETAILED ANALYSIS OF THE TECHNICAL ISSUES

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 — May 1984
Date Published — July 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
CONTRACT NO. DE-AC02-76CH00016
FIN. NO. A-3219

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

TABLES

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 nonexclusive 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 the field studies.

1. TECHNICAL ANALYSIS OF THE SUCCESS LIKELIHOOD INDEX METHODOLOGY (SLIM) AND ITS IMPLEMENTATION THROUGH MULTI-ATTRIBUTE UTILITY DECOMPOSITION (MAUD)

1.1 Introduction

This two-volume 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 presented an overview of SLIM, discussed the results of the experiment and field test, the linking of SLIM to MAUD, and outlined a Test Plan for the next phase of research.

Volume II, presented here, provides detailed theoretical and technical information to supplement the non-technical overview given in Volume I. The goal of the four chapters comprising this volume is to demonstrate the theoretical and technical depth of the SLI methodology.

Chapter 1 of this volume begins with a discussion of the criteria which can be used to assess human reliability assessment (HRA) techniques. This is followed by a detailed description of the Success Likelihood Methodology (SLIM) including a number of important technical issues. These include calibration, the determination of uncertainty bounds and consistency, the type of judges to be used in SLIM assessments, training aspects, modeling capability, and generalization using the techniques. A detailed technical treatment of the theoretical structure underlying SLIM is then provided, followed by a description of the implementation of SLIM through the use of MAUD. Finally, evaluation of this implementation in terms of the criteria originally specified is discussed.

Chapter 2 provides a detailed description of Phase I (experimental evaluation) and Phase II (field study) of the SLIM research program, including the presentation of additional analyses from the data obtained.

Chapter 3 presents the stand-alone procedures for implementing SLIM through MAUD in the form of a detailed frame-by-frame example of a SLIM-MAUD session.

Finally, Chapter 4 presents a detailed plan for the next phase of research, phase IV, a test of the MAUD-based implementation of SLIM.

1.2 Assessment Criteria for Human Reliability Techniques

In this section key criteria will be discussed which can be used as a basis for comparing and assessing HRA techniques. After a technical description and analysis of SLIM and its implementation through MAUD, SLIM-MAUD will be evaluated in terms of these criteria in Section 1.12.

The criteria for the evaluation of HRA techniques can be divided into three broad categories: practicality, acceptability, and usefulness. These three criteria combine concerns over theoretical and technical considerations with issues concerned with the actual application of such methodologies. The last consideration reemphasizes the fact that no matter how good any method may be from a theoretical or technical standpoint, it may never be put into practice if it is not acceptable to potential users.

1.3 Practicality

The fundamental concern of this criterion is with the ease or difficulty of implementing a judgment technique with respect to resource needs and constraints.

1.3.1 Cost

This is probably the most important criterion within the practicality category. Cost in this context is basically the number of person-hours required on average to carry out an evaluation, plus the cost of any computer support needed to carry out the analysis, plus any other resource support requirements. A further consideration is the type of judge required. Obviously, the time of some individuals will be more expensive than others. For example, plant operators may be severely constrained in terms of their availability for long periods. As a rule of thumb, the resources required to carry out a human reliability assessment should not exceed those necessary for a hardware assessment of comparable complexity. Costs in terms of time will obviously be greater at first, because of the need to train users in the application of the technique.

....2 Training Requirements

It should be possible for nonspecialists to be trained in the use of the technique within a reasonable length of time. Since it is likely that prime users of a HRA technique will be engineers, it is important that specialist expertise, such as a comprehensive knowledge of human factors, which would require extensive training, will not be essential to exercise the technique. Techniques with a "built-in" training capability will be particularly attractive.

1.3.3 Breadth of Application

This criterion consists of several dimensions. The approach should be applicable to all types of tasks likely to be encountered in nuclear power applications. It should be able to deal with the full range of Skill-, Rule-, and Knowledge-based tasks as defined by the Rasmussen et al., (1981) taxonomy. These include control room operations, testing, maintenance, diagnostic, and decision making in ill-defined, unfamiliar situations, etc. The approach should also be applicable to human actions over the whole life cycle of the system. This will include design, construction, quality assurance, commissioning, operation, and decommissioning.

1.3.4 Data Requirements

Most human reliability techniques are limited by their requirements for data, usually in the form of error frequencies, together with the associated Performance Shaping Factors (PSFs), which can be collected from operational situations. In view of the difficulty of obtaining such data, techniques will be rated higher on this criterion to the extent to which they require fewer data of this type.

1.3.5 Capability to Consider Socio-technical and Organizational Factors

There is considerable evidence from interviews with operators and from reports of major incidents--e.g., the Presidential Commission Report (1979) on the Three Mile Island (TMI) accident--that factors such as organizational policies, management structures, etc. have a considerable impact on operator errors. Most human reliability assessment techniques are based on psychological theories of individual performance which do not consider such broad socio-technical factors. A technique which can incorporate these factors into the assessment process will therefore rate highly on this criterion.

1.3.6 Difficulty of Exercising Procedure

All other things being equal, preference is likely to be given to a technique which is relatively easy to use. With regard to judgmental techniques, ease of use will be determined primarily by the level of difficulty of the judgments that have to be made. This is partly influenced by the training of the judges, which has implications for the cost criterion discussed previously.

1.3.7 In-house Capability

Ideally, an organization should be able to use a technique with minimum inputs from outside specialists. In other words, the greater the "stand-alone" capability of a methodology, the more attractive it will be to a potential user.

1.4 User Acceptability

Any HRA methodology will be viable to the extent the technique is adopted and actually used by a full range of experts concerned with nuclear safety. In particular, the methodology should be acceptable to the U.S. Nuclear Regulatory Commission (NRC), the nuclear utilities, HRA and human factors experts, and the expert judges who actually use the methodology to arrive at human error estimates.

1.4.1 Scrutability

This criterion is related to the concept of face validity in psychology and the auditability criteria discussed in Section 1.5.2. If a HRA technique makes sense to the individuals who are using it, then it is much more likely to be acceptable than an approach which invokes obscure theories or complex mathematics. Individuals are more highly motivated to perform if they feel that they are engaged in a purposeful activity to achieve a clearly defined goal. In the HRA context, this means that it should be possible to explain in simple terms the meaning of the procedures that are necessary to exercise a technique.

1.4.2 Relationship of Technique to PRA Approaches and Techniques

It is essential that the procedures and outputs of the technique are compatible with the practices, expertise, and needs of the PRA community. There is a requirement to produce probabilistic data, preferably with uncertainty bounds, which can be readily applied in fault tree analyses and other existing PRA techniques.

1.5 Usefulness

Ultimately, the test of any methodology is its capability of producing coherent, reliable, and valid results. A methodology can be deemed to pass such a test if it withstands criticism among peers in the scientific community.

1.5.1 Accuracy and Validity

In a PRA context, the validity of a technique is likely to be judged by the praymatic criterion of the degree to which the human error probabilities (HEPs) predicted by the method concur with empirical probabilities estimated by frequency data obtained from "real" situations. This is usually referred to as "accuracy" in PRA, and is similar to the idea of empirical or external validity in psychology. A difficult problem arises here, because for many of the scenarios of interest in a PRA, the events beiny quantified are likely to be extremely infrequent. In fact, there will often be no previous occurrences of the event being assessed.

Because of this fact, the validity of any HRA technique cannot be established on a purely empirical basis. It is therefore important that other

types of validity, in particular content and construct validity, be considered. In the context of HRA techniques, this means that the underlying model utilized in the technique should be appropriate for the types of situation to which it will be applied. This is particularly important for the qualitative, human error model evaluation aspects of HRA, discussed in Section 1.6. For an HRA technique to be valid, the model which specifies the likely human error modes and the process whereby numerical estimates of HEPs are obtained must both be valid. A discussion of validity in a human-machine system context is available in Hollnagel (1981), and Cronbach and Meehl (1955) consider the validity issue within the context of psychological testing. A likely basis for the evaluation of the results from the SLIM test is convergent validity. In effect, this will mean comparing the numerical outputs of SLIM with those provided by other methods (e.g., THERP, absolute probability judgments [SNL, 1983], Influence Diagrams [Phillips, Humphreys, and Embrey, 1983]).

1.5.2 Auditability

The technique should have a clear and explicit structure such that the procedures used in obtaining the outputs can be readily traced for auditing purposes. This implies that the technique should include facilities for documenting the operation of the method. Implicit within the auditability criterion is the requirement that the technique should be systematic and relatively invariant across different applications.

1.5.3 Modeling Capability

There is some ambiguity regarding the use of the term "modeling" in human reliability assessment. One meaning which is applied in Section 1.5.7, refers to the underlying psychological model of human cognition and behavior which is assumed in the analysis. This will be reflected in the types of human error which are likely to be considered by the judges. The process of identifying errors of commission or omission in conjunction with the hardware failure analysis is also referred to as modeling. This form of modeling will be driven by the assumed underlying psychological model, but is obviously not the same as this model. The outputs from the error mode modeling will depend on the analyst's perception of the interaction between the assumed psychological model and the specific conditions of the situation being evaluated.

The function of modeling is not only to derive the whole range of credible error modes, but also to eliminate the incredible ones. The basic requirements for the modeling phase of HRA are that it should be complete yet parsimonious.

Identification of the credible error modes depends partly on the degree of experience of the judges in situations and plants similar to those of the scenario being considered. On the other hand, it is likely that some explicit and systematic procedure for carrying out the modeling process will lead to a more realistic identification of the important error modes. As has been shown by Hollnagel et al. (1981), the use of an underlying model of human

performance can produce such a structure. It will be apparent that the modeling phase in fact includes some degree of implicit quantification, in that the judge makes a dichotomous decision that a particular error mode is either "incredible" and will be excluded, or that it is "credible" and will therefore be subsequently quantified. Thus, a sufficiently flexible HRA technique could be applied directly at the modeling phase to perform a preliminary screening prior to detailed quantification.

Incorrect modeling will inevitably affect the correctness of the final human error estimate that is produced. In the case of decomposition techniques such as Technique for Human Error Rate Prediction (THERP), some branches may be missing in the event tree that models the human errors, resulting in a probability estimate more optimistic than it should be. Similarly, in subjective judgment methods such as SLIM, if the judges are not aware of the full range of human error modes that are possible in the situation being assessed, then their global subjective estimates of success or failure likelihoods will probably be erroneous.

To summarize, the existence of an explicit and comprehensive modeling procedure will enhance the acceptability of a HRA technique to both potential users and the scientific community.

1.5.4 Reliability

Reliability, in the psychological testing sense, means that a technique will generate similar human error probabilities if the same human actions are assessed on different occasions. Specifically, this is test-retest reliability and can be quantified with the test-retest correlation coefficient. This is the correlation between the obtained scores (e.g., human reliability estimates) from the same group of individuals assessing the same set of tasks on two different occasions. Other measures of reliability are discussed in Cronbach (1964).

A number of factors could contribute to variability in an individual's assessments. In the field of psychological testing, there are similar variations in test scores, and lists have been compiled (e.g., Thorndike, 1949) of factors that can affect reliability. In the context of HRA techniques, the use of systematic procedures and trained judges will be important for enhancing reliability. Since validity in this context is difficult to establish empirically, reliability becomes correspondingly more important.

1.5.5 Uncertainty Bounds Determination

For PRA purposes, a HRA technique should be capable of generating uncertainty bounds. This facilitates the use of error probability estimates in bounding analyses, Monte Carlo simulations, etc.

1.5.6 Sensitivity Analysis Capability

The technique should make it possible to identify which characteristics of a situation have the greatest effect on the probability of error, and the impact of this effect on risk so that design recommendations and cost-benefit analyses can be carried out.

1.5.7 Justifiability of Underlying Model

A technique will be preferred if it is based on a theoretical structure or model which is well established and has been experimentally validated. The model should be appropriate to handle the full range of situations to which the technique will be applied.

1.6 The Basic SLI Methodology

The approach underlying SLIM was originally derived from the common sense observation that the likelihood of successfully accomplishing an action or task was a function of various characteristics of the individual, the situation, and the task itself--performance shaping factors (PSFs). PSFs are presumed to combine together in some manner to determine the probability of success. It seemed intuitively reasonable that success likelihood was a function of how good or bad these factors were in a particular situation, weighted by their relative importance in affecting success.

Embrey (1979) proposed that these importance weights could be obtained by carrying out multiple regression analyses using task error rates as the dependent variable and measurements of the factors influencing these error rates as predictors. However, the absence of a large data base of human error probabilities and associated measurements of influencing factors precluded the application of this technique.

Nonetheless, this approach led to a consideration of procedures described in the decision analysis literature, specifically Multi-Attribute Utility Theory (MAUT). MAUT is used in social decision making and many other applications to assist judges in making choices between alternatives which have different values on a number of dimensions or attributes (Edwards, 1977; Johnson and Huber, 1977). For example, if a group of experts had to decide on which of several alternative locations to build a nuclear power plant, the relevant attributes to be considered could be seismic considerations, proximity to large centers of population, availability of cooling water, etc. These factors are weighted to reflect their relative importance, and then each alternative site is rated numerically (or scaled) in terms of these attributes, e.g., Site A might score high on seismic suitability, but have inadequate water resources, etc. whereas with Site B, the situation is reversed. The product of the importance weights and the ratings for each factor is formed and these products are then summed for each alternative being evaluated. This procedure generates a quantity for each alternative which represents a measure of the judges' perception of its "utility" or "value." The alternative with the highest utility is usually the one that is chosen.

In the context of human reliability assessment, the alternatives being considered are often human actions in a nuclear power plant accident scenario. Examples of such scenarios are a Small-Break Loss-of-Coolant Accident (LOCA) and a Steam Generator Tube Leak. In some cases, the judges might decide which of several alternative courses of action is best to achieve a particular goal in the situation being analyzed. For example, several alternative strategies can be adopted to achieve emergency cooling at a critical phase in a Small-Break LOCA. In this case, the interest is in deciding upon the "best" alternative in terms of maximizing the likelihood of success, where success is defined as achieving emergency cooling.

Such a decision is equivalent to placing all the actions under consideration onto a scale of likelihood of success. It is assumed that this scaling can be achieved by procedures directly analoguous to the scaling of alternatives on an expected utility scale in MAUT. Further, it is assumed that these scale values can be subsequently converted to probabilities, as will be discussed in Section 1.6.5. A comprehensive discussion of the formal axiomatic structure underlying SLIM is provided in Section 1.10.

The scale value on the success likelihood scale which is subsequently transformed to a probability is known as the Success Likelihood Index (SLI). The SLI is directly analogous to the overall measure of utility in MAUT.

A step-by-step description of the various stages in carrying out a SLIM evaluation for a task is given in Section 3.1 of Volume I. It is repeated here with an extended technical commentary.

1.6.1 Derivation of Performance Shaping Factors

Explicit consideration of Performance Shaping Factors (PSFs) is the basic underpinning of SLIM. The term PSF is used to denote both human traits and conditions of the work setting that are perceived by the judges to have predominant influence on the success likelihood in the scenario being evaluated. Not <u>all</u> the factors influencing success are expected to be identified. Rather, what is required is the identification of a relatively small set of PSFs that account for the major part of the variability of success likelihood that will be encountered in a realistic range of situations. Typical "human traits" PSFs might include quality of training, psychological state, motivations, etc. Conditions of the work setting that might shape performance include time available to complete a task, task performance aids, etc.

It should be noted that the term "Performance Shaping Factors" is used here in a more general sense than in THERP as described in Swain and Guttmann (1983). In THERP, PSFs are generally used to modify baseline probabilities obtained in NUREG/CR-1278. By contrast, PSFs in SLIM are defined as follows: "Those factors which acting alone or in combination determine the probability of success of a human action in a human-machine system." Thus, PSFs are at the very foundation of SLIM, not simply adjustments to baselines.

The derivation of PSFs will normally be a consensus process carried out by the participating judges. In some cases, definitions of PSFs to be considered may be provided a priori to the judges. Using MAUD to implement SLIM allows the set of PSFs to be developed via an interactive dialogue (see Chapter 3 of this volume). As described in Section 3.1.1 of Volume I, the modeling of the various error modes which may affect the likelihood of success being achieved is also carried out at this stage.

1.6.2 Weighting of PSFs

It is necessary to determine the perceived relative importance of the PSFs in terms of their effects on the likelihood of success. This procedure is described in Section 3.1.2 of Volume I. The method has been revised slightly from that described in earlier accounts of SLIM, e.g.--Embrey (1981). The revised technique, described in Volume I, Section 3.1.2, was adopted because some judges found the original weighting procedure somewhat difficult to apply in the earlier questionnaire studies reported in Embrey (1983). The revised technique, based upon direct weighting of the PSFs by the judges, is still open to some theoretical objections, as discussed in Humphreys (1977) and Section 1.11. However, in many applications it has produced reasonable results (e.g., Kneppreth et al., 1978; Huber, 1974). MAUD obtains importance weights via an alternative procedure, described in Section 1.10, that overcomes these objections. Although there are several techniques for obtaining PSF weights, it is highly recommended that, whenever possible, the techniques described in Section 1.11 be built into any implementation of SLIM.

1.6.3 Rating of PSFs

Judges rate the PSFs by assigning a value to each PSF on an equal interval scale to represent the degree to which each factor is minimal or maximal in promoting success in the situation being evaluated. It is desirable to supply some anchor points or examples on the scales to ensure that all judges have an equivalent understanding of what constitutes the endpoints and midpoints of the scale. This method of scaling suffers, however, from some theoretical problems which will be discussed in Section 1.11.

1.6.4 Calculation of the SLI

SLIM assumes that the SLI is a linear additive function of the products of the PSF weights and ratings. Hence the equation for computing the SLIs for j tasks being evaluated on i PSFs is:

$$SLI_j = W_i \cdot R_{ij}$$

where:

SLI_j = the combined utility of the various PSF in enhancing the likelihood of success for task j (i.e., the SLI for the task),
W_i = normalized importance weight for the i'th PSF ($\Sigma W_i = 1$),
R_{ij} = scale value (rating) of the j'th task on the i'th PSF.

The linear additive model used in SLIM has been extensively employed and has been shown to be very robust, and suitable for application to a wide variety of settings (Dawes and Corrigan, 1974). The primary assumption of the model is that the PSFs being considered are independent of one another. It is therefore important that users of SLIM ensure that this assumption is met, by, for example, deleting or combining PSFs which appear to be similar in meaning. In the SLIM-MAUD procedure the interactive computer program continuously monitors the degree of dependence during the interaction process and prompts and aids the user to restructure the PSFs, where necessary, to preserve the validity of the additive, independence assumption.

1.6.5 Transformation of the SLI to Probability

As discussed in Section 3 of Volume I, the conversion of SLI values to probabilities (or "calibration") is achieved in SLIM by the use of a linear logarithmic calibration equation of the form:

$$\log \text{ of the probability of success} = a\,SLI + b$$

where:

a and b are empirically derived constants.

These constants are determined by asking the judges to assess at least two actions for which empirically determined probabilities (from error frequency data) are available. The existence of a reasonably consistent relationship between the SLI and success (or failure) probability is obviously important if SLIM is to be used to generate point estimates of human error probabilities (HEPs) for the purpose of PRA. However, the exact nature of this relationship is primarily an empirical question. Work by Pontecorvo (1966), Hunns (1982), and Embrey (1983a) support the logarithmic relationship as do the results presented here. However, the generality of the relationship can only be established by a comprehensive program of research. This issue is discussed further in Section 1.10.1.

1.7 Experts to be Used in SLIM and SLIM-MAUD Assessments

Experience in the use of SLIM suggests that the most important requirement is the composition of the group of judges. The groups should be composed of judges representing a range of expertise. At least one individual should have some direct knowledge and experience with the specific plant being assessed. This is because there are usually plant-specific characteristics both at the hardware level, and the team and organizational level, which will have a substantial effect on the reliability of the operators. Although it is sometimes difficult to obtain control room staff from an operating plant, due to shift and other constraints, experience indicates that trainers familiar with the plant being assessed are equally effective. In terms of understanding the detailed nature of the situation the operator will have to handle, it is also useful if one individual in the group has some knowledge of plant thermohydraulics. The group should also include at least one individual

with human reliability or human factors expertise in order to ensure an adequate assessment of factors influencing operator performance. These areas of experience may come from any of the following populations of experts: operators/supervisors/trainers, plant designers, PRA specialists, and HRA and human factors specialists.

As the background of these judges will be different, one would expect their detailed expertise with regard to the PSFs being evaluated to also differ. These differences and their significance for the reliability of assessments made using SLIM-MAUD can be investigated empirically. An experimental procedure for carrying out such an investigation will be described in detail in Chapter 4 of this volume.

1.8 Applicability of Data Generated by SLIM-MAUD to PRA

Because SLIM allows tasks to be analyzed at any level of aggregation, it can be used at whatever level of analysis is required within a PRA. Thus, probabilities can be generated for individual task elements such as opening valves, or global assessments can be made concerning the likelihood of success of courses of action which comprise complex aggregations of simple task elements, e.g., the likelihood that an operator will successfully transfer to the recirculation mode of core cooling after the depletion of the refueling water storage tank during a PWR LOCA.

Since SLIM is not a reductionist* technique, the problem of how to combine failure probabilities for use at higher levels of analysis within a PRA does not arise. The methodology is applied directly at whatever level of disaggregation of tasks is desired. In the Test Plan described in Chapter 4, the use of a MAUD-based implementation of SLIM is proposed to analyze directly tasks at two distinct levels of aggregation--Level A vs Level B tasks from the list of 35 risk analysis tasks developed by the NRC and Sandia National Laboratories (SNL) (SNL, 1983). A detailed list of these tasks is contained in Appendix B.

1.9 Implementation and Procedure

The preliminary results from the experiment described in Chapter 2 of this volume indicate that with minimal training judges were able to make acceptably accurate predictions in the majority of the cases studied when using the early version of the basic SLIM approach. Additional experience with the use of SLIM are consistent with these results.

1.9.1 The Role of the Facilitator

The facilitator is the individual who conducts a SLIM session. Preliminary indications are that the facilitator is essential only during the first few sessions with a particular set of judges. The main questions to be addressed during the initial stages of SLIM involve deciding on definitions of

*That is, it does not rely upon the decomposition of tasks into subtasks and sub-subtasks, etc. as is done, for example, with THERP.

PSFs, explaining the mechanisms of the rating and weighting procedures, and carrying out the necessary calculations. A facilitator will be neccessary in every case where Integrative Group Process (IGP) techniques are used to resolve differences between individual judges or where SLI assessments are to be based on group rather than individual judgments. In this case the facilitator's role will be that described in published accounts of the application of IGP (Gustafson et al., 1983). In the MAUD-based implementation of SLIM, MAUD itself assumes much of the facilitator's role in the manner described in the foregoing section. During an actual implementation, however, a facilitator may be useful in assisting judges.

1.9.2 Modeling Capability

Although the MAUD-based implementation of SLIM provides sophisticated procedures for generating models of the judges' perceptions of the effects of PSFs on the likelihood of success, neither it nor any other implementation of SLIM provides explicit procedures to facilitate the other aspect of modeling--i.e., the identification of error modes in the scenario being evaluated. In practice, however, it has been found that the derivation, weighting, and rating of PSFs provides very useful input to the modeling aspect of HRA. Although quantification is usually carried out subsequent to the modeling phase, the judges often reevaluate their earlier assessments of the likely failure modes after having carried out the SLIM or SLIM-MAUD procedure. The reentrant editing facilities available within the MAUD-based implementation of SLIM may then be employed to revise the judges original SLIM assessments to bring them in line with these conceptual changes. Then, MAUD automatically computes the SLIs.

1.10 Technical Issues Within SLIM and SLIM-MAUD

In previous sections, an overview of the basic SLI methodology was presented. In this section, a number of technical issues, central to SLIM, will be explored in greater detail.

1.10.1 Calibration

The question of calibration becomes important only when absolute probability estimates are required. It should be emphasized that SLIM can be used very effectively as a prescriptive technique for design or to carry out sensitivity analyses, without necessarily employing calibration. Thus, if the objective is to compare the relative error likelihoods of two or more different actions, or the same action after changes are made in the operator-system interface, then the SLIs of these actions can be compared directly in determining the degree to which one rather than the other of the actions meets their objectives. The relative position of the tasks on the Success Likelihood scale will remain invariant regardless of the calibration procedures used to convert the SLIs to probabilities. This property of the SLI can be used prescriptively in many safety study applications. Where absolute probabilities are required, in fault tree analysis for example, calibration is necessary.

The function of calibration is to convert the SLI values for the set of actions to be compared into an absolute probability value. This can be achieved if two or more reference tasks are available for which absolute probabilities are known and which can be assessed using the same PSFs used to evaluate the set of actions of unknown probability. This requirement is to ensure that the reference tasks belong to the same population (in the sense of being influenced by the same PSFs) as the target tasks.

Another requirement is the existence of a consistent monotonic relationship exists between the SLI scale and the probability of success. In SLIM, this is assumed to be of the form:

log of the Probability of Success = a SLI + b

where:

a and b are empirically derived constants.

As discussed in Embrey (1983a) and Seaver and Stillwell (1983), the log relationship receives some support in the literature, but the evidence is insufficient to conclude that it is firmly established. Nonetheless, an intuitive justification has been produced by Hunns (1982), and the experiment reported in Section 2.2 provides further empirical support for the appropriateness of the relationship. A further discussion of the nature of the calibrated process within an "expected regret" structure is contained in Section 1.11.8. However, as emphasized in the discussion in Section 4.1 of Volume I, the validity of the SLIM method does not stand or fall by the "correctness" of the logarithmic or any other specific conversion equation. There are various approaches to calibration and these will now be considered in the following sections.

1.10.2 Use of Two Calibration Tasks

Problems arise with this approach if the calibration tasks are not evaluated by the judges in a manner which is consistent with the rest of the data set. This difficulty can be minimized if selection of reference tasks has been based on taxonomic considerations to ensure that they belong to the same category as the tasks being evaluated. The experimental test of SLIM, described in Chapter 2, demonstrated that there are considerable problems in using a task taxonomy without a systematic method for assigning tasks to appropriate categories. Stage 1 of the Test Plan described in Section 4.5 proposes procedures that will allow tasks to be categorized in a structured way into clusters for which specific calibration tasks can be provided.

1.10.3 The Regression Approach

The availability of a number of reference tasks allows the use of the regression approach, which is the approach used in the experimental test of SLIM and discussed more fully in Chapter 2 of this volume. The method has the advantage of taking into account the inevitable variability associated with the judges' SLI estimates of the reference tasks. Another advantage is that it allows the coefficient of determination (the square of the correlation coefficient) to be calculated, which provides a measure of the amount of variability in the log probability accounted for by the SLI values. If the coefficient of determination is low, then this is a warning that the calibration of the judges may be inadequate. It may be that one or more judges in the consensus group has some misunderstanding about the meaning of the PSFs or the nature of the tasks. One solution is to provide the judges with feedback about how these possible misunderstandings may have led to inadequate calibration. The tasks can then be re-assessed to see if calibration has improved. This may require more than one iteration. The process of providing feedback to SLIM-MAUD users to improve calibration is completed with relative ease because of the interactive-based nature of the MAUD program. A more drastic solution to miscalibration is to convene a new group of judges who may be capable of more consistent assessments.

Even if only three calibration tasks are available, this will allow some check to be made on the consistency of the judges' calibration. If lack of coherence is detected, some of the methods discussed above could still be employed to improve consistency.

1.10.4 The Use of Absolute Probability Judgments for Endpoints

This method was employed in the SLIM field study, described in Section 5.2 of Volume I and Section 2.2 of this volume. The technique requires the judges to make absolute probability judgments of the best and worst cases for the scenario being evaluated, i.e., the situation where all the PSFs are as bad as they can credibly be in an operating licensed plant, and, conversely, where they are all as good as they can credibly be in a real plant. These two scenarios are assigned SLI values of 0 and 100 respectively; in other words, they are used to define the endpoints of the SLI continuum. Substitution of these boundary conditions into the general SLIM equation (given in Section 1.6.5) produces the following calibration function:

$$P_f = [(LP)^{SLI/100}] \cdot [(HP)^{(1 - SLI/100)}]$$

where:

P_f = probability of failure,
HP = judged probability of failure under worst conditions (higher probability),
LP = judged probability of failure under best conditions (lower probability).

The reason for using this procedure is that in many situations, particularly rare-event scenarios, calibration tasks estimated by frequency data may not be available. By defining the reference situations as the endpoints of the SLI scale, the need to elicit SLI values for reference tasks from the judges is avoided. In using this approach, it is important to emphasize that the reference situations refer to the specific plant and the specific scenario being evaluated. Also, the judges are asked to make estimates for "credible" as opposed to hypothetical conditions, to enable them to make reasonable extrapolations from the present plant state.

This calibration method effectively uses the SLI to interpolate between two probabilities on the basis of the PSFs for the tasks being assessed. This method of calibration may appear to be inferior to that based on reference tasks because it employs absolute probability judgment. However, there is evidence that experienced judges can make well-calibrated probability estimates in some situations, e.g., weather forecasting (see, for example, Lichtenstein et. al., 1981). On the other hand, the probabilities that occur when estimating the worst and best cases in rare-event scenarios are likely to be much more extreme than in the applications cited in the above reference. Techniques exist that allow judges to estimate very low probability events (Selvidge, 1980; Stael von Holstein and Matheson, 1979) through consideration of the probability of their occurrence being contingent on the occurrence of other infrequent (though, less rare) events for which probabilities have been established. These multistep absolute judgment techniques may be of use in the assessment of human error probabilities for rare events. However, they do not have facilities for conducting sensitivity analysis nor the internal checking capabilities of SLIM and SLIM-MAUD. Another technique which can be used to generate reference probabilities for use in SLIM is the Influence Diagram approach, described in Embrey (1983b).

1.10.5 Discussion

The preceding sections have indicated several alternative approaches to calibration. Which approach is the best is an empirical question which must be subject to further research. In the case of rare events, it is obvious that the effectiveness of calibration cannot be verified by a comparison of the HEP estimates with frequency data. The most effective approach may be to aim for convergent validity by comparing the results produced from the use of different techniques, e.g., SLIM, the Influence Diagram, Absolute Probability Judgment and Paired Comparisons.

1.10.6 Uncertainty Bounds Determination

PRAs often require the assignment of uncertainty bounds around estimates of human error probabilities. To some extent, the concept of uncertainty is more appropriate to error rates estimated by frequency data, than to subjectively derived HEPs. However, measures of uncertainty can be derived by various means. One approach is for the judges to assign uncertainty bounds via absolute judgment. Procedures for achieving this are described in

Seaver and Stillwell (1983). They suggest the use of an odds response scale on which each expert is asked to mark the upper and lower bounds of the HEP estimates. An example of such a scale is presented as Figure 3.1 of Volume I.

Seaver and Stillwell (1983) also describe a method for calculating uncertainty bounds when individual SLI values are available for each judge. This involves calculating the variance of the log HEP estimates across judges as follows:

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

where m is the number of judges and HEP_{ij} is the HEP estimate for event i by judge j. The standard error of these estimates is then calculated as follows:

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

The 95% uncertainty bounds are then given by $\log HEP_i \pm 2$ s.e. This method was used in the field study of SLIM discussed in Chapter 2 of this volume.

1.10.7 Inter-judge Consistency

Questions of inter-judge consistency arise only when the mathematical aggregation of individual judge data is employed. Where this method is used, Seaver and Stillwell (1983) provide procedures for estimating inter-judge consistency using the intraclass correlation coefficient as shown in Section 2.2.2 of this volume. Generally speaking, however, whenever a group of judges is available, SLIM should be exercised in a consensus mode, as this provides the optimal means of using information which different judges may possess. Experience gained in the development of SLIM and its preliminary applications indicate that the use of mathematical techniques is not optimal for aggregating the assessments of multiple judges. This recommendation also applies to any other indirect assessment technique that relies on expert judgment (e.g., Influence Diagrams, Paired-Comparisons, etc.), as much diagnostic information is lost through regression effects. Use of Integrative Group Process Techniques (IGP) (Gustafson et al., 1983) is considered superior in measuring differences between judges because they permit the experts themselves, not the technique, to achieve consensus. Although the consensus mode requires the use of additional resources in terms of bringing experts together, its advantages outweigh these additional costs.

1.11 The SLIM-MAUD Approach: Detailed Technical Considerations

Experience gained in the development of SLIM, discussed in preceding sections, has provided the basis for further improvements in the methodology. For example, it was discovered that the procedures used in the early version of SLIM (described in NUREG/CR-2986) to obtain weights and ratings are not theoretically optimal (Humphreys, 1977), even though they are still capable of producing usable results, as is evidenced by the studies described in Chapter 2 of this volume.

Subsequent sections are devoted to the development of the detailed axiomatic basis underlying SLIM. As part of this development, those aspects of the early SLIM technique, which were theoretically sub-optimal, will be identified. Also discussed will be how the resolution of this sub-optimality can be achieved through improved implementation of SLIM, such as the MAUD-based SLIM-MAUD.

1.11.1 The Foundation of SLI Methodology: Multi-Attribute Utility Theory

SLIM is founded on the assumption that the courses of action evaluated are possible alternatives, which could be chosen for implementation within the situation for which the assessments are made. Rational theories of choice between alternatives are founded on the notion of preference (Savage, 1954; Fishburn, 1970; Keeney and Raiffa, 1976): one should choose the alternative for which one has the greatest preference, given one's current goal. The SLI methodology assumes that preferences should be formulated in accordance with the goal of success: the greater the likelihood of success of a course of action, should it be implemented, the greater will be the relative preference for it, compared with the other alternatives under consideration.

Within the SLI methodology the alternatives assessed are actions (or the success of actions). In the case of nuclear power plants (NPPs) it is not necessary that these alternative actions be restricted to the set of alternative actions open to an operator at just one point in a particular sequence. Rather, the alternatives may range over those which would be assessed according to the following hypothetical question: Consider a situation where action A has to be performed and a situation where action B has to be performed. Assuming that the outcomes of these situations, given their success or failure, are equally positive or negative, which situation would you prefer?

The SLI methodology finds its application, however, in contexts where neither success likelihoods nor relative preferences can be estimated directly with any degree of confidence. The methodology is specifically designed to overcome this problem by providing a decomposition method for identifying the set of Performance Shaping Factors (PSFs) which contribute independently and collectively to the overall likelihood of success, together with a composition rule which enables the (decomposed) ratings of each course of action on the set of PSFs to be transformed into a single number--the Success Likelihood Index for that course of action.

Success likelihoods are expressed on a scale of relative likelihoods of success of each course of action in relation to others. Where absolute values of likelihoods of success are available for two or more courses of action under consideration, the relationships described in Section 1.10 may be used to calculate absolute likelihoods (on a probability of success scale) for all the courses of action under consideration.

1.11.2 Multi-Attribute Utility Theory Axiomatization of Decomposition of Alternatives

The appropriate decomposition of preferences between alternative courses of action into ratings on PSFs is that developed within Multi-Attribute Utility Theory, described in detail in von Winterfeld and Fischer (1975), and Humphreys (1977). Outlined here are the major assumptions which together form an axiomatization prescribing a simple additive composition rule for computing success likelihood indices.

The decomposition depends on the assumptions of connectedness and transitivity of choices (Arrow, 1952; Fischer, 1972), fundamental to all theories of rational choice, together with the certain critical monotonicity and independence assumptions, discussed below.

1.11.3 Monotonicity Assumption

Given the adoption of an ordered scaling method describing positions of alternative courses of action on PSFs, the monotonicity assumption requires that each PSF should be scaled in such a way that:

$$x_{ij} \succcurlyeq x_{ik} \text{ iff } f(x_{ij}) > f(x_{ik})$$

where $f(x_{ij})$ is the numerical SLI value assigned to the jth course of action on the ith PSF, and x_{ij} represents the relative preference for performance at level x_{ij} on PSF i.

$\succcurlyeq$ denotes "is preferred at least as much as", and $>$ denotes "is numerically greater than or equal to." That is, on each PSF, larger numerical values should imply greater preference for performance at the levels they index.

Use of a scaling metric is simply a device to allow the use of numbers to represent preference orderings (Beals, Krantz & Tversky, 1968). When scale values "as obtained" do not represent this interpretation, the "folding" technique described in Section 1.10.6 may be used to rescale the values in such a way that the monotonicity assumption is met. (This technique is automatically applied within MAUD, in such a way that the monotonicity assumption is always met.)

1.11.4 Preference Independence Assumption

In the SLI methodology, scores on PSFs contribute additively to the SLI index for each course of action. It is therefore important that any set of

PSFs comprise factors which contribute independently of one another to overall likelihood of success. When SLIM is implemented through MAUD, this is achieved by testing the assumption of preference dependence between ratings of courses of action on all factors. If the independence assumption is violated, MAUD detects the violation and restructures the set of factors in such a way that the retained factors meet the independence condition.

The method adopted for testing preference independence within MAUD is based on testing for Weak Conditional Utility Independence (WCUI) between each factor in the set and all other factors in turn (this is called testing n-WCUI for each factor [see Raiffa, 1969; and von Winterfeld and Fischer, 1975]). The definition of independence contained in WCUI is weaker than that contained in definitions of statistical independence (for example, that employed within multiple regression methods). Hence, tests of statistical independence are too strong in this context. Nonetheless, they may be used as a stringent test of the possibility of a violation of WCUI. MAUD performs checks for statistical independence as a guide for further action which may involve structural reordering of the set of PSFs. MAUD's statistical checking procedure monitors potential failures of n-WCUI between each PSF introduced into the set and every other PSF already in the set. Should the statistical check fail, the offending pair of PSFs are presented to the user, and a thought experiment is then conducted between SLIM-MAUD and the user to ascertain whether n-WCUI has actually been violated. If it has, the user is prompted to identify a new PSF to replace the offending pair, and the structure is reordered appropriately.

Technically, ensuring n-WCUI in this way guarantees the adequacy of the decomposition and the correctness of the additive composition rule used in computing the SLIs. If there is no uncertainty concerning which course of action will actually be adopted, that is in cases where the MAUD results will be taken as prescriptive, the course of action with the highest SLI will be that which is actually adopted in the given context.

However, there are cases where assessments have to be made where the results will be valid in any eventuality, i.e., where the SLI can be interpreted in relation to the absolute probability of success of a course of action, rather than just relative to the other courses of action considered in the SLIM assessments. In such a case, the Weak Conditional Utility success assumption has to be strengthened within Multi-Attribute Utility Theory to a Strong Conditional Utility (SCUI) assumption. Direct interactive testing of SCUI is very difficult (Raiffa, 1969; Keeney and Raiffa, 1976; Humphreys, 1977). However, there is a more straightforward way of ensuring SCUI than searching for appropriate direct test procedures. In every case where n-WCUI is satisfied but SCUI may not be, a "prescriptive" decomposition procedure may be used providing that (i) the preference functions are expressed as utility functions U_i adequate for use in conditions where there is uncertainty about the course of action to be chosen, and (ii) a "marginality" assumption can be made (Raiffa, 1969; Fishburn, 1970).

MAUD adopts this approach using an assessment procedure for U_i based on an axiom system for "allocation of importance" devised by Sayeki (1972).

Within SLIM, the resulting U_i values constitute Success Likelihood Indices for courses of action for transformation into probabilities of success in all cases.

This procedure, described in Section 1.11.7, generates a set of weights to be assigned to ratings on the various PSFs which ensures the use of correct scaling factors which meet SCUI requirements. Other direct weighting techniques for PSFs do not meet this requirement. Hence, despite their apparent simplicity, they may not be appropriate for use within SLIM in applications where the resulting SLI values are to be transformed into assessed probabilities of success for defined courses of action.

1.11.5 Additive Composition Rule for Computing SLIs

Once the independence assumptions described above have been met, the following model may be used as the composition rule aggregating ratings on PSFs in computing the SLIs of the alternative causes of action:

$$X_j \succ X_k \text{ iff } SLI(X_j) = \sum_{i=1}^{n} u_i(x_{ik}) \geq \sum_{i=1}^{n} u_i(x_{ij}) = SLI(X_k).$$

Given a scaling procedure which yields values on PSFs $g_i(x_{ij})$ (i.e., the ratings assigned to the PSFs), monotonically related to $u_i(x_{ij})$, a procedure based on Sayeki's (1972) axiomatization of allocation of importance may be employed to construct the $u_i(x_{ij})$ directly.

The relation is of the form:

$$u_i(x_{ij}) = \lambda_i [g_i(x_{ij})] \text{ where } \lambda_i = 1.$$

The λ_i are in fact products of

(value-wise importance weight) x (relative scaling factor)

w_i q_i

Hence, in "separated" form:

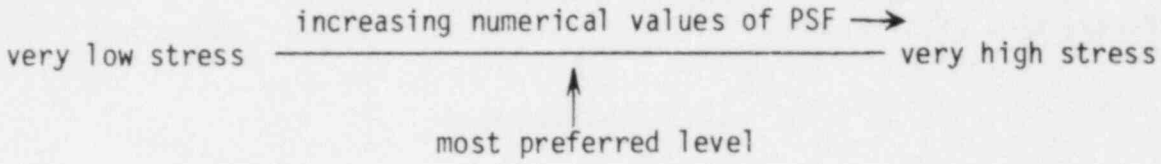
$$u_i(x_{ij}) = w_i q_i [g_i(x_{ij})].$$

From a conjoint measurement perspective, the separation of λ_i into $w_i g_i$ is both unnecessary and pointless since, in practice, w_i and q_i cannot be assessed separately from one another. Hence, the procedure employed within MAUD for the assessment of λ_i does not attempt any such separation.

1.11.6 Folding Procedure for J-scaled Assessments on PSFs

Sometimes, level preferences on PSFs may not be monotonically related to linearly increasing numerical assessments on the scale.

For example, on a PSF scaled from:



The most preferred level does not occur at either pole on the factor. Preferences for stress values start off by increasing monotonically with stress, as at the very low end, increasing the stress associated with a cause of action may be functional through enhancing motivation and vigilance. However, as stress is increased further, preference for these higher levels of stress starts to decrease, as higher levels of stress are no longer functional, leading to fragmentation of action and ineffective behavior associated with high anxiety (O'Brien, Rosa, and Stengrevics, 1983).

This example illustrates what is known as a single-peaked preference function on a PSF. Coombs (1964) has described how these functions frequently arise in practice. Elicited PSFs tend to be identified in "J-scales" where J stands for "joint"-shared across people, whose common language serves to identify "natural" poles for the scales, like "very low stress" and "very high stress." However, in preference-technology, such as that employed in SLIM methodology, the scales required are I-scales (Coombs, 1964; Dawes, 1972). I-scales are individual preference scales, where the most preferred value, which must correspond to the pole indexing the largest numerical value on the scale, will depend on the individual context and the individual goal operative in the application concerned. Coombs developed "folding" techniques, whereby a single peaked preference function on a J-scale can be "folded" about the "ideal point" (the most preferred level) on a J-scale PSF to yield I-scaled PSF values, appropriate for use as g_{ij} in the additive composition rule described in Section 1.11.5. When used to implement SLIM, MAUD ascertains the ideal points on each PSF, and folds the J-scaled ratings given by the user about this point as the first step in the process of testing independence assumptions and constructing the SLIs. In other words, the ratings are rescaled in terms of their distance from the ideal point.

1.11.7 Compensation Method for Assessing Weights for PSFs

The full computational procedure for this method employed in SLIM-MAUD is quite complex, as it requires a preliminary cluster analysis to determine the optimal sequence of the assessments required for the practical implementation of Sayeki's (1972) axiomatization of allocation of importance. Complete details of the procedure are given in Humphreys and Wisudha (1983). Presented here is only the form of the key operations involved in computing the λ_i.

In early implementations of the procedure (e.g., von Winterfeldt and Edwards, 1973), each $\lambda'_i (=W_i g_i)$ was determined by observing how a decision maker's wholistic u_j ratings of hypothetical courses of action changed when values of their (hypothetical) levels on attributes equivalent to PSFs were changed from the "worst" to "best" levels.

Consider the effect of switching from worst (0) to best (1) on PSF 1. According to the conjoint measurement model used here (Krantz et al., 1971):

$$\Delta F_j = \left[\sum_{i=2}^{n} \lambda'_i g(x_{ij}) + \lambda'_1(1)\right] - \left[\sum_{i=2}^{n} \lambda'_i g(x_{ij}) + \lambda'_1(0)\right] = \lambda_1,$$

where ΔF_j is the change in the wholistic rating of outcome j, similarly for all other PSFs.

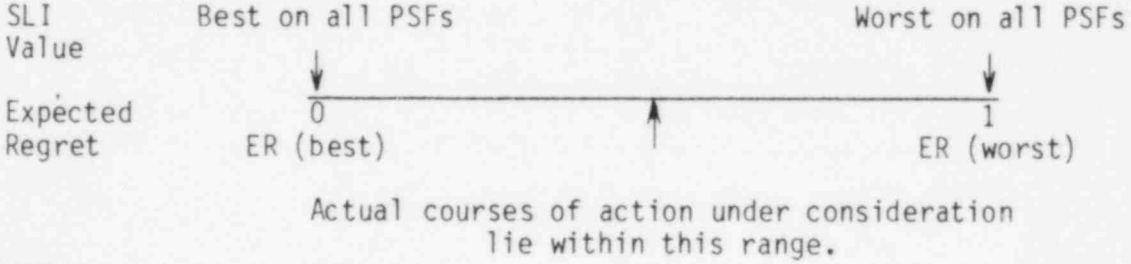
In the development of the method used in MAUD, preferences between alternative switches from worst to best on pairs of PSFs are assessed, (a much clearer and more sensitive measure than the "overall change" method used in earlier assessment procedures based upon Sayeki's method). The optimal sequence of pairs is generated on the basis of a cluster analysis of I-scaled ratings on PSFs and only n-1 "compensation" assessments need be made in determining the values of a complete set of λ_i', i=1 to n (n is the number of pairs).

1.11.8 Conversion of SLIs to Numerical Probabilities

Determination of the precise form of the calibration equation to transform SLIs into probabilities is an empirical question. Discussed here is a calibration equation which makes use of the notion of "expected regret" (Savage, 1954). SLIs are scaled in such a way that for any course of action:

$$SLI_j = 1 - ER_j ,$$

where ER_j is the expected regret associated with selecting course of action j, rather than the hypothetically "most preferable" course of action, which could be represented by a rating at the most preferred level, on every PSF in the set. This most preferable course of action would be SLI value 1 (or 100 depending on the scales used). Conversely, a hypothetically least preferable course of action could be represented by a rating at the least preferred level on every PSF giving an SLI of 0. Hence, the measure of expected regret implicitly constructed within SLIM is scaled as follows:



where:

$$\log (ER_j) = a\,(SLI_j) + k \ . \qquad (1)$$

a and k can be found from the equivalence shown in the above diagram.

Assume that actual regret which would accrue if any course of action chosen was successful is 0, i.e., no regret is experienced. However, if any course of action fails, then the regret experienced is R, the regret at having failed to avert a failure. The logarithmic form of the equation follows from Bernoulli's (1954) proposal that the increase in expected regret, contingent upon a unit decrease in SLI (the converse of an increase in expected utility from increase in SLI), is inversely proportional to the extent of the SLI value (the degree of regret already experienced) below that indicating certain success (see Galanter [1962] and Lee [1971] for discussions of empirical support for this assumption).

R is a constant over all courses of action, given an independence of path assumption, which says that the degree of regret associated with the consequences of a specified failure is independent of the sequence of actions which were tried without success to prevent its occurrence. Thus, according to the expected regret theory (which is based on the same axiom set as expected utility theory, see Savage [1954]).

For any course of action i:

$$ER_j = RP_j \ , \qquad (2)$$

where P_j is the probability of failure associated with that course of action. Substituting this relationship in Eq. (1):

$$\log (RP_j) = a\ SLI_j + c \ .$$

Hence:

$$\log P_j = a\ SLI_j + c = \log R \ , \qquad (3)$$

but since R and thus log R is a constant, we can express Eq. (3) as:

$$\log P_j = a\ SLI_j + b \ ,$$

where b is the constant c - log R.

SLIM uses this relationship to ascertain values of log P_j (and hence P_j) for all courses of action under consideration.

1.12 Implementation of SLIM Through MAUD: SLIM-MAUD

MAUD5 (the latest version of MAUD, described in detail in Humphreys and Wisudha [1984]) is a general interactive computer-based system for the assessment of choice alternatives that has been extensively developed and tested in decision analysis settings (Humphreys and McFadden, 1980; Humphreys and Wooler, 1981; John, von Winterfeldt, and Edwards, 1983; Kobashi, 1983).

When MAUD is used to implement SLIM, up to 10 tasks can be evaluated simultaneously during one session. As with any implementation of SLIM,

conversion of the SLIs to probabilities requires the inclusion of at least two and preferably three reference tasks for which success probabilities are known. Calibration can also be achieved through the use of other calibration techniques discussed in Section 1.10.1. An example of a SLIM-MAUD session together with commentary is given in Section 3.4 of this volume. This example can be referred to as an aid in reading the following sections.

1.12.1 How MAUD Works With the Judges Using SLIM

This section considers the details of what occurs after a SLIM-MAUD user, the expert-judge, has selected a set of tasks for which he or she wishes to assess SLI values. It should first be noted that MAUD carries out the rating of PSFs prior to the derivation of the importance weights which is the reverse of the procedural order of other techniques for implementing SLIM. After asking the judge for a set of actions to be evaluated, SLIM-MAUD then elicits the PSFs. Each action is then rated on each PSF separately. MAUD gives the judge flexible editing facilities for changing information already given to the computer. These are needed because new ideas and insights often occur to the judge during the interaction process. MAUD tests the coherence of the data supplied by the judge, and derives all the information necessary to apply an algorithm, based on Multi-Attribute Utility Theory, for decomposing the data into overall assessments across the courses of action being evaluated.

MAUD works entirely with the judge's own inputs, asking him or her for words or phrases which describe all the important elements of his or her understanding of the problem. Its operation and text is geared toward a new user who is not necessarily experienced in interacting with computer systems.

1.12.2 Eliciting and Rating the PSFs

MAUD starts by asking the user to name the actions under consideration. It then proceeds to help the judge elicit PSFs relevant to evaluating these courses of action by asking him or her to specify differences and similarities between triads of alternatives, following Kelly's "difference" method (Kelly, 1955; Fransella and Bannister, 1977). The words thus elicited are used to represent the poles or endpoints of PSFs, and MAUD will allow changes if the judge is not satisfied with the definitions he or she has given as poles. The judge is next asked to rate all the courses of action on a scale between these poles, and to specify the ideal (most preferred in terms of maximizing success) point on the scale. MAUD then "folds" the elicited J-scale ratings about the ideal point into an I-scale and rescales the I-scale so that the least preferred course of action on the folded scale receives the value 0 and the most preferred course of action on the scale receives the value 1 (see Section 3.4.5, Frames 10-13).

When the judge has successfully generated two PSFs which are significant to him or her for distinguishing between the courses of action in terms of the likelihood of success, MAUD allows poles of PSFs to be specified directly using a heuristic known as the "opposite" method without explicitly

going through the consideration of similarities and differences between triads of alternatives (Epting, Suchman and Nickeson, 1971) (see Section 3.4.5, Frame 21).

MAUD can resume its presentation of triads of alternatives as a means of eliciting further PSFs from the judge at any time he or she requires assistance in considering further important aspects of the situation yet to be explored. This has been found to help in drawing out fresh insights about the factors which may affect the likelihood of success of different courses of action. This procedure can effectively be carried out as a group process.

1.12.3 Editing and Restructuring Rating Assessments

MAUD's difference method and opposite method are structure-eliciting heuristics, originally developed for application in clinical psychology, and are very effective in eliciting material from the judge. There is, however, no guarantee that such material (or material generated by any other elicitation technique) will be coherent or optimal from the perspective of Multi-Attribute Utility Theory, which is the underlying basis of SLIM.

MAUD overcomes this problem in the following ways. It provides considerable facilities for editing material whenever the judge (or MAUD) becomes dissatisfied with the way in which he or she has represented some aspect of the situation. This may be due to incoherence of ratings of courses of action on a PSF, owing to inappropriate specification of poles, failure to find an ideal point, and so on. Editing may involve restructuring the user's view of the situation by changing the ratings or ideal points on PSFs, renaming the poles, deletion of courses of action or PSFs, and replacement by others (see Frames 22 and 23).

Alternatively, restructuring can be initiated by MAUD in interaction with the judge. MAUD monitors the I-scaled ratings on the PSFs input by the judge, checking each set as soon as it is elicited with the sets of I-scaled ratings on all other PSFs currently in the preference structure. It is important to ensure that conditional utility independence is maintained between these sets of ratings (see Section 1.11.4). However, checking this assumption directly involves asking a number of rather difficult and very repetitive questions. MAUD therefore takes an indirect approach, capitalizing on the fact that tests for statistical nonindependence are stronger than those for violations of conditional utility independence. MAUD monitors the statistical associations between pairs of sets of I-scaled ratings, and only questions the judge about utility independence when the statistical test indicates that there is a reasonable chance that the conditional utility independence requirement may have been violated, i.e., that there may be interactions between the PSFs (see Frame 18).

The data which form the basis for the computation of SLI should, whenever possible, be expressed in such a way as to permit the use of a simple additive composition rule. If the aspects of the problem are expressed so as to make this impossible, then the problem should be restructured to permit the use of such a rule, in preference to the adoption of a more complex rule.

When there is a violation of PSF independence, restructuring is accomplished through the deletion of the offending PSFs and their replacement with a PSF more appropriately expressing their shared meaning (see Frames 34 and 35).

1.12.4 Assessing Relative Importance Weights of PSFs

When the judge thinks that a sufficient number of PSFs representing all the important aspects of the situation have been specified, MAUD can then investigate importance weights and relative scaling factors for all PSFs in the SLI structure. These quantitites must be determined in order to be able to apply an additive composition rule (von Winterfeldt and Fisher, 1975). Since SLIM is founded on Multi-Attribute Utility Theory, it is important that the procedure for applying this rule also be a foundation of MAUD. In early versions of MAUD, this was achieved by constructing reference gambles or "basic reference lottery tickets" (BRLTs) (Raiffa, 1969).

As reported by John et al. (1983) many people find the BRLTs difficult to assess and so an alternative procedure for assessing importance weights is offered in MAUD5 (Humphreys and Wishuda, 1984) which is the recommended version of MAUD. The procedure built in MAUD5 was developed from Sayeki's (1972) "compensation" method, a theoretically optimal procedure for use under conditions of "riskless" choice (von Winterfeldt, Barron, and Fisher, 1980), which has been compared with BRLT-based procedures by von Winterfeldt and Edwards (1973) and Humphreys (1977). In the version of this technique implemented in MAUD, a clustering algorithm is used to determine the n-1 comparisons required to allocate relative importance weights to the n PSFs used to characterize the courses of actions being assessed. Each comparison involves adjusting the position of an explicitly specified option on one of the two PSFs on which it is defined to find indifference in preference with respect to a "reference option," which has a fixed, explicit definition on the same two PSFs (see Frame 55 onwards).

1.12.5 Assessing SLIs and Probabilities of Success

At the end of each session, or at any other time at the judges' request, MAUD produces a summary showing the assessed PSFs currently under consideration, the judges' ratings of the courses of action in terms of the likelihood of success, and their ratings on the PSFs. When relative importances of the PSFs have been investigated, these are shown, together with the SLI values for each course of action (see Figure 3.2).

When reference probabilities have been supplied for two courses of action, which have also been rated by the judge on the PSFs, an extension to MAUD developed for use in SLIM applications can be used to compute and display the probabilities of success for all the courses of action included in the assessment. (A description of this extension together with its computer code is given in Section 3.5 of this volume.) The judge may then wish to carry out further restructuring, introducing new alternative courses of action, removing old ones, or changing PSFs in interaction with MAUD. At all times, MAUD provides comprehensive editing facilities, so that a user can correct errors and

restructure the problem as he or she wishes. The system is fully re-entrant, which means that restructuring and evaluation activities can be carried out by the user in any order until the final result has been achieved.

1.12.6 Resources Required to Implement SLIM Through MAUD

SLIM may be implemented in a number of ways with or without computer support. Phase IV of the SLIM research program, described in Chapter 4 of this volume, is concerned with a full MAUD implementation of SLIM, using the stand alone procedures described in Chapter 3. With this implementation, the resources required to assess SLIs for any given set of tasks would be as follows:

- Judges. Availability of a group of four to six judges with collective experience stemming from HRA, PRA, plant design, and operations. These experts should meet together in a single group to perform the assessments in interaction with MAUD.

- Software. Availability of MAUD5 configured to implement SLIM in the manner described in Section 3.2 of this volume.

- Hardware. Availability of a single stand-alone microcomputer. This must be either an IBM-compatible personal computer using DOS 2.0 or a CP/M-based system. In either case, the microcomputer must be equipped with two double-density floppy disks, a memory of at least 64K RAM, a display screen, and a printer.

- Office space. Availability of a room where the group of assessors meet to interact with SLIM-MAUD free of outside disturbance.

- Time. A typical SLIM-MAUD session begins with about 30 minutes of introductory discussion and classification of the tasks to be discussed into subsets of 4-10 tasks each (see Section 1.13). Each subset of tasks is then assessed in a session with MAUD, which on average lasts about 45 minutes. Hence, if for example, 14 tasks were to be assessed, divided into two subsets of seven tasks each, the total time required would be 30 + (2 x 45) = 120 minutes. Additional time may be assigned for the initial formalities included in convening the group and for debriefing the judges afterwards.

1.13 Need for Appropriate Classification Scheme

In the earlier work on SLIM reported in Embrey (1983a) the importance of developing a task classification procedure for use within SLIM was emphasized. This is required for several reasons. The SLIM-MAUD procedure develops a common set of PSFs and their associated relative importance weights for all the tasks that are being evaluated together in a particular SLIM-MAUD session. Therefore, there is an underlying assumption that for all the tasks in a particular session the likelihood of success will be influenced by the same PSFs with the same relative importance weights. In order that calibration can be carried out, it is necessary that the reference tasks included within the

SLIM-MAUD session are also sensitive to the same set of PSFs and their associated importance weights. The experimental study of the basic SLIM technique (discussed in Chapter 2 of this volume) noted the problems that can arise if generic PSF weights are not applied to subsets of tasks on a systematic basis. The lack of homogeneity within the tasks assigned to the categories meant that judges had difficulty in applying the pre-defined PSFs.

In addition, it seems likely that the poor correlation obtained in the field test of SLIM between the log HEPs and the SLIs calculated using the generic weights could be attributed to the tasks being inappropriately classified. These considerations suggest the need for a taxonomy which contains categories to which tasks can be assigned on the basis of their homogeneity with respect to the PSFs which influence the likelihood of success.

As part of the research program described in this report, a wide range of taxonomies which had been employed in the human reliability and other areas were surveyed. Thirteen error taxonomies and 12 task/performance taxonomies were reviewed for possible use as task categorization approaches in SLIM. These taxonomies are presented in Table 1.1. (The Altman classification appear in both columns of the table because they address both performance and error classification.)

Table 1.1 Error and task/performance taxonomies reviewed.

Error Taxonomies	Task/Performance Taxonomies
Meister and Rabideau (1965)	Berliner et al. (1964)
Altman (1964a)	Altman (1964a)
Altman (1964b)	Altman (1964b)
Altman (1967)	Miller (1967)
Rook (1962)	Miller (1971)
Meister (1964)	Alluisi (1967)
Swain and Guttman (1983)	Fleishman (1967)
Edwards (1981)	Farina and Wheaton (1971)
Norman (1981)	Theologus and Fleishman (1971)
Reason (1979)	Levine and Teichner (1971)
Adams et al. (1980)	Fleishman (1975)
Metwally et al. (1982)	Levine et al. (1971)
Rasmussen et al. (1981)	

The results of the survey were disappointing as far as the applicability of the taxonomies to SLIM was concerned. Although many of the approaches appeared to be viable methods for grouping together tasks or errors from the point of view of various psychological models, none of them contained systematic procedures for assigning tasks to categories, and very few of them

considered complex tasks with significant cognit ve content, such as those encountered in nuclear power plant safety analyses. It was therefore concluded that some alternative approach to task classification needs to be developed for use with SLIM.

The Test Plan described in Chapter 4 addresses this objective, and incorporates procedures for developing a taxonomy of the tasks to be assessed. The essential requirement for a SLIM taxonomy is that the classification method be based upon the systematic use of expert judgment and specifically oriented toward the types of tasks encountered in nuclear PRA applications. The taxonomy developed for use in a specific context should differentiate between tasks on the basis of PSFs which expert judges generally agree are the major determinants of success or failure in that context.

1.14 Evaluation of SLIM and Its Implementation Through MAUD

In Sections 1.3, 1.4, and 1.5, a number of criteria were proposed for the evaluation of HRA techniques. These were grouped under the following three major headings:

- Practicality
  - Cost
  - Training requirements
  - Breadth of application
  - Data requirements
  - Capability of considering socio-technical and organization factors
  - Difficulty of exercising procedure
  - In-house capability
- Acceptability
  - Scrutability
  - Relationship of technique to PRA approaches and techniques
  - Interface with human reliability data bank
- Usefulness
  - Accuracy and validity
  - Auditability
  - Modeling capability
  - Reliability
  - Uncertainty bound determination
  - Sensitivity analysis

The SLIM technique and its implementation with MAUD will now be discussed in terms of these criteria.

1.14.1 Practicality

Cost. The costs associated with exercising SLIM are difficult to determine at this stage, since there are no experienced users of the technique who could be evaluated on this criterion. In the experiment

described in Section 2.1 of this volume, 21 tasks were evaluated in one day. However, this was perceived to be an excessive workload by the inexperienced judges involved. Also, the tasks were generally simpler than those encountered in PRA. In the field study, it was found that for nuclear power plant situations probably 80% of the time devoted to carrying out an assessment was taken up with technical discussions concerning the nature of the phenomena that the operators would have to handle, and identifying the likely success and failure routes to be expected. Taking this time into account, it is unlikely that more than two hours were used in quantifying each operator action using the original form of SLIM. The use of MAUD has not yet been tested in the field in this way. Given that the length of time was typical and that a team of 10 judges participated in an assessment, this implies a cost of approximately five person-hours per action quantified. However, in the example cited, the judges were relatively unfamiliar with the technique and experienced judges might be expected to be quicker. The resources necessary to exercise SLIM-MAUD will depend on the number of tasks assessed in each session. Using the estimates given in Section 1.12.6, and assuming five judges, this suggests that approximately 45 person minutes per task will be required. Thus, the capability of SLIM-MAUD to handle many tasks simultaneously considerably reduces the overall cost of assessments.

Training Requirements. As discussed in Section 1.3.2, the training requirements for SLIM are unlikely to be excessive, particularly true for the MAUD procedure which has a built-in self-training capability.

Breadth of Application. SLIM can be applied to any set of actions for which evaluation is desired. These include actions at all stages of the system life cycle, including design errors, maintenance, and testing in addition to control room operations. The practicality of this can be assessed within the Test Plan.

Data Requirements. The data requirements for SLIM are less stringent than for most other techniques because data are required only for calibration and not for individual task elements. Only absolute judgment techniques require fewer data.

Capability to Consider Socio-technical Factors. There is no formal constraint on the nature of the PSFs considered within SLIM, which may include socio-technical factors, such as motivation, group and organizational characteristics, etc.

Difficulty of Exercising Procedure. Feedback from judges indicates that the procedures involved in using SLIM, with or without MAUD, are not perceived to be difficult or complex to exercise.

In-house Capability. SLIM can be implemented for use in-house by an organization after a relatively short period of familiarization with

the procedures described in detail in Chapter 3 of this volume. Obviously, some prior training of the assessment group and the individual leading an assessment session will be necessary.

1.14.2 Acceptability

Scrutability. Experience gained from the experiment and the field studies indicate that the procedures involved in implementations of SLIM are regarded as comprehensible in common sense terms. The underlying theory is, in its axiomatic form, quite complex. However, the user need not be aware of this complexity.

Relationship to PRA Approaches. The field study has indicated that the data produced by SLIM are acceptable for PRA purposes. The technique seems to interface well with the other aspects of hardware modeling carried out within the PRA.

Interface with Human Reliability Data Banks. The data generated from a SLIM assessment could be incorporated in any of the human reliability data banks that have been proposed (e.g., Comer et al. 1983). However, one of the major strengths of the SLIM technique is that it allows the context within which an action is assessed to be taken properly into account in forming the assessment. Differences in context (e.g., differences in plant, operating characteristics, etc.) are explicitly reflected in the set of PSF weights used in calculating the SLI that are appropriate in a particular context. Therefore, we consider that each SLI value (or probability of failure derived from the SLI value) should be entered in any data bank together with information about the set of PSF weights employed in its calculation. Without this contextual information, the usual problems concerning the application of failure probability data to a specific application will arise once again.

To avoid such problems, it is recommended that eventually a SLIM data bank should be set up, which contains context-specific information for easy use when making SLIM assessments in any particular context. The data bank could contain (1) prespecified "frames" of PSFs with predefined weights, which would apply to particular categories of tasks and contexts, and (2) a library of machine-readable summaries of SLIM sessions, as produced by SLIM-MAUD (see Sections 3.4.5 and Figure 3.2 of this volume), which can be retrieved for revision and extension of SLI assessments in appropriate contexts. This library could be held on 5-1/4-inch floppy disks, and the appropriate disk could be borrowed from the library each time it is needed in a particular application. SLIM-MAUD has fully re-entrant editing capabilities, and hence a SLIM-MAUD session starting from a "library-held" summary would need only to comprise editing and final assessment stages, without the initial PSF generating stages, as the results for these stages will be readily available on the library disk.

1.14.3 Usefulness

Accuracy and Validity. Accuracy can be checked only for situations for which objective data are available, and such situations tend to be very different from those encountered in high risk nuclear power scenarios. The limited data set which was available for the experimental test of SLIM indicates an acceptable level of accuracy, particularly in view of the fact that an equal-weight model was used. In the absence of "rare-event" data, the validity of the SLIM approach must be judged on the basis of the coherence of its underlying model, the degree to which its predictions are confirmed by other techniques, and by any empirical data that become available in the long term.

Auditability. Within the SLIM procedure, the routes via which the probability estimates are generated are clearly traceable. There is, however, a need to build in an effective documentation process whereby the discussion that leads to the assignment of particular weights and ratings can be preserved. In SLIM-MAUD, all the transactions with the computer are recorded on disk and can subsequently be replayed for auditing purposes.

Modeling Capability. At the moment, the SLIM technique does not provide any specific additional modeling structure apart from the usual task analyses that are conducted as part of a HRA. However, the use of SLIM generates a structure which assists the judges in deciding which error modes are credible, and which are not. Thus, there is a two-way interaction between the modeling and the quantification process.

Reliability. The only formal reliability test of SLIM was conducted as part of the Phase II research, the field test of SLIM. As shown in Section 2.2.2 of this volume, the inter-judge reliability was reasonably high.

Uncertainty Bounds Determination. As described in Section 1.10.6 of this volume there are a number of techniques which can be used to generate uncertainty bounds in SLIM.

Sensitivity Analyses. The availability of importance weights within SLIM constitutes a built-in sensitivity analysis capability. The effects on the success likelihood of varying the quality of the various PSFs in different scenarios can be readily assessed.

Justifiability of Underlying Model. The model underlying SLIM-MAUD has been justified in considerable detail in Section 1.10 of this volume. This justification emphasized the high degree of theoretical rigor upon which the model is based.

2. PHASES I AND II OF THE SLIM RESEARCH PROGRAM

This chapter discusses the results of Phase I, the experimental study of SLIM, and of Phase II, the SLIM field test.

2.1 Phase I Research: Experimental Study Using Basic SLIM

The experimental evaluation of the basic SLIM approach was described in Section 4 of Volume I. Further details concerning the objectives of the experiment, the procedures adopted, and the results are discussed in this section.

2.1.1 Experimental Objectives

The evaluation of SLIM was part of a larger study in which three subjectively based human reliability assessment techniques were compared. However, since this report is concerned primarily with the SLIM approach, only the results from this part of the experiment will be discussed here. The objectives of the experiment, with respect to SLIM, were as follows:

- To test the hypothesized logarithmic relationship between the log (probability of success) and the SLI.
- To evaluate the possibility of using generic PSF weights applicable to broad categories of task, rather than individual weights for each task.
- To compare the probability estimates generated by SLIM with the known empirical probabilities of error for the task set used in the experiment.

The existence of a consistent monotonic relationship, such as a logarithmic function, is an important assumption underlying SLIM and hence it was deemed useful to test this assumption as part of the investigation. The possibility of using generic weights relating to broad taxonomy categories was seen as a potential means of reducing the amount of time necessary to exercise the technique, thereby reducing resource requirements. The comparison of the probability estimates produced by SLIM with the empirical task error probabilities is an important aspect of validating the technique as a whole, at least within the range of probabilities considered in this experiment.

2.1.2 Experimental Procedure

As discussed in Volume I, Section 4, 21 tasks were used in this study, consisting of seven each within the categories of Skill, Rule, and Knowledge based behaviors. Eight expert judges participated in the study: four reliability analysts, two operators, and two human factors specialists. A document sent to these judges several days before the experiment (Appendix A of this volume) introduced the concept of PSFs and described the six PSFs to be used in the experiment. It then discussed the three categories of behavior

(Skill, Rule, and Knowledge based behaviors) as described by Rasmussen (1980). The judges were also provided with comprehensive descriptions of the tasks to be used in the study, divided into the three categories. Each judge was asked to perform a weighting exercise for each PSF for the three generic categories of behavior mentioned above. A total of 18 PSF weights; (six PSF x three task categories) were thus generated by each judge prior to the experimental session.

The session began in the morning with the judges further acquainting themselves with the task descriptions. Following a general discussion of the tasks with the experimenters, the judges individually rated each PSF for each task (6 PSF x 21 tasks = 216 ratings), which took approximately two hours. After a lunch break, the group of eight judges was divided into two groups of four, balanced with respect to expertise (i.e., each group had one operator, one human factors specialist, and two reliability analysts). Each group generated a set of 18 generic consensus weights, based on their original individual weights and subsequent discussion and interaction with other members of the groups.

It was also intended to derive consensus ratings at this stage, but lack of time precluded this. Finally, each judge was given a general questionnaire about the technique to complete.

In summary, the experiment generated eight sets of preconsensus (individual) generic weights, eight (individual) preconsensus sets of ratings, and two sets of (consensus) generic weights. From this, three aggregated sets of SLI values could be generated, using preconsensus, and the two sets of consensus weights, although all three used the same ratings. The raw data on weights and ratings are presented in Tables 2.1 and 2.2. The task descriptions, their sources, and the error probabilities, together with other experimental material, are contained in Appendix A of this volume.

2.1.3 Results and Discussion

2.1.3.1 Test of the Logarithmic Hypothesis

As discussed in Volume I (Section 4.1), the logarithmic hypothesis was tested by first plotting the log (empirical error probabilities) against the SLI values calculated from the generic PSF weights for the Rule, Skill, and Knowledge-based categories and the individual task ratings on these PSFs. The individual judges' SLI values were calculated first and the median of these values was used to obtain the overall SLI for each task. No signficant differences were found between the SLIs calculated from the preconsensus and the consensus importance weights.

Low, nonsignificant correlations were obtained between the log HEPs and all three groups of SLIs (i.e., calculated using preconsensus and both sets of consensus generic weights) and it was suspected that the use of generic weights was responsible for this. The SLIs were therefore recalculated using

Table 2.1 Ranks and Weights for PSFs Within Generic Task Categories.

		KBB						RBB						SBB					
Judge	PSF	A	B	C	D	E	F	A	B	C	D	E	F	A	B	C	D	E	F
1. L.A.	Rank	3	4	5	6	1	2	3	5	4	6	1	2	1	3	5	6	2	4
	Wt.	25	30	60	75	10	20	50	80	70	95	10	40	10	60	90	100	30	65
2. I.W.	Rank	2	4	5	6	1	3	4	2	3	6	1	5	6	5	4	2	1	3
	Wt.	10	50	50	70	10	40	50	30	50	150	10	100	100	70	50	20	10	50
3. B.J.	Rank	2	4	5	6	1	3	6	2	4	5	3	1	1	6	5	2	3	4
	Wt.	15	75	100	150	10	30	100	30	40	75	35	10	10	200	150	20	25	100
4. C.F.	Rank	3	6	5	4	1	2	6	3	4	5	1	2	2	6	5	3	1	4
	Wt.	80	300	150	100	10	60	300	200	250	250	10	35	30	300	200	80	10	100
5. A.C.	Rank	2	5	4	6	1	3	5	4	3	6	2	1	5	6	2	3	4	1
	Wt.	40	200	100	300	10	50	250	100	50	300	30	10	200	300	20	80	100	10
6. G.B.	Rank	1	5	4	6	2	3	4	5	2	6	1	3	4	5	2	1	3	6
	Wt.	10	60	45	70	20	30	45	55	20	65	10	25	30	40	15	10	25	50
7. N.H.	Rank	1	6	4	5	3	2	4	5	3	6	2	1	5	6	3	4	1	2
	Wt.	10	500	100	300	50	30	200	250	150	300	60	10	80	100	20	50	10	10
8. J.M.	Rank	5	6	2	4	1	3	6	5	2	4	1	3	3	4	2	6	1	5
	Wt.	100	75	20	50	10	40	75	50	20	40	10	35	40	60	20	120	10	80
Consensus																			
Group I	Rank	2	6	4	5	1	3	6	3	4	5	2	1	1	3	6	4	2	5
Ss 1-4	Wt.	20	150	100	130	10	25	250	50	100	150	20	10	10	60	140	120	20	130
Group II	Rank	3	5.5	4	5.5	1	2	1	5	3	6	1	2	5	6	2	4	1	3
Ss 5-8	Wt.	25	100	50	100	10	20	60	80	40	100	10	20	180	200	20	120	10	50

Table 2.2a SLIM Ratings Judges 1 to 4.

	J1						J2						J3						J4					
Tasks	A	B	C	D	E	F	A	B	C	D	E	F	A	B	C	D	E	F	A	B	C	D	E	F
1	70	50	80	30	70	90	90	100	70	20	0	60	80	70	45	80	0	75	80	30	10	20	20	70
2	65	65	78	45	55	70	50	50	20	50	10	20	80	75	35	50	0	80	30	60	50	50	80	70
3	80	65	42	45	50	80	80	90	70	50	80	80	60	80	40	50	0	65	70	90	50	70	75	80
4	60	65	45	70	58	75	80	90	50	90	10	70	60	65	50	85	0	70	60	50	65	80	50	60
5	80	80	40	48	55	40	50	90	30	60	60	80	50	50	50	10	50	50	60	80	50	75	75	50
6	65	65	90	40	43	45	90	70	90	80	90	80	60	50	100	30	60	75	75	80	80	70	40	60
7	80	90	40	80	50	70	90	70	100	10	0	60	75	65	80	80	0	50	50	60	80	80	90	50
8	80	80	35	45	50	55	90	80	90	40	0	70	--	70	50	60	0	50	70	85	60	70	70	75
9	80	75	40	40	65	80	70	70	40	50	0	90	60	70	40	80	0	50	60	80	40	20	20	70
10	70	70	60	20	50	30	50	90	50	60	10	90	65	80	50	85	30	65	70	90	50	70	20	80
11	55	68	65	65	50	47	60	90	70	90	10	90	65	80	50	85	30	65	70	90	80	90	80	80
12	50	65	57	35	50	47	60	90	70	10	10	90	65	80	50	20	30	65	40	90	80	40	80	80
13	60	70	55	55	55	50	60	90	70	20	10	90	65	80	50	70	30	65	40	90	80	40	40	80
14	78	70	65	60	50	75	40	50	40	20	10	90	40	70	50	60	50	50	60	85	30	65	75	80
15	30	70	30	40	50	65	70	50	70	90	0	80	50	50	60	70	0	60	75	90	30	85	50	80
16	30	70	40	25	50	35	40	50	30	20	0	90	60	60	50	75	0	60	60	75	40	70	50	70
17	50	50	42	25	50	55	50	60	50	80	0	60	70	65	60	70	50	60	75	40	50	60	60	50
18	35	55	35	35	55	80	50	60	50	80	0	40	50	70	40	80	0	50	60	80	50	20	20	70
19	45	55	60	45	50	70	50	70	90	50	10	60	70	70	70	65	0	30	60	75	30	75	75	30
20	55	65	30	35	65	55	50	40	70	50	0	60	50	40	40	50	0	55	80	75	30	70	50	30
21	65	70	30	30	50	78	50	70	20	20	10	70	50	60	25	60	30	50	70	75	30	40	80	70

Table 2.2b SLIM Ratings Judges 5 to 8.

	J5						J6						J7						J8					
Tasks	A	B	C	D	E	F	A	B	C	D	E	F	A	B	C	D	E	F	A	B	C	D	E	F
1	70	50	20	80	50	70	65	70	25	60	0	40	100	50	100	90	0	60	50	50	30	40	20	40
2	25	75	80	80	50	30	50	80	80	15	60	40	40	50	70	50	75	75	50	50	85	50	80	60
3	80	80	80	80	20	80	80	65	50	30	10	55	80	90	25	100	0	75	50	50	30	50	10	50
4	90	90	25	80	75	80	50	40	25	70	50	70	70	75	50	80	60	50	60	50	60	30	30	60
5	75	75	25	50	75	75	30	65	35	10	60	50	25	70	60	20	80	50	50	50	50	30	30	50
6	90	90	100	90	75	75	20	20	95	20	50	50	80	50	75	50	80	70	50	60	80	60	60	70
7	80	80	25	90	80	50	50	50	95	10	0	35	80	60	90	80	0	70	50	50	70	50	25	50
8	75	90	50	80	80	50	90	60	50	20	0	20	90	80	40	70	0	60	50	50	70	30	20	50
9	80	80	15	75	75	50	55	60	50	80	0	60	60	80	40	50	0	60	50	50	50	20	0	50
10	80	90	25	80	15	75	10	50	50	20	10	65	40	70	50	50	75	60	25	50	50	0	20	50
11	80	90	25	80	15	75	10	50	50	70	10	65	60	70	50	80	90	60	40	50	60	60	20	50
12	80	90	25	60	15	75	80	50	50	30	10	65	40	70	50	40	80	70	20	50	50	80	20	50
13	80	80	25	40	15	50	5	50	50	40	10	50	50	70	75	60	80	50	30	40	50	25	20	50
14	75	90	40	60	75	75	90	50	20	40	60	60	40	70	80	50	60	75	50	50	50	20	50	50
15	90	15	50	90	80	50	10	10	50	40	0	50	90	50	50	90	0	80	50	30	50	50	35	50
16	100	60	40	80	75	80	20	0	50	50	50	65	70	40	40	80	0	70	50	50	50	35	35	35
17	100	75	75	90	25	80	80	50	65	10	70	45	50	25	50	90	20	70	50	50	50	75	50	50
18	80	80	15	75	75	50	50	50	50	80	0	60	60	80	30	50	0	90	50	65	50	20	0	50
19	100	90	80	60	80	80	50	70	20	60	0	60	70	50	50	70	0	80	50	50	70	50	50	50
20	50	40	50	50	50	50	10	50	50	10	0	50	50	30	40	70	0	60	40	40	50	20	20	30
21	80	50	20	30	75	50	90	50	20	50	0	50	50	60	25	60	0	80	50	50	25	50	20	60

an equal weights assumption, which is tantamount to not using the generic weights. The additive scale resulting from this process is similar to the Likert Scale technique which has been extensively employed in attitude and personality testing (Edwards, 1957; Dawes, 1972). In many situations, the equal weights model can be superior to the use of weights estimated by regression analysis, (Dawes and Corrigan, 1974, Einhorn and Hogarth, 1975).

Using the equal weight data, the correlation increased to a significant value of r= -0.60 (p less than 0.005, d.f.=19). A content analysis was carried out to determine the degree of information that was present in the task descriptions supplied to the subjects (see Appendix A of this volume). This, together with the judges' comments, suggested that three of the tasks should be eliminated from the analysis. The analysis was recalculated with data from the remaining 18 tasks and the correlation coefficient increased to -0.71, (p less that 0.001, d.f.=16). The logarithmic assumption of SLIM is therefore supported by this result. This result must be considered parsimonious in that equal weights methods can always be improved by combining them with appropriate prior information (Einhorn and Hogarth, 1975 op.cit.). In other words, if the judges have real knowledge concerning the relative importance of PSFs, the weighting information should be used.

It should be emphasized that the use of generic PSF weights which apply to groups of tasks is not necessarily negated by the results of this experiment. However, it is important that tasks which are grouped together are sufficiently homogeneous, such that the relative importance of the PSFs, in terms of their effects on success likelihood, is identical for all tasks included within a particular set. A possible reason for the lack of success of SLIM when generic weights were used was that they could not be meaningfully applied to the three groupings of tasks the judges were instructed to use. If the tasks had been properly categorized into categories for which common PSFs and weights applied, then it seems likely that the use of the weighting information would have been more effective. Phase I of the Test Plan is designed to develop procedures which would allow a task to be assigned to a category for which common PSF weights apply (see Chapter 4 of this volume).

2.1.3.2 Comparison of SLIM Error Probability Estimates with Empirical Error Probablities

This comparison required the conversion of the SLI values into probabilities in order to compare them with the empirical data. A number of alternative methods are available for this process (see discussion in Section 2.3). The first method investigated was the technique which is usually employed in the Paired Comparison approach (Smith et. al., 1969; Seaver and Stillwell, 1983) of using the highest and lowest error probabilities and substituting these into the basic SLIM relationship to produce a calibration equation. This equation was then used to derive the Human Error Probabilities (HEPs) for the various tasks from the corresponding SLI. The first comparison was to calculate a product-moment correlation between these probability estimates and the empirical probabilities. A low non-significant correlation was obtained.

The reason for this can be seen by looking at Figure 2.1. The use of two calibration points will produce meaningful results only if the method used by the judges for generating SLI values for these tasks is the same as the method for all the other tasks being evaluated.

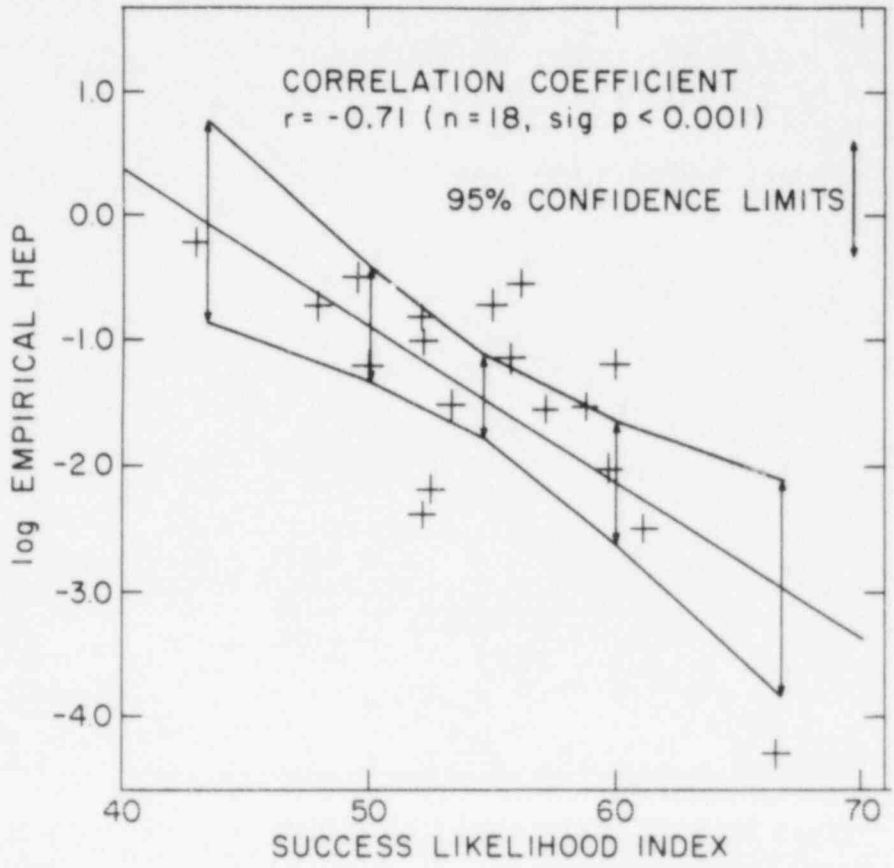


Figure 2.1 Graph of SLI (calculated using equal weights) vs log empirical human error probabilities showing the best fitting linear regression line.

The lower HEP calibration point lies some distance from the regression line for the data as a whole. The use of a calibration equation which is based solely on the lowest and the highest HEP probability point will therefore not adequately represent the whole data set regarding the relationship between HEPs and SLIs. The alternative calibration approach is to use the regression equation which can be calculated if the number of tasks available with known HEPs is sufficiently large. The regression line and equation which was calculated for all 18 data points is shown in Figure 4.2 in Volume I, which is reproduced for convenience as Figure 2.1 in this Chapter. It is apparent that the quality of the predictions made from a regression equation approach will depend on several factors. The first is the degree of scatter of the points around the regression line, which is a measure of the judges'

consistency in generating SLI estimates. The size of the sample used to generate the regression line is also important, and the degree to which the sample typifies the population for which the regression equation will be used for prediction. In the present experiment, the population of data can be defined as the 18 tasks which were assessed. The regression equation calculated from all these tasks can be used to calculate an estimate of the log HEPs from the original SLI values generated by the judges. These estimates represent the normative case, because they are calculated from the largest and most representative sample (the entire data set). The differences between these estimates and the empirical log HEPs are due to the scatter of the SLI estimates around the regression line.

The estimates and their associated confidence limits are compared in Table 2.3 with the empirical log HEP estimates together with their confidence limits. The confidence limits for the empirical data points were calculated from the normal approximation to the binomial distribution. The error probabilities for the tasks used in this study were calculated from much larger denominators than would be available for the high risk scenarios considered in nuclear power plant PRAs. Hence, the 95% confidence limits around the estimates are quite small.

Table 2.3 Comparison of Empirical and SLIM Estimated Error Probabilities for 18 Tasks.

			log p (error) Actual			log p (error) Estimated		
Task	Numerator	Denominator	UB	X	LB	UB	X	LB
1	47	1368	-1.50	-1.50	-1.50	-0.93	-1.29	-1.65
2	53	272	-0.70	-0.71	-0.72	-1.14	-1.49	-1.84
3	114	12000	-2.02	-2.02	-2.02	-1.60	-2.08	-2.56
4	596	9306	-1.19	-1.19	-1.19	-1.63	-2.12	-2.61
6	Not Known	Not Known	?	-4.30	?	-2.10	-2.95	-3.80
7	492	16800	-1.53	-1.53	-1.53	-1.53	-1.97	-2.41
9	13	80	-0.77	-0.79	-0.81	-0.74	-1.13	-1.52
10	17	2631	-2.19	-2.19	-2.19	-0.80	-1.18	-1.56
11	8	2631	-2.51	-2.52	-2.52	-1.72	-2.27	-2.82
13	11	2631	-2.37	-2.38	-2.38	-0.77	-1.15	-1.53
14	6	207	-1.51	-1.54	-1.56	-1.38	-1.77	-2.16
15	13.5	133	-0.97	-0.99	-1.01	-0.77	-1.15	-1.53
16	9	140	-1.17	-1.18	-1.22	-0.40	-0.86	-1.32
17	22	300	-1.12	-1.13	-1.15	-1.24	-1.60	-1.96
18	22	64	-0.45	-0.47	-0.48	-0.32	-0.80	-1.28
19	47	160	-0.52	-0.53	-0.54	-1.28	-1.64	-2.00
20	23	36	-0.18	-0.19	-0.21	+0.84	+0.04	-0.80
21	2	10	-0.56	-0.70	-0.91	-0.04	-0.60	-1.16

With the HEPs estimated by SLIM, the largest 95% confidence interval is 1.7 log units and the smallest 0.7 log units. For 11 out of the 18 tasks, the HEP estimates generated by SLIM include the empirical HEP point estimates within the 95% confidence interval. Another way to assess the accuracy of SLIM is to consider the degree to which the estimated mean HEPs fall within the range of one order of magnitude about the empirical HEP means. In 12 of the 18 tasks, the SLIM estimates reach this level of precision, which is generally considered to be adequate for PRA purposes. Given the inexperience of the judges and the use of equal weights, these results must be regarded as being reasonably promising.

It may be argued that the level of precision attainable by this procedure is unrealistic, because in a real application the regression equation would have to be estimated by a subset of calibration tasks. We can simulate this situation by randomly choosing samples from the existing data set and calculating the regression line in each case. Figure 2.2 illustrates this using 10 random samples of 12 tasks.

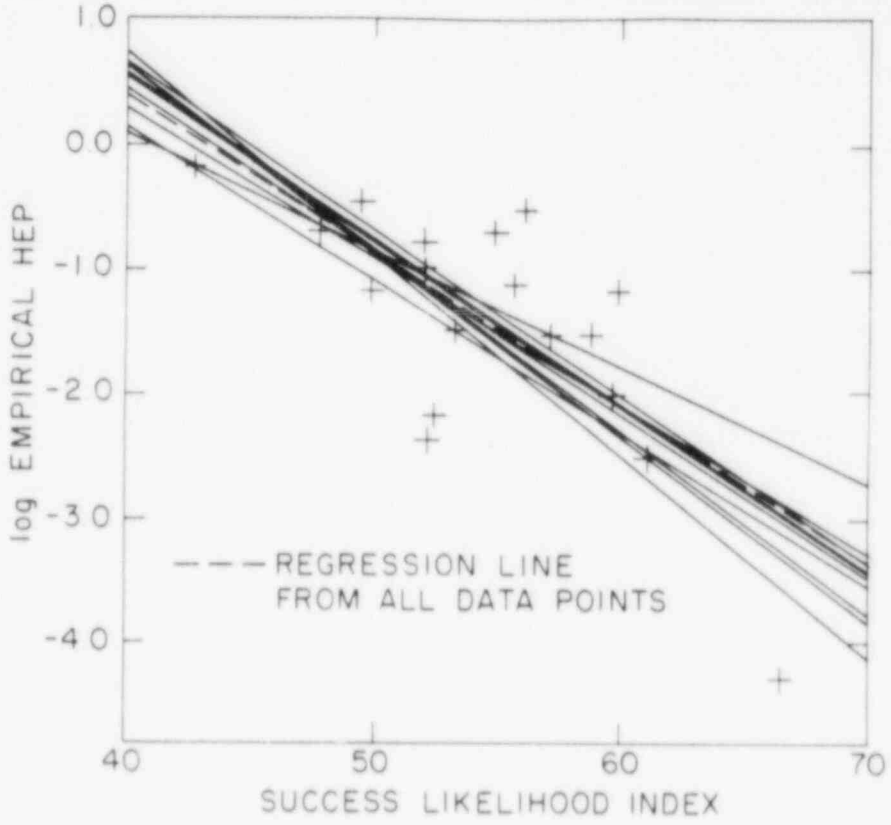


Figure 2.2 Regression lines generated by random selection of 12 calibration points from a data set of 18.

It can be seen that if the SLIs were converted to probabilities using one of the subset regression lines, the results would be very similar to using the global regression line. Obviously, the less the scatter of the original data points about the regression line, the closer the approximation of any regression line fitted through a subset of the data to the general regression line.

2.1.4 Conclusions Regarding the Phase I Experimental Study

Among the insights produced by the study using basic SLIM, the first was the critical importance of an effective classification scheme. Detailed procedures for the development of such a scheme are given in Chapter 4 of this volume.

Although it was necessary to calculate the SLIs without the weighting data, the study demonstrated that the assumption of a logarithmic relationship between SLIs and HEPs was a reasonable one. This result provides increased confidence for the assumption that the logarithmic relationship may be employed generally for the conversion of SLIs to HEPs.

The SLI values obtained in the experiment were converted to log HEPs using the regression equation obtained from the overall data set. Given the lack of experience of the judges, and the use of rating data alone, the log HEPs obtained by this method showed a reasonably good correspondence with the empirically determined HEPs for most of the tasks considered.

Although this exercise cannot in any sense be regarded as a validation study, it further confirms the viability of SLIM as an approach to human reliability assessment.

2.2 Phase II Research: Field Study Using Basic SLIM

An overview of the field study utilizing the basic SLIM technique has already been presented in Section 5.2 of Volume I. This study involved the assessment of eight critical human actions in five severe accident sequences for two BWRs and two PWRs. In this section, this study will be considered in more detail.

The pool of judges used for the field study consisted of 12 individuals including PRA specialists, a human factors engineer, a thermohydraulics expert, and simulator trainers who had experience in some of the plants being assessed.

The seven PSFs used in the study, defined and described in Appendix A, were as follows:

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

The three design dimensions considered in the first PSF were the quality of the displays, the degree to which operators were involved in design modifications, and the automation of routine functions. Meaningfulness of procedures was assessed in terms of their realism, the provision of location aids to indicate the position of items referred to in the procedures, the extent to which they kept the operators in touch with the plant, and the degree of operator involvement in their preparation. An additional dimension of "comprehensibility" was added by the assessment group.

The role of operations was a global PSF referring to the degree to which operations was prominent. Typical dimensions included the amount of paperwork and the relationship of operations to other departments in the plant, e.g., maintenance. The team's PSF included dimensions such as the existence of shifts which allowed teams to stay together and clearly defined operational roles. Stress was defined in terms of the resources available to the crew to meet demands, as affected by time constraints and the effects of shift systems on individuals. Conflicts between safety and availability goals were also regarded as producing negative stress.

Morale and motivation were regarded as a function of the professional status of the operating team, and the existence of a career structure. The competence PSF was assessed in terms of the amount of appropriate training received together with the extent to which an effective certification process existed.

Five plant-specific scenarios involving eight operator actions were evaluated in the field study. The scenarios and actions are summarized in Table 2.4 Because actions 7 and 8 were evaluated within a single scenario (i.e., for the same nuclear plant), judges believed the PSF weights and ratings would be identical for each action. As a result, the calculated SLI value for both actions is also identical. However, different HEPs were obtained by using different best and worst case estimates in each scenario. Thus, although eight actions were evaluated, only seven distinct data points are available and this is reflected in the various analyses performed.

Several aspects of this study are of particular methodological interest and will therefore be discussed in more detail.

2.2.1 Comparison of Alternative Aggregation Procedures

Two procedures for aggregating the individual judgments to arrive at overall HEP values for each human action were used. In the first method, individual SLIs were derived from the individual weights and ratings for the seven PSFs. These were converted to log HEPs using each individual's absolute probability estimates of the best and worst case HEPs to derive separate calibration equations for each scenario. The resulting individual estimates of the log HEP for the action being evaluated were then aggregated by taking their geometric mean to arrive at the overall log HEP. (This is the

Table 2.4 Scenarios Evaluated and Critical Actions Quantified in SLIM Field Study.

Scenario				
No.	Reactor*	Accident Sequence Code	Accident Description	Operator Actions
1	PWR1	S_2H	Small-break loss-of-coolant accident with failure of emergency core coolant system recirculation.	1. Operator fails to respond to low-low refueling water storage tank (RWST) alarm. 2. Operator fails to recover above failure.
2	BWR1	TQUV	Transient event initiated by loss of off-site power with failure of all reactor inventory makeup.	3. Operator fails to use automatic depressurization system (ADS) and low pressure injection (LPI) following loss of high pressure injection.
3	PWR2	S_2H	Small-break loss-of-coolant accident with failure of emergency core coolant system recirculation.	4. Operator fails to prevent refueling water storage tank (RWST) from emptying before recirculation is achieved. 5. Operator fails to recover above failure by depressurization and low pressure recirculation (LPR).
4	BWR2	TC	Transient initiating event with failure to achieve reactor subcriticality.	6. Operator fails to recover from anticipated transient without Scram (ATWS).
5	BWR2	TW	Transient iniating event with failure of residual heat removal (RHR) system to remove heat from suppression pool.	7. Operator fails to recover when main steam isolation valve (MSIV) is isolated but power conversion system (PCS) is available. 8. Operator fails to recover when power conversion system (PCS) is lost but control rod drive mechanism (CRDM) is available.

*Judges evaluated the five scenarios with reference to specific operating reactors. Names of specific reactors are not presented here for proprietary reasons.

method for aggregating log HEPs; see Seaver and Stillwell, 1983.) The required HEP was then derived by taking the antilog. This procedure was repeated for each action being evaluated.

The alternative procedure was carried out during the SLIM session itself. The arithmetic means of the individual weights and ratings for the seven PSFs were obtained and these were combined to give an overall SLI. Using the consensus absolute judgments for the best and worst case HEPs as calibration points, the HEP value for the action being assessed was derived.

It will be apparent that although the first of these procedures involves purely mathematical aggregation, the second is only a consensus procedure to the extent that it involves the use of consensus values for the calibration points. A correlated t-test was performed to investigate whether the HEPs

derived by these two procedures differed significantly. The obtained t value ($t = 0.210$) indicates no significant difference between the HEPs derived by these two methods (for significance at $p < 0.01$, t must exceed 3.143 at 6 d.f.).

2.2.2 Inter-judge Consistency

Although a total of 12 judges participated in assessing tasks, only 3 judges took part in all SLIM sessions. It was therefore only possible to calculate inter-judge consistency measures for these three judges. The procedure described in Seaver and Stillwell (1983) was used to calculate inter-judge consistency. This first involved carrying out a two-way analysis of variance (ANOVA) using the individual log HEPs as the dependent variable and actions evaluated and the judges as the factors. The resulting ANOVA table is shown below:

Source	SS	df	MS	F Ratio
Action evaluated (A)	27.06	6	4.51	15.24*
Judges (J)	0.10	2	0.05	0.169
A x J	3.55	12	0.296	--
Total	30.71	20	--	--

*$p < 0.001$

The results of this analysis indicate that most of the variability in the log HEPs is due to differences between the actions evaluated. There are no significant differences between judges.

The intraclass correlation coefficient, representing the average correlation between the estimates of each pair of experts, can also be calculated as follows:

$$r = \frac{F - 1}{F + (n + 1)} = \frac{15.24 - 1}{15.24 + 5} = 0.6704$$

This result approaches significance ($p = \leq .10$, d.f. = 5) and indicates moderate agreement between judges.

2.2.3 Uncertainty Bounds

The uncertainty bounds on the HEPs derived from data of the three judges considered in the last section were obtained using the method described in Section 2.3.6. The average uncertainty (± 2 standard errors) about the log HEP estimates was 1.04 log units, the range being 2.14 to 0.34. An uncertainty of one order of magnitude would generally be regarded as acceptable for PRA purposes. Only one estimate exceeded this criterion.

2.2.4 Sensitivity Analysis

To make design recommendations, it is important to be able to identify which PSFs judges perceive to have the greatest effect on the probability of success or failure. SLIM's ability to provide this information is an important advantage over other approaches.

A two-way analysis of variance was performed using the PSF weights as the dependent variable and the PSF categories and the actions evaluated as the two factors. The results of this analysis are reproduced below:

Source	SS	df	MS	F Ratio
PSF (P)	26,360.81	6	4,393.47	31.57*
Actions evaluated (A)	859.94	6	143.32	1.03 N.S.
[P x A]	5,009.52	36	139.15	--
Total	32,310.15	48	--	--

*p <.001

The ANOVA suggests that there are significant differences between the importance weights assigned to the different PSFs. To investigate these differences further, multiple comparisons were carried out using Scheffe's method.

This indicated the following statistically significant differences in the importance weights:

- Competence was perceived to be more important than design, stress, morale, and role of operations.
- Competence, teams, and procedures were all perceived to be more important than stress, morale, and the role of operations.
- Design was more important than morale and role of operations.
- Stress was more important than role of operations.

All these differences in importance are statistically significant. The ranking of the various PSFs in terms of importance was as follows:

PSF	Mean Weight	Normalized Mean Weight
Competence	93.80	0.20
Teams	86.91	0.19
Procedures	85.71	0.19
Design	68.47	0.14
Stress	58.24	0.13
Morale	35.80	0.08
Role of operations	31.60	0.07
	Σ = 460.53	Σ = 1.00

This analysis indicates that, in terms of cost-effectiveness, the greatest improvements in HEPs for all the actions considered would be gained by investing resources in the areas of training (to improve competence), followed by providing for effective team structures and improvements in procedures.

The finding within the ANOVA that there was no significant interaction between the PSF categories and the actions evaluated indicates that the pattern of PSF weights did not differ significantly between the actions being evaluated. This suggests that it would have been possible to use a generic set of PSF weights in this case. However, this could not have been established a priori without the existence of a valid task classification scheme.

2.2.5 Analysis of Rating Data

The data were subjected to an analysis similar to that applied to the importance weights. In this instance, the PSF ratings are the dependent variables, with the PSF categories and the actions evaluated the two factors. The ANOVA table is given below:

Source	SS	df	MS	F-Ratio
PSF (P)	3780.39	6	630.07	5.70*
Action evaluated (A)	3993.21	6	655.54	6.02*
[PXA]	3981.37	36	110.59	--
Total	32,310.15	48	--	--

*$p<0.01$

Multiple comparison tests were carried out to investigate the nature of the PSF main effect. The ranking of the mean ratings is given below:

PSF	Mean Rating
Teams	72.68
Competence	69.26
Design	66.49
Procedures	64.16
Role of Operations	61.92
Morale	57.81
Stress	43.84

As might be expected for the scenarios being evaluated, the stress rating PSF is significantly lower (worse) than for the teams, competence, and design PSFs. The results also suggest that some attention should be paid to improving morale, although the importance weights analysis suggest that this does not have a large impact on the probability of success for the actions being assessed.

The existence of significant differences between the actions being evaluated indicates that the mean ratings, when all the PSFs are aggregated together, differ between the scenarios. The mean of the PSF ratings can be

regarded as a measure of the overall "quality" of the plants with regard to the scenarios under consideration. It should be noted, however, that the ranking of quality is not necessarily identical to the ranking of likelihood of success for the actions being evaluated, since the ratings are subsequently combined with the importance weights to give the Success Likelihood Indices (SLIs).

2.2.6 Conclusions from the Field Study

The analyses discussed in the preceding sections indicate the versatility of the SLIM technique. The numerical values obtained for the actions evaluated are summarized in Tables 2.5 and 2.6. As discussed in Section 5 of Volume I and Section 2.2.1 of this volume, the SLI values in this study were converted to HEPs using absolute probability judgments for the best and worst case conditions. Because the judges were not unanimous in their judgments of the best and worst case HEPs for these conditions, alternative estimates for the derived HEPs are presented in Table 2.5. In Table 2.6, single HEP estimates are derived for all the scenarios by taking the geometric mean of the absolute probability judgments for the best and worst case HEPs where complete consensus was lacking. Table 2.6 also contains 95% confidence limits estimated by the techniques discussed in Section 1.10.6 of this volume.

In addition to HEP estimates, the technique produces a wide variety of other outputs which could be used for systems analysis and design purposes. For example, the weights assigned to the PSFs for each action evaluated indicate the relative importance of design factors such as training and procedures in reducing the likelihood of error. This could be used to evaluate the merit of different design solutions to improving human reliability.

2.3 Overview and Discussion of Phase I and Phase II Research

The results of the Phase I experimental study provided support for the assumed logarithmic calibration function between SLIs and HEPs. It also demonstrated that even with inexperienced judges, SLIM was capable of generating HEPs comparable to empirically determined HEPs for a high proprotion of the tasks considered in this exercise. It seems reasonable to assume that these results could be improved upon if reliable weighting data were available and if the judges were experienced in the use of the technique.

The Phase II study indicated that the technique could be used as part of a PRA study for evaluating HEPs for critical actions in nuclear power plant accident sequences. The technique appeared to have a high degree of user acceptability, and the judges felt they had gained new insights into the nature of the human actions which could impact on safety in the scenarios considered.

Phases I and II constituted a learning period in the development of SLIM and indicated not only SLIM's strengths, but also areas where the basic

Table 2.5 Summary of SLIM Quantification Results.

Scenario	Operator Action	Best Case Pf	Worst Case Pf	SLI	HEP
1. PWR 1	1. Fails to respond to low-low RWST alarm.	10^{-4}	10^{-2}	76	3.0×10^{-4}
	2. Fails to recover from failure.	10^{-3} 5×10^{-2}	5×10^{-2} 5×10^{-1}	64 64	9.5×10^{-3} 1.2×10^{-1}
2. BWR 1	3. Fails to use ADS and LPI following loss of off-site power and HPI.	10^{-4} 10^{-5}	10^{-2} 10^{-2}	65 65	5.0×10^{-4} 1.0×10^{-4}
3. PWR 2	4. Fails to prevent RWST emptying before recirculation achieved.	5×10^{-3}	2×10^{-1}	70	1.5×10^{-2}
	5. Recover from failure.	5×10^{-2}	5×10^{-1}	46	1.7×10^{-1}
4. BWR 2	6. Failure to recover from ATWS.	5×10^{-2}	5×10^{-1}	55	1.4×10^{-1}
5. BWR 2	7. Failure to recover TW when MSIVs isolated but PCS available.	10^{-4} 10^{-6}	10^{-3} 10^{-3}	70 70	2.0×10^{-4} 8.0×10^{-6}
	8. Failure to recover when PCS lost but CRDM available.	10^{-2}	10^{-1}	70	2.0×10^{-2}

Table 2.6 HEP Values and Confidence Limits From SLIM Using Geometric Means of Nonconsensus Boundary Conditions.

Scenario	Operator Action	HEP	95% Confidence Limits	
1. PWR 1	1. Fails to respond to low-low RWST alarm.	3.0×10^{-4}	1.2×10^{-4}	7.6×10^{-4}
	2. Fails to recover from failure.	3.3×10^{-2}*	1.2×10^{-2}	9.1×10^{-2}
2. BWR 1	3. Fails to use ADS and LPI following loss of off-site power and HPI.	2.4×10^{-4}*	4.8×10^{-5}	1.2×10^{-3}
3. PWR 2	4. Fails to prevent RWST emptying before recirculation achieved.	1.5×10^{-2}	6.9×10^{-3}	3.5×10^{-2}
	5. Recovery from failure.	1.7×10^{-1}	1.2×10^{-1}	2.6×10^{-1}
4. BWR 2	6. Failure to recover from ATWS.	1.4×10^{-1}	1.2×10^{-2}	1.0
5. BWR 2	7. Failure to recover TW when MSIVs isolated but PCS available.	4.0×10^{-5}*	1.2×10^{-5}	1.4×10^{-4}
	8. Failure to recover when PCS lost but CRDM available.	2.0×10^{-2}	5.8×10^{-3}	6.9×10^{-2}

*Obtained by taking geometric mean of boundary conditions.

technique appeared to have weaknesses. In particular, it appeared that a more systematic method for eliciting PSFs from judges was needed. Difficulties were also encountered in ensuring that the PSFs utilized were truly independent as is required by the additive MAUT model underlying SLIM.

Another area of concern was the question of how to deal with PSFs such as stress, where the effects on the likelihood of success are not linear (i.e., the likelihood of success would probably be degraded by very high or very low values of stress, but facilitated by moderate levels). Finally, it appeared that a more sophisticated approach to deriving the PSF weights was necessary to ensure that a common baseline was used for the evaluation of all the actions in a set.

These problems can be solved with the MAUD implementation of SLIM using the MAUD technology (SLIM-MAUD) as discussed in Chapter 1. Chapter 3 provides a detailed description of the procedures for using MAUD, together with an example of a typical SLIM-MAUD session. It is recommended that, wherever possible, SLIM be implemented using MAUD in future applications of the approach.

3. STAND-ALONE PROCEDURES FOR IMPLEMENTING SLIM WITH MAUD5

3.1 Obtaining MAUD5

MAUD5 is software proprietary to, and marketed by, the Decision Analysis Unit, London School of Economics. Versions of MAUD5 exist for a wide variety of microcomputers. The microcomputer configuration needed to run MAUD requires a minimum of 64K random-access memory, two 5-1/4-inch floppy disks (or one floppy disk and a hard disk), and a printer. It must either run under the CP/M operating system (and accept standard CP/M files) or conform to IBM/PC DOS standards.

MAUD5 may be obtained by the Decision Analysis Unit by completing the form in Appendix D, which requests specific information about the computer on which MAUD will be run. The user is required to sign an end-user license for MAUD5 (details of this license are given in Appendix D). The license is valid indefinitely and the purchase price reflects the number of copies of MAUD5 the end user may have in operation at any one time.

On receipt of assigned end user license, the Decision Analysis Unit will supply the user with a 5-1/4 inch diskette containing the complete set of compiled programs and system files which comprise MAUD5. The end user should treat this disk as a master disk. The first step on receipt of the disk should thus be

- Format a blank disk on the microcomputer
- Copy the computer's operating system to this disk
- Copy the whole of the master disk to this disk

This results in the MAUD5 working disk, ready for use.

3.2 Configuring MAUD5 to Implement SLIM

MAUD5 is supplied as a general purpose system for aiding expert judges in making assessments and decisions. To implement SLIM using MAUD5, the program M5CONFIG supplied with MAUD5 must be used to change the default text used within MAUD to that required to implement SLIM. M5CONFIG need only be run when first installing MAUD5 for implementation. The SLIM-MAUD text will then be presented to the user each time MAUD5 is run (until and unless M5CONFIG is run to change the text again).

To run M5CONFIG, proceed as follows:

- Insert the MAUD5 working disk in drive A of your computer
- After you have loaded the computer's operating system the following prompt will appear:

  A >
  Type M5CONFIG followed by pressing the RETURN or ENTER key.

M5CONFIG will then proceed to guide you interactively through the configuration procedure. The text below shows this procedure frame-by-frame when implementing SLIM-MAUD. All user inputs are shown underlined to distinguish them from the text generated by M5CONFIG.

o Frame 1

Do you want to re-configure the screen control functions for MAUD5 on your console? NO

(Please type YES or NO, using the keyboard like a typewriter, and then press the key marked return. If you prefer, you can type Y for "YES" and N for "NO." Make sure that the "caps lock" key is depressed on the keyboard as MAUD will not accept lower case command characters.)

[Answering "Y" to this question is only necessary when moving MAUD to a computer with a different type of VDU from that for which MAUD5 was supplied to the end user.]

• Frame 2

Do you want to revise the text used within MAUD5? Y. You can:

(1) Revise the current text
(2) Restore the default text
(3) Revise the default text.

Which would you like to do?
Please type your option number ; 3

[If at some future time revisions need to be made to the text implementing SLIM-MAUD, then option 1 can be used instead of option 3.]

• Frame 3

MAUD5 starts with a 15-line introductory frame. At present, it appears like this;

```
>
>
>
>
>
>
>         This computer is set up to help you think about
>                   a decision you wish to make.
>
>
>
>
>
>
>
```

Do you wish to change this? Y

- Frame 4

Please type in your new text (a maximum of 15 lines)

>This computer is set up to help you assess the likelihood of success of
>various courses of action in the situation you wish to consider.
>
>The assumption of the success likelihood index methodology (SLIM) used here
>is that the likelihood of task success is a function of certain key factors
>in the situation, known as Performance Shaping Factors (PSFs).
>
>Here, SLIM is implemented through MAUD5, a system which helps you identify
>the important PSFs on which the courses of action are to be assessed.
>
>MAUD5 also provides techniques for identifying the relative importance of
>these factors and for assessing the SLI for each course of action. Each SLI
>can be transformed into a probability of success provided you can supply
>probability of success assessments for two reference courses of action among
>the set of actions you have been considering.

Is this frame OK? Y

[Each of the 12 lines of text to the right of > were typed in by the user.]

- Frame 5

Currently, MAUD5 aims to help the user in
>making your decision

Do you wish to change this? Y
Please type in your new text
>determining likelihood of success

- Frame 6

The user is expected to develop his or her
>preferences

Do you wish to change this? Y
Please type in your new text
>assessments of likelihood of success

- Frame 7

The alternatives are rated on dimensions, each of which is
currently called an
>aspect

Do you wish to change this? Y
Please type in your new text
>performance shaping factor

- Frame 8

  Currently, the user is asked to define his or her own name
  for the set of alternatives under consideration

  Do you wish to change this? Y

  Do you wish the user to define his or her own input? N

  Please supply a new name for the set of alternatives
  In singular form, each alternative is a >course of action
  In plural form, all alternatives are >courses of action

- Frame 9

  The user is given an example of a performance shaping factor
  where some alternatives are characterized as
  >interesting
  and other alternatives are characterized as
  >boring

  Do you wish to change this? Y

  Please supply your new example:
  Some alternatives should be characterized as >simple
  and other alternatives should be characterized as >complex

- Frame 10

  Numerical values assigned by MAUD5 to alternatives are currently called
  >preference values

  Do you wish to change this? Y

  Please type in your new text
  >success likelihood indices

- Frame 11

  The verb describing the basis for choosing an alternative
  is to >prefer

  Do you wish to change this? Y

  Please type in your new text
  >assess as most likely to succeed

- Frame 12

  MAUD5 currently asks for the title for the session with the question
  WHAT IS > YOUR NAME

  Do you wish to change this? Y
  Please type in your new text
  WHAT IS > THE TITLE FOR THIS SESSION?

- Frame 13

  Are all the changes you have made to the text OK? Y

[At this point, the configuration of MAUD5 to implement SLIM is complete. M5CONFIG saves all the complete changed text on the MAUD5 diskette and returns the user to the operating system which will return the prompt:]

A>

[Sessions with MAUD5 implementing SLIM may be run now, or at any subsequent time.

The above configuration sets up MAUD to implement SLIM using text in accord with the protocols defined in the Test Plan described in Section 4 of this volume. These protocols may be changed as a result of practical experiences with MAUD implementations of SLIM during the test plan and subsequently. Advice on the most appropriate configuration of MAUD to implement SLIM in any particular context will continue to be available at all times from Human Reliability Associates, Ltd.]

3.3 Ascertaining Tasks to be Analyzed

The implementation of SLIM using MAUD has the advantage that MAUD provides interactive guidance for users unfamiliar with the system. The only prerequisite for using MAUD in a SLIM-MAUD session is that the user has ascertained a homogeneous set of tasks for which he or she wishes to assess SLI values. A set of tasks is considered "homogeneous" if the successful performance of each task in the set depends upon a common set of PSFs. Homogeneity of tasks can be determined in practice in one of two ways: (1) by reference to a pre-constructed task taxonomy or (2) by grouping of tasks according to direct judgments of homogeneity by the user. Under option 1, the judge has simply to ensure that the tasks selected for assessment in the SLIM-MAUD session share common features which place them all in the same cell of a pre-constructed task taxonomy. If the complete set of tasks to be assessed falls into more than one cell in this taxonomy, then the required procedure is to form subsets of tasks, each comprised of all the tasks that were classified into a particular cell. The user then proceeds to use the MAUD implementation in assessing each subset in turn. This sequence is summarized in Figure 3.1.

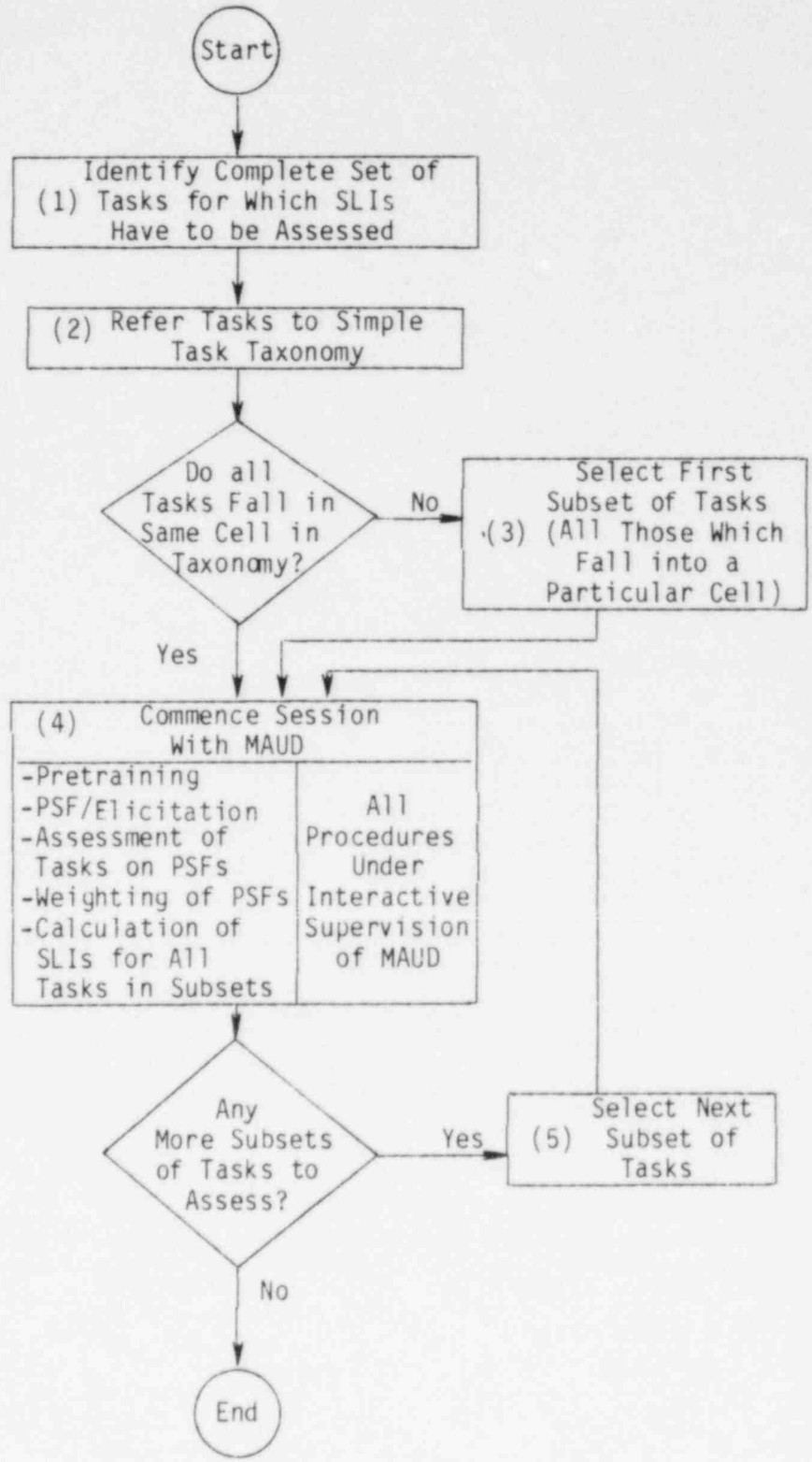


Figure 3.1 Outline of procedure for assessment of SLIs for a set of tasks in a MAUD-based implementation of SLIM.

The Test Plan described in Chapter 4 provides methodology for the development of simple task taxonomies for the level A and B tasks described in Appendix C. This methodology may be used to develop a number of preconstructed task taxonomies, but there are still likely to be many instances where the judge is confronted with a set of tasks to assess for which no preconstructed task taxonomy exists.

In this case the recommended procedure, replacing step 2 in Figure 3.1, is as follows:

1. Copy a short description of each task situation onto a file card (4"x6" or 5"x7").

2. Present the complete set of cards to the SLIM-MAUD user with the following instructions:

   In order to use MAUD to implement SLIM, the tasks to be assessed must first be sorted into a number of groups, with each group comprising not less than 4, and not more than 10 tasks. Sort the tasks so that the likelihood of success of performance of any and all of the tasks you sort into any one group depends on their ratings on a common set of performance shaping factors. It is not necessary to identify these factors at the present time. Please examine all the tasks on these cards (give the user the set of cards) and sort them into groups on the table in front of you, arranging and rearranging the groups until you are satisfied that the tasks sorted into each group meet this criterion.

3. Allow the user to develop the subsets of tasks to be assessed within MAUD through following these instructions. When he or she is satisfied with the final grouping of tasks, take each subset in turn and assess them as a group through the use of MAUD (step 4 in Figure 3.1).*

The example given in Section 3.4 (below) of the use of MAUD to implement SLIM on a subset of tasks was based upon the prior division of the set of 15 level A tasks given in Appendix C into three subsets by the user, following the instructions given above. The subsets identified were as follows:

- Task subset I, comprising tasks 1, 2, 3, 10.
- Task subset II, comprising tasks 4, 5, 6, 7, 8, 15
- Task subset III, comprising tasks 9, 11, 12, 13, 14.

---

*If the user isolated a group of less than four tasks, then reference tasks or variants of the tasks specified in this "isolated" group, which meet the homogeneity criterion, may be added to increase the number of tasks to be assessed through the use of MAUD to four or more.

Section 3.4 describes the procedure used within step 4 of Figure 3.1 for subset I of the above three subsets. Descriptions of the tasks on this subset are given in Table 3.1. A complete assessment of all 15 level tasks described in Appendix C would require that step 4 be executed three times (i.e., once for each of the subsets.)

Table 3.1 Subset I of Level A Tasks Selected by the "SLIM-MAUD" User.

Task I: EDS MAN ACT

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?

Task 2: RCIC MAN

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?

Task 3: NIS INSERT

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 decreased following the scram. What is the likelihood that the operator will fail to actuate the nuclear instrumentation system correctly?

Task 10: BLK OUT MALOP

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?

3.4 Using MAUD to Implement SLIM

3.4.1 Overview of an Assessment Session: MAUD-directed Interaction

MAUD takes control of all the interaction with the user required to execute the activities shown in Step 4 of the sequence shown in Figure 3.1. MAUD is user friendly: it provides keyboard training to teach the user how to enter data, and it provides menus at each option with clear descriptions of operations in common language to guide the user efficiently through the procedure and any options he or she wishes to select. Most importantly, MAUD has comprehensive facilities for error detection, and informative corrective procedures. If a user makes an obvious mistake (e.g., selects a nonexistent option, enters a data value out of range, etc.) then MAUD explains the mistake to the user and asks for a correction. If a conceptual mistake is made (e.g., a PSF is identified which the user later wishes to delete, a data value is assessed which the user later wishes to change, or a set of PSF weights is assessed, but the user is not satisfied with the results, etc.) MAUD's reentrant editing procedures can be used to restructure any or all aspects of the work completed to correct this conceptual mistake. MAUD will actively assist in editing and correcting a conceptual mistake at the moment the user becomes aware of it.

In addition, MAUD continuously monitors the coherence of the user's performance in the session. So long as all goes well, this monitoring is invisible to the user, but as soon as a coherence problem is detected, MAUD intervenes, inviting the user to think about the problem it has identified in his or her performance, and specifying the alternative options which can be selected to put things right and continue the session. In this way MAUD will usually spot the conceptual mistakes which the user has not detected.

Hence MAUD provides good guidance and training facilities for the naive user. Validation studies of MAUD (e.g., John, Von Winterfield and Edwards, 1983) indicated that these facilities were approximately on a par with those that could be provided by the continual presence throughout the session of a facilitator skilled in use of the technique.

3.4.2 Training Required of Users

The procedure followed in the MAUD implementation of SLIM requires minimal pretraining. MAUD itself undertakes the training by interactively guiding a naive user through the program's procedures.

However, like any other assessment methodology, SLIM follows the rule "garbage in = garbage out." Therefore, it is presumed that the SLIM-MAUD user will be expert with respect to knowledge of the relevance and relative importance of PSFs and the assessment of tasks on these PSFs. The aim of the SLIM-MAUD procedure is to capture this expertise in an efficient and coherent way as a basis for determining the SLIs of the set of tasks being assessed.

3.4.3 Requirements for Running a Session to Implement SLIM Through MAUD5

The judge must be supplied with a computer configuration which meets the specification given in Section 3.1. The MAUD5 working disk must have previously been configured to implement SLIM, following the instructions given in Section 3.2. A formatted disk which will hold the results of the session (SLIs, rating on PSFs, etc.) must also be available. This disk will be referred to in the following discussion as the "SLIM-MAUD data disk." The end user may adopt whatever policy he or she wishes concerning the brand of diskette and procedures for cataloging them. They can be formatted as required using the standard procedures for disk formatting or the end user's microcomputer sessions. Each data disk can hold the results of between 15 and 60 SLIM-MAUD scenarios (depending upon the microcomputer's disk drive capacity). Each session is given a name (which should be unique if a new record is required on the SLIM-MAUD disk), and the session record can be recalled at any time for subsequent use in interaction with MAUD. Copies of these disks may be deposited in the SLIM data bank described in Section 1.14.2.

3.4.4 Instructions for Starting a Session Implementing SLIM Through MAUD5

Users must have passed through the procedures in Figure 3.1 and reached Step 4. They are then ready to commence the session with MAUD5. The following instructions describe how to start the session.

- Insert the MAUD5 working disk in drive A of your computer.
- Insert a SLIM-MAUD data disk in drive B of your computer.
- Make sure that the printer and VDU are properly connected to the microcomputer and are on line.
- After you have loaded the computer's operating system, the following prompt will appear:

  A>

- Respond by typing MAUD5 on the keyboard. MAUD5 will then guide you through the SLIM assessment.

3.4.5 Example of a SLIM-MAUD Session

MAUD is a stand-alone set of procedures, with the details of the steps in the procedures self-contained within the MAUD program. Thus, rather than present a further description of these procedures here, it is more useful to provide a detailed example of MAUD's frame-by-frame interaction with a user. Accordingly, the remainder of this chapter reproduces a sequence of frames (VDU images) which constituted a session with MAUD, with minimal commentary.

The first frame shown is that which appears on the VDU screen directly after the user, following the instructions in Section 3.4.4, has responded:

A>MAUD5

and pressed RETURN. In this frame, and in all subsequent frames in the session, any text input by the user has been underlined for purposes of clarity (it was not underlined on the original VDU image). All other text was supplied by MAUD. As the system is dynamically interactive, often developing a frame in interaction with the user before moving to the next, a static sequence of frames does not give a very good impression of the way MAUD operates and, as a linear sequence, does not illustrate the range of branching and re-entrant options available. However, it does give some idea of the type of dialogue employed when implementing SLIM.

- Frame 1

```
**********************************<MAUD5>**********************************
--------------------------------------------------------------------------------
>This computer is set up to help you assess the likelihood of success of
>various courses of action in the situation you wish to consider.
>
>The assumption of the success likelihood index methodology (SLIM) used here
>is that the likelihood of task success is a function of certain key factors
>in the situation, known as Performance Shaping Factors (PSFs).
>
>Here, SLIM is implemented through MAUD5, a system which helps you identify
>the important PSFs on which the courses of action are to be assessed.
>
>MAUD5 also provides techniques for identifying the relative importance of
>these factors and for assessing the SLI for each course of action. Each SLI
>can be transformed into a probability of success provided you can supply
>probability of success assessments for two reference courses of action among
>the set of actions you have been considering.
>
>Have you used the MAUD system before? NO

(Please type YES or NO, using the keyboard like a typewriter, and then press
the key marked RETURN. If you prefer, you can type Y for "YES" and N for
"NO.")
```

• Frame 2

MAUD5 will work with you, using the words you give it.
It will ask you questions which are relevant
in determining likelihood of success.

It will allow you to make changes in your description of the alternatives under consideration and your assessment of likelihood of success as you wish.

When MAUD asks you a question, for example:

WHAT IS THE TITLE FOR THIS SESSION?

You should type in your answer from the keyboard.
Enter your answer now, and then press the key marked RETURN

LEVEL A TASKS SUBSET I

• Frame 3

The title for this session will be LEVEL A TASKS SUBSET I

If you make a typing mistake, there are two ways of making corrections:

- If you notice the mistake before you press the RETURN button, you can correct it by pressing the backspace button, which rubs out the characters one at a time. You can then type the correct characters, followed by pressing RETURN.

- If you notice the mistake after you have pressed the RETURN button, you will have to wait until the next opportunity MAUD5 offers you to make changes. However, these opportunities occur quite frequently.

Press RETURN to continue.

• Frame 4

Are you starting this MAUD5 session from scratch
(rather than starting with data on file from a previous session)? Y

- Frame 5

  Please think about the courses of action which are important in determining likelihood of success

  You must consider not less than 4 and not more than 10 courses of action within the set of courses of action under consideration. The name you give to each course of action can be whatever you like, so long as you are clear to what it refers.

  You may wish to consider not only courses of action which are immediately available, but others which may become available, or those which are useful for preference purposes.

  Please keep the description of each course of action short; type just one or two words.

- Frame 6

  Please type in the name of a course of action you want to consider

  Its name is TASK 1: EDS MAN ACT

  Now the next course of action you want to consider

  Its name is TASK 2: RCIC MAN

  Now the next course of action you want to consider

  Its name is TASK 3: NIS INSERT

  Now the next course of action you want to consider

  Its name is TASK 10: BLK OUT MALOP

- Frame 7

  Is there another course of action you want to consider? N

- Frame 8

  The courses of action currently under consideration are:

  (1) TASK 1: EDS MAN ACT
  (2) TASK 2: RCIC MAN
  (3) TASK 3: NIS INSERT
  (4) TASK 4: BLK OUT MALOP

  Do you want to change anything? N

- Frame 9

You are now going to be asked about differences
between courses of action. Try to think about differences
which are important to you in determining likelihood of success.
For instance, some people feel that certain courses of action
are simple while other courses of action are complex
and some courses of action are in-between.

This is just one example and it may not be relevant to you.
There are no right or wrong answers; it is important to try
to include those differences which you
think are important in determining likelihood of success
within the particular situation you are considering.

Press RETURN to continue.

- Frame 10

Can you think of a performance shaping factor in which one of these
courses of action

(1) TASK 2: RCIC MAN
(2) TASK 3: NIS INSERT
(3) TASK 10: BLK OUT MALOP

is different from the other two in a way that matters to you in
determining likelihood of success? (Please answer YES or NO) Y

What is the number next to the course of action that is different? 3

- Frame 11

You have said that TASK 10: BLK OUT MALOP is different from
TASK 2: RCIC MAN AND TASK 3: NIS INSERT
In not more than three words each time, please describe
how they differ.

First describe TASK 4: BLK OUT MALOP

TASK 10: BLK OUT MALOP IS: HIGHLY ABNORMAL

On the other hand,
TASK 2: RCIC MAN and
TASK 3: NIS INSERT are : FAIRLY NORMAL

Are you reasonably happy with this description? Y

- Frame 12

It should be possible to give each course of action
a rating from 1 to 9 according to its position
on the scale

```
HIGHLY ABNORMAL
 1                Your rating of    TASK 1: EDS MAN ACT      is 3
 2                Your rating of    TASK 2: RCIC MAN         is 5
 3                Your rating of    TASK 3: NIS INSERT       is 7
 4                Your rating of    TASK 4: BLK OUT MALOP    is I
 5   to
 6                                  Are these ratings OK ? Y
 7
 8
 9
FAIRLY NORMAL
```

[If at this point the user is dissatisfied with the ratings, MAUD provides the opportunity to edit the ratings or cancel the scale. Since in the above example the user was satisfied with the ratings, MAUD continued by eliciting the ideal point on the relevant PSF.]

- Frame 13

Thinking only about the performance shaping factor below, what position
on the scale would be IDEAL for a course of action
in the present context?

```
HIGHLY ABNORMAL
 1
 2                Your best possible value is : 9
 3
 4
 5  to            Is this rating OK? Y
 6
 7
 8
 9
FAIRLY NORMAL
```

[MAUD has not obtained sufficient information to rescale the ratings about the ideal point in this PSF, so it returns to the method of triad comparisons to obtain the second PSF.]

- Frame 14

Can you think of a performance shaping factor in which one of these courses of action

(1) TASK 1: EDS MAN ACT
(2) TASK 2: RCIC MAN
(3) TASK 3: NIS INSERT

is different from the other two in a way that matters to you in determining likelihood of success? (Please answer YES or NO) Y

What is the number next to the course of action that is different? 1

- Frame 15

You have said that TASK 1: EDS MAN ACT is different from TASK 2: RCIC MAN and TASK 3: NIS INSERT
In not more than three words each time, please describe how they differ.

First describe TASK 1: EDS MAN ACT

TASK 1: EDS MAN ACT is: HIGHLY STRESSFUL

On the other hand,
TASK 2: RCIC MAN and
TASK 3: NIS INSERT are: LESS STRESSFUL

Are you reasonably happy with this description? Y

- Frame 16

It should be possible to give each course of action a rating of 1 to 9 according to its position on the scale

HIGHLY STRESSFUL
1 Your rating of TASK 1: EDS MAN ACT is 3
2 Your rating of TASK 2: RCIC MAN is 5
3 Your rating of TASK 3: NIS INSERT is 7
4 Your rating of TASK 4: BLK OUT MALOP is 1
5 to
6 Are these ratings OK? Y
7
8
9
LESS STRESSFUL

- Frame 17

Thinking only about the performance shaping factor below, what position on the scale would be IDEAL for a course of action in the present context?

```
HIGHLY STRESSFUL
 1
 2                    Your best possible value is : 9
 3
 4
 5   to               Is this rating OK? Y
 6
 7
 8
 9
LESS STRESSFUL
```

- Frame 18

Your assessments of likelihood of success for the courses of action under consideration in terms of their ratings on the scale ranging from HIGHLY ABNORMAL TO FAIRLY NORMAL seem very much the same as your assessments of likelihood of success for the courses of action in terms of their ratings on the scale ranging from HIGHLY STRESSFUL to LESS STRESSFUL.

Does this mean that these two scales have very similar meanings for you? N

OK

Press RETURN to continue

[MAUD automatically checks for preference independence (in terms of promoting likelihood of success) between the ratings on each PSF, as it is entered, and all other PSFs already entered. In this case, MAUD discovers that the preference (I-scaled) ratings on the PSF just entered were very nearly the same as those on the first PSF entered (scaled from HIGHLY ABNORMAL to FAIRLY NORMAL). MAUD, having detected a possible violation of preference independence, asked the question shown in Frame 18 to see whether these two PSFs should be restructured to maintain preference independence.

The judge, however, answered MAUD's question in the negative, because the source of nonindependence lay in the environment (abnormal situations tend also to be stressful), rather than in the nature of PSFs elicited.

Hence, MAUD allows the two PSFs to stand, gives the judge the opportunity to review and edit the information already put in, and will allow direct input of information (without the method of triad difference) if the user indicates a preference for this type of elicitation (Frames 18 to 20).]

- Frame 19

  Would you like to be reminded of the information you have put in so far? N

[If the user had answered "yes" to this question, MAUD would have produced an annotated summary on the line printer.]

- Frame 20

  Do you want to alter any of the information you have put in so far? N

[If the user had answered "yes" to this question, MAUD would have offered the user comprehensive editing facilities for courses of actions (renaming, deletions, and insertions) and for PSFs (renaming, rescaling, deletions), until the user indicated that all the alterations required at that time had been completed. MAUD then continues as follows:]

- Frame 21

  Can you think of any other relevant performance shaping factor in which the courses of action differ from each other in a way that matters in determining likelihood of success? Y

  In not more than three words each time, please describe how some of them differ from the others.

  Some courses of action are OPERATOR COMPETENCY

  Whereas other courses of action are OPERATOR INCOMPETENCY

  Are you reasonably happy with this description? Y

- Frame 22

It should be possible to give each course of action a rating of 1 to 9 according to its position on the scale

```
OPERATOR COMPETENCY
 1              Your rating of     TASK 1: EDS MAN ACT       is 1
 2              Your rating of     TASK 2: RCIC MAN          is 5
 3              Your rating of     TASK 3: NIS INSERT        is 5
 4              Your rating of     TASK 10: BLK OUT MALOP    is 3
 5   to
 6                                Are these ratings OK? N
 7
 8
 9
OPERATOR INCOMPETENCY
```

[Whenever the user indicates he or she is not satisfied with some assessment, MAUD moves into an editing mode, presenting the appropriate menu of options for the types of editing that would be appropriate at that moment. At this point, the menu presented was as follows:]

- Frame 23

You can now

(1) Cancel this performance shaping factor (and all ratings on it)
(2) Change your ratings on this performance shaping factor

Which would you like to do?

Please type the number: 1

Are you sure? Y

OK

Press RETURN to continue

- Frame 24

Would you like to be reminded of the information you have put in so far? N

- Frame 25

Do you want to alter any of the information you have put in so far? N

- Frame 26

Can you think of any other relevant performance shaping factor in which the courses of action differ from each other in a way that matters in determining likelihood of success? Y

In not more than three words each time, please describe how some of them differ from the others

Some courses of action are STANDARD TRAINING

Whereas other courses of action are SPECIAL TRAINING

Are you reasonably happy with this description? Y

- Frame 27

It should be possible to give each course of action a rating from 1 to 9 according to its position on the scale

```
STANDARD TRAINING
 1                Your rating of     TASK 1: EDS MAN ACT        is 9
 2                Your rating of     TASK 2: RCIC MAN           is 5
 3                Your rating of     TASK 3: NIS INSERT         is 1
 4                Your rating of     TASK 10: BLK OUT MALOR     is 5
 5   to
 6                                Are these ratings OK? Y
 7
 8
 9
SPECIAL TRAINING
```

- Frame 28

Thinking only about the performance shaping factor below, what position on the scale would be IDEAL for a course of action in the present context?

```
STANDARD TRAINING
 1
 2                Your best possible value is : 1
 3
 4
 5   to           Is this rating OK? Y
 6
 7
 8
 9
SPECIAL TRAINING
```

- Frame 29

  Would you like to be reminded of the information
  you have put in so far? N

- Frame 30

  Do you want to alter any of the information you
  have put in so far? N

- Frame 31

  Can you think of any other relevant performance shaping factor in
  which the courses of action differ from each other in a way that
  matters in determining likelihood of success? Y

  In not more than three words each time, please describe
  how some of them differ from the others

  Some courses of action are HIGHLY SKILLED OPERATION

  Whereas other courses of action are LESS SKILLED OPERATION

  Are you reasonably happy with this description? Y

- Frame 32

  HIGHLY SKILLED OPERATION

1	Your rating of	TASK 1: EDS MAN ACT	is 1
2	Your rating of	TASK 2: RCIC MAN	is 4
3	Your rating of	TASK 3: NIS INSERT	is 6
4	Your rating of	TASK 10: BLK OUT MALOP	is 4
5 to			
6		Are these ratings OK? Y	
7			
8			
9			

  LESS SKILLED OPERATION

- Frame 33

Thinking only about the performance shaping factor below, what position on the scale would be IDEAL for a course of action in the present context?

```
HIGHLY SKILLED OPERATION
 1
 2                  Your best possible value is: 9
 3
 4
 5   to             Is this rating OK? Y
 6
 7
 8
 9
LESS SKILLED OPERATION
```

- Frame 34

Your assessments of likelihood of success for the courses of action under consideration in terms of their ratings on the scale ranging from STANDARD TRAINING to SPECIAL TRAINING seem very much the same as your assessments of likelihood of success for the courses of action in terms of their ratings on the scale ranging from HIGHLY SKILLED OPERATION to LESS SKILLED OPERATION.

Does this mean that these two scales have very similar meanings for you? Y

[As in Frame 18, MAUD has checked for preference independence between the ratings on the PSF scaled from HIGHLY SKILLED OPERATION to LESS SKILLED OPERATION and all other PSFs currently assessed, finding a possible violation of preference independence with the PSF scaled from STANDARD TRAINING to SPECIAL TRAINING. (Although the input ratings were dissimilar--in fact they were in opposite direction on the scale--the fact that ideal points were in different positions on each scale led to scale transformations which resulted in a very high positive correlation between the two sets of PSF ratings.)

However, unlike Frame 18, the judge answered positively to MAUD's question. In this case, MAUD proceeds as follows:]

- Frame 35

You are now going to be asked to name a new performance shaping factor which will substitute for those two similar scales.

Please type in a word (or phase of not more than three words) which has the same meaning as both SPECIAL TRAINING and HIGHLY SKILLED OPERATION

Your new word(s): NEED HIGHOP TRAINING

Now please type in a word (or phrase of not more than three words) which has the same meaning as both STANDARD TRAINING and LESS SKILLED OPERATION

Your new word(s): NEED LOWOP TRAINING

[At this point, MAUD restructures the set of PSFs by deleting the PSFs scaled between STANDARD TRAINING to SPECIAL TRAINING and between HIGHLY SKILLED OPERATION to LESS SKILLED OPERATION. These are replaced by the new PSF running from NEED HIGHOP TRAINING to NEED LOWOP TRAINING, which is now scaled in the usual way:]

- Frame 36

It should be possible to give each course of action a rating from 1 to 9 according to its position on the scale

```
NEED HIGHOP TRAINING
 1                Your rating of    TASK 1: EDS MAN ACT       is 1
 2                Your rating of    TASK 2: RCIC MAN          is 5
 3                Your rating of    TASK 3: NIS INSERT        is 7
 4                Your rating of    TASK 10: BLK OUT MALOP    is 3
 5   to
 6                                  Are these ratings OK? Y
 7
 8
 9
NEED LOWOP TRAINING
```

- Frame 37

Thinking only about the performance shaping factor below, what position on the scale would be IDEAL for a course of action in the present context?

```
NEED HIGHOP TRAINING
 1
 2                  Your best possible value is: 9
 3
 4
 5   to             Is the rating OK? Y
 6
 7
 8
 9
NEED LOWOP TRAINING
```

- Frame 38

Would you like to be reminded of the information you have put in so far? N

- Frame 39

Do you want to alter any of the information you have put in so far? N

- Frame 40

Can you think of any other relevant performance shaping factor in which the courses of action differ from each other in a way that matters in determining likelihood of success? Y

In not more than three words each time, please describe how some of them differ from the others

Some courses of action are GOOD PROCEDURES

Whereas other courses of action are POOR PROCEDURES

Are you reasonably happy with this description? Y

- Frame 41

It should be possible to give each course of action a rating from 1 to 9 according to its position on the scale

..

```
GOOD PROCEDURES
 1                Your rating of      TASK 1:  EDS MAN ACT     is 1
 2                Your rating of      TASK 2:  RCIC MAN        is 4
 3                Your rating of      TASK 3:  NIS INSERT      is 8
 4                Your rating of      TASK 4:  BLK OUT MALOP   is 6
 5  to
 6                                Are these ratings OK?  Y
 7
 8
 9
POOR PROCEDURES
```

- Frame 42

Thinking only about the performance shaping factor below, what position on the scale would be IDEAL for a course of action in the present context?

```
GOOD PROCEDURES
 1
 2                Your best possible value is:  1
 3
 4
 5   to           Is this rating OK?  Y
 6
 7
 8
 9
POOR PROCEDURES
```

- Frame 43

Would you like to be reminded of the information you have put in so far? N

- Frame 44

Do you want to alter any of the information you have put in so far? N

• Frame 45

Can you think of any other relevant performance shaping factor
in which the courses of action differ from each other in a way
that matters in determining likelihood of success? Y

In not more than three words each time, please describe
how some of them differ from the others

Some courses of action are LITTLE TIME AVAILABLE

Whereas other courses of action are MUCH TIME AVAILABLE

Are you reasonably happy with this description? Y

• Frame 46

It should be possible to give each course of action
a rating from 1 to 9 according to its position
on the scale

LITTLE TIME AVAILABLE

1	Your rating of	TASK 1: EDS MAN ACT	is 2
2	Your rating of	TASK 2: RCIC MAN	is 3
3	Your rating of	TASK 3: NIS INSERT	is 7
4	Your rating of	TASK 4: BLK OUT MALOP	is 1
5 to			
6	Are these ratings OK? Y		
7			
8			
9			

MUCH TIME AVAILABLE

• Frame 47

Thinking only about the performance shaping factor below, what position
on the scale would be IDEAL for a course of action
in the present context?

LITTLE TIME AVAILABLE

1	
2	Your best possible value is : 9
3	
4	
5 to	Is this rating OK? Y
6	
7	
8	
9	

MUCH TIME AVAILABLE

- Frame 48

Your assessments of likelihood of success for the courses of action under consideration in terms of their ratings on the scale ranging from HIGHLY ABNORMAL to FAIRLY NORMAL
seem very much the same as your assessments of likelihood of success for the courses of action in terms of their ratings on the scale ranging from LITTLE TIME AVAILABLE to MUCH TIME AVAILABLE

Does this mean that these two scales have very similar meanings for you? N

OK

Press RETURN to continue.

- Frame 49

Your assessments of likelihood of success for the courses of action under consideration in terms of their ratings on the scale ranging from HIGHLY STRESSFUL to LESS STRESSFUL
seem very much the same as your assessments of likelihood of success for the courses of action in terms of their ratings on the scale ranging from LITTLE TIME AVAILABLE to MUCH TIME AVAILABLE

Does this mean that these two scales have very similar meanings for you? N

OK

Press RETURN to continue.

[As in Frames 18 and 34, MAUD has checked for preference independence between the ratings of the PSF scaled from LITTLE TIME AVAILABLE to MUCH TIME AVAILABLE, and all other PSFs currently assessed, and this time has found possible violations of preference independence with two previously entered PSFs; those scaled from HIGHLY ABNORMAL to FAIRLY NORMAL and from HIGHLY STRESSFUL to LESS STRESSFUL.

However, the judge, as in Frame 18, answered MAUD's question in the negative for both cases, indicating that here also the source of nonindependence lay in the environment. Thus, MAUD allows all these PSFs to stand as they are and continues as follows:]

- Frame 50

Would you like to be reminded of the information you have put in so far? N

- Frame 51

  Do you want to alter any of the information you
  have put in so far? N

- Frame 52

  Can you think of any other relevant performance shaping factor
  in which courses of action
  differ from each other in a way that matters in
  determining likelihood of success? N

[At this point, the user indicates that MAUD should regain control of the session, and so MAUD responds by offering a series of options:]

- Frame 53

  You can

  (1) Obtain your overall assessments of likelihood of success
      on the basis of your current ratings
  (2) Get MAUD to help you again in thinking of other
      performance shaping factors on which the courses of action
      differ from each other
  (3) Have a summary printed of what has happened
      in the MAUD5 session so far
  (4) Save your data for future use
  (5) Stop

  Which would you like to do?
  Please type 1, 2, 3, 4, or 5 : 1

[Obtaining success likelihood indices requires that weights be assigned to PSFs. MAUD proceeds to assess these weights, in interaction with the user as follows:]

- Frame 54

  In order to find out which are the most important factors for you
  in determining likelihood of success, you will now have to choose
  between hypothetical courses of action which MAUD5
  will describe to you, using some of your own words.

  Would you like some instructions on how to make your choice? Y

- Frame 55

You are going to be asked to choose between 2 hypothetical courses of action which differ on just two of your scales.

Let's look at the first course of action

course of action A scores as follows:

HIGHLY STRESSFUL LESS STRESSFUL

X.....|.....|.....|.....|.....|.....|.....|.....|

HIGHLY ABNORMAL FAIRLY NORMAL

|.....|.....|.....|.....|.....|.....|.....|.....X

- On the first scale, the "X" indicates that course of action A scores the same as the worst course of action you rated on this scale.

- On the second scale, the "X" indicates that course of action A scores the same as your "ideal" course of action on this scale.

Press RETURN to continue.

- Frame 56

Now the second course of action

course of action B scores as follows:

HIGHLY STRESSFUL LESS STRESSFUL

|.....|.....|.....|.....|.....|.....|.....|.....X

HIGHLY ABNORMAL FAIRLY NORMAL

X.....|.....|.....|.....|.....|.....|.....|.....|

- On the first scale, the "X" indicates that course of action B scores the same as your "ideal" course of action on this scale.

- On the second scale, the "X" indicates that course of action B scores the same as the worst course of action you rated on this scale.

Press RETURN to continue.

- Frame 57

These two hypothetical courses of action will be presented to you together, and you will be asked which one you assess as most likely to succeed.

> If you choose course of action A, its rating on the second scale will be moved down half a unit to make a new, slightly less attractive hypothetical course of action A.
>
> If, on the other hand, you choose course of action B, then its rating on the first scale will be moved down a half a unit to make a new, slightly less attractive hypothetical course of action B.

You will then be asked again to choose between the two courses of action until MAUD finds the point at which your preference between courses of action A and B changes over.

Press RETURN to continue.

- Frame 58

This process will be repeated 2 times, using hypothetical courses of action described on various pairs of performance shaping factors, after which MAUD5 will know enough about which factors are important to you in choosing between courses of action to work out your assessments of likelihood of success for the real courses of action which you have been considering up till now.

Press RETURN to continue.

- Frame 59

Imagine you had to choose between course of action A which scores as follows:

HIGHLY STRESSFUL LESS STRESSFUL

X.....|.....|.....|.....|.....|.....|.....|.....|

HIGHLY ABNORMAL FAIRLY NORMAL

|.....|.....|.....|.....|.....|.....|.....|.....X

and course of action B which scores as follows:

HIGHLY STRESSFUL LESS STRESSFUL

|.....|.....|.....|.....|.....|.....|.....|.....X

HIGHLY ABNORMAL FAIRLY NORMAL

X.....|.....|.....|.....|.....|.....|.....|.....|

Which would you choose, A or B? A

[MAUD is seeking the point where the user is indifferent between A and B. Because the user chose A, A is made slightly less attractive by moving the position of the "X" for A on the scale HIGHLY ABNORMAL to FAIRLY NORMAL half a scale interval away from the ideal point. The user is now again asked whether he or she prefers A or B. This procedure is repeated until a point is found where the user changes from A to B, which in this case was as follows:]

- Frame 60

Imagine you had to choose between course of action A which scores as follows:

```
HIGHLY STRESSFUL                          LESS STRESSFUL

X.....|.....|.....|.....|.....|.....|.....|.....|

HIGHLY ABNORMAL                           FAIRLY NORMAL

|.....|.....|.....|.....|.....X.....|.....|.....|
```

and course of action B which scores as follows:

```
HIGHLY STRESSFUL                          LESS STRESSFUL

|.....|.....|.....|.....|.....|.....|.....|.....X

HIGHLY ABNORMAL                           FAIRLY NORMAL

X.....|.....|.....|.....|.....|.....|.....|.....|
```

OK. Now which would you choose, A or B? B

Are you reasonably sure? Y

[If the user is not sure about the changeover point, MAUD gives the option of moving the "X" again until an indifference point is found, repeats the sequence of weight assessments from the beginning, or abandons the attempt to estimate PSF weights, going instead to edit courses of action, PSFs, etc., before attempting to assess weights again. In this case, however, the user was reasonably certain of the changeover point between A and B, so MAUD continued with the weights assessment sequence.]

- Frame 61

Now imagine you had to choose between course of action A which scores as follows:

LITTLE TIME AVAILABLE MUCH TIME AVAILABLE

X.....|.....|.....|.....|.....|.....|.....|.....|

HIGHLY ABNORMAL FAIRLY NORMAL

|.....|.....|.....|.....|.....|.....|.....|.....X

and course of action B which scores as follows:

LITTLE TIME AVAILABLE MUCH TIME AVAILABLE

|.....|.....|.....|.....|.....|.....|.....|.....X

HIGHLY ABNORMAL FAIRLY NORMAL

X.....|.....|.....|.....|.....|.....|.....|.....|

which would you choose, A or B? A

[Several movements of the "X" on the scale occurred until the changeover between A and B was found to be at the point shown in Frame 62.]

- Frame 62

Now imagine you had to choose between course of action A which scores as follows:

```
LITTLE TIME AVAILABLE          MUCH TIME AVAILABLE

X.....|.....|.....|.....|.....|.....|.....|.....|

HIGHLY ABNORMAL                          FAIRLY NORMAL

|.....|.....|.....|.....|.....|..X..|.....|.....|
```

and course of action B which scores as follows:

```
LITTLE TIME AVAILABLE          MUCH TIME AVAILABLE

|.....|.....|.....|.....|.....|.....|.....|.....X

HIGHLY ABNORMAL                          FAIRLY NORMAL

X.....|.....|.....|.....|.....|.....|.....|.....|
```

OK. Now which would you choose, A or B? B
Are you reasonably sure? Y

[If the user was not sure about the changeover point, MAUD gives the option of moving the "X" again until an indifference point is found, repeats the sequence of weight assessments from the beginning, or abandons the attempt to estimate PSF weights, going instead to edit courses of action, PSFs, etc., before attempting to assess weights again. In this case, however, the user was reasonably certain of the changeover point between A and B, so MAUD continued with the weights assessment sequence.]

- Frame 63

Now imagine you had to choose between course of action A which scores as follows:

HIGHLY ABNORMAL FAIRLY NORMAL

X.....|.....|.....|.....|.....|.....|.....|.....|

NEED HIGHOP TRAINING NEED LOWOP TRAINING

|.....|.....|.....|.....|.....|.....|.....|.....X

and course of action B which scores as follows:

HIGHLY ABNORMAL FAIRLY NORMAL

|.....|.....|.....|.....|.....|.....|.....|.....X

NEED HIGHOP TRAINING NEED LOWOP TRAINING

X.....|.....|.....|.....|.....|.....|.....|.....|

Which would you choose, A or B? B

[This time the user initially preferred course of action B (rather then A as in Frames 59 and 61), and so MAUD makes course of action B progressively less attractive by moving the position of "X" down the scale towards HIGHLY ABNORMAL until the user changes over from preference for B to preference for A. This was found to be at the point shown in Frame 64 which, as can be seen, required only a slight movement towards ABNORMAL before the user switched preferences. This indicates that the relative importances of the two PSFs assessed here are fairly evenly matched.]

• Frame 64

Now imagine you had to choose between course of action A which scores as follows:

HIGHLY ABNORMAL FAIRLY NORMAL

X.....|.....|.....|.....|.....|.....|.....|.....|

NEED HIGHOP TRAINING NEED LOWOP TRAINING

|.....|.....|.....|.....|.....|.....|.....|.....X

and course of action B which scores as follows:

HIGHLY ABNORMAL FAIRLY NORMAL

|.....|.....|.....|.....|.....|.....|.....|..X..|

NEED HIGHOP TRAINING NEED LOWOP TRAINING

X.....|.....|.....|.....|.....|.....|.....|.....|

OK. Now which would you choose, A or B? A

Are you reasonably sure? Y

- Frame 65

Now imagine you had to choose between course of action A which scores as follows:

NEED HIGHOP TRAINING NEED LOWOP TRAINING

X.....|.....|.....|.....|.....|.....|.....|.....|

POOR PROCEDURES GOOD PROCEDURES

|.....|.....|.....|.....|.....|.....|.....|.....X

and course of action B which scores as follows:

NEED HIGHOP TRAINING NEED LOWOP TRAINING

|.....|.....|.....|.....|.....|.....|.....|.....X

POOR PROCEDURES GOOD PROCEDURES

X.....|.....|.....|.....|.....|.....|.....|.....|

Which would you choose, A or B? B

[This time course of action B was initially choosen, and several movements were made on the scale from NEED LOWOP TRAINING towards NEED HIGHOP TRAINING before the judge's changeover point for preference for A was found at the point shown in Frame 66.]

- Frame 66

Now imagine you had to choose between course of action A which scores as follows:

NEED HIGHOP TRAINING NEED LOWOP TRAINING

X.....|.....|.....|.....|.....|.....|.....|.....|

POOR PROCEDURES GOOD PROCEDURES

|.....|.....|.....|.....|.....|.....|.....|.....X

and course of action B which scores as follows:

NEED HIGHOP TAINING NEED LOWOP TRAINING

|.....|.....|.....|.....|.....|.....|.....|.....X

POOR PROCEDURES GOOD PROCEDURES

X.....|.....|.....|.....|.....|.....|.....|.....|

OK. Now which would you choose, A or B? A

Are you reasonably sure: Y

[At this point, MAUD has sufficient information to compute PSF weights and Success Likelihood Indices. MAUD does this and then provides an updated summary of the results on the line printer.]

- Frame 67

The summary of your session which is being prepared now will show your overall assessments of likelihood of success for the courses of action under consideration.

[In this case, the summary was that given in Figure 3.2. The success likelihood indices for all the courses of action are given, followed by an annotated record of the material that was elicited in the MAUD session as a basis for computing those SLIs.

When the summary has been presented, MAUD asks the user:]

- Frame 68

  Do you want to alter any of the information you
  have put in so far? N

[If the user had answered "Y," MAUD would have offered the opportunity to edit courses of action and performance shaping factors. Any editing destroys the validity of any PSF weights computed earlier, and so after editing, SLIs are not available for the course of action under consideration until MAUD has gone through the PSF weighting procedure in interaction with the user again. When the user wishes to alter nothing, MAUD offers a chance to add new material which may have come to mind during the weights assessment procedure before returning to the main menu (Frames 69 and 70).]

SUMMARY OF SESSION LEVEL A TASKS SUBSET I SO FAR:

Current order of assessments of likelihood of success of courses of action from best to worst
(success likelihood indices are given in brackets)

TASK 3: NIS INSERT (0.84) BEST

TASK 2: RCIC MAN (0.59)

TASK 1: EDS MAN ACT (0.33)

TASK10: BLK OUT MALOP (0.13) WORST

Ratings of courses of action on the scales you are currently using:

Rating scale number	TASK 1 EDS MAN ACT	TASK 2 RCIC MAN	TASK 3 NIS INSERT	TASK 10 BLK OUT MALOP	performance shaping factor
(1)	3	5	7	1	HIGHLY ABNORMAL(1) to FAIRLY NORMAL(9) Ideal value = 9
(2)	3	5	7	1	HIGHLY STRESSFUL(1) to LESS STRESSFUL(9) Ideal value = 9
(6)	1	5	7	3	HIGHOP TRAINING(1) to LOWOP TRAINING(9) Ideal value = 9
(7)	1	4	8	6	GOOD PROCEDURES(1) to POOR PROCEDURES(9) Ideal value = 1
(8)	2	3	7	1	LITTLE TIME AVAILABLE(1) to MUCH TIME AVAILABLE(9) Ideal value = 9

Figure 3.2 Summary of the SLIM-MAUD SLI assessment sessions produced by the MAUD on the line printer.

```
* The following information shows your assessments of likelihood of success
  for the courses of action under consideration
  1.00 represents the best course of action and 0.00 represents the
  worst course of action on each performance shaping factor.
* If you wish to change anything, you may do so at the end
  of this summary.

                 T    T    T    T
                 A    A    A    A
                 S    S    S    S
                 K    K    K    K
                                1
                 1    2    3    0
                 :    :    :    :
                                B
                 E    R    N    L
                 D    C    I    K
                 S    I    S
                      C         O
                 M              U
                 A    M    N    T
                 N    A    S
                      N    E    M
                 A         R    A
Rating           C         T    L
scale            T              O   performance shaping factor
number                          P

 (1)     0.33 0.67 1.00 0.00 HIGHLY ABNORMAL to FAIRLY NORMAL
                              relative importance = 0.24
 (2)     0.33 0.67 1.00 0.00 HIGHLY STRESSFUL to LESS STRESSFUL
                              relative importance = 0.17
 (6)     0.00 0.67 1.00 0.33 HIGHOP TRAINING to LOWOP TRAINING
                              relative importance = 0.24
 (7)     1.00 0.57 0.00 0.29 GOOD PROCEDURES to POOR PROCEDURES
                              relative importance = 0.16
 (8)     0.17 0.33 1.00 0.00 LITTLE TIME AVAILABLE to MUCH TIME AVAILABLE
                              relative imprtance = 0.19

The following scales are no longer in use for the reasons given below

(3)  OPERATOR COMPETENCY to OPERATOR INCOMPETENCY
     You canceled this scale after trying to rate the courses of action on
     it.
(4)  STANDARD TRAINING to SPECIAL TRAINING
     This scale was canceled because of its
     similarity with scale number 5
(5)  HIGHLY SKILLED OPERATION to LESS SKILLED OPERATION
     This scale was canceled because of its
     similarity with scale number 4
END OF SUMMARY.
```

Figure 3.2 (Cont'd) Summary of the SLIM-MAUD SLI assessment sessions produced by the MAUD on the line printer.

- Frame 69

  Can you think of any other relevant way that the courses of action differ from each other in a way that matters in determining likelihood of success? N

- Frame 70

  You can

  (1) Obtain your overall assessments of likelihood of success on the basis of your current ratings
  (2) Get MAUD to help you again in thinking of other performance shaping factors on which the courses of action differ from each other.
  (3) Have a summary printed of what has happened in the MAUD5 session so far
  (4) Save your data for future use
  (5) Stop

  Which would you like to do?
  Please type 1, 2, 3, 4, or 5 : 4

[When the user indicates that the data from the session are to be saved, MAUD first asks for the name of the file in which the data are to be saved and then saves the data in the file of that name on the SLIM-MAUD data disk currently in disk drive B of the computer. If the file name does not already exist in the directory of the SLIM-MAUD data disk, then a new file is automatically created. If however, the file name already exists (i.e., data has been placed in the file previously), then the old data are automatically replaced by the new data.]

- Frame 71

Please type the name of the file in which you want
to keep the material from this session: LASUB1
[a max. of 6 characters]

Is this name O.K.? Y

Session now stored in LASUB1

A>

[The prompt A> indicates that MAUD has completed its task and has returned the user to the microcomputer operating system, where it is now possible to turn off the computer without damaging any data recorded during the assessment session. Alternatively, responding

A>MAUD5

will cause MAUD to be reentered for use in a new session, or for further work on the data in file LASUB1, or in any other MAUD-produced file previously stored on the disk inserted in the microcomputer's disk drive B.]

3.5 Conversion of SLI Values into Probabilities

As described in Section 1.10.8 of this volume, it is possible to convert SLI values into asessed probabilities of failure given (1) probabilities of success estimates for reference courses of action among the set assessed within a SLIM-MAUD session, and (2) the assumption of a functional relationship between SLI values and probability values which is presently defined as:

$$\log P_i = a(SLI_i) + b$$

where:

a & b = empirical constants
SLI_i = SLI value computed by MAUD5 for the i^{th} course of action
P_i = assessed probability for the i^{th} course of action

MAUD5 itself does not include a routine for converting SLIs into probabilities of failure, but the computer code given in Figure 3.3 is an example of a simple program that can be appended to MAUD5 to achieve this task. To use this program, the user loads BASIC on computer, and then types in the program and saves it under the name SLIPROB.

```
5   '*****************************SLIPROB*************************************
6   '***************SAMPLE PROGRAM TO CONVERT SLIs TO PROBABILITIES***********
10  '
20  FOR I=1 TO 24: PRINT: NEXT
30  PRINT "Please give the name of the first reference course of action :"
40  LINE INPUT">",R1$:PRINT
50  PRINT "What is the probability of failure";:INPUT P1
60  PRINT "What is its SLI value";:INPUT SLI1:PRINT
70  PRINT "Please give the name of the second reference course of action :"
80  LINE INPUT ">",R2$: PRINT
90  PRINT "What is the probability of failure";:INPUT P2
100 PRINT "What is its SLI value";:INPUT SLI2:PRINT
110 LINE INPUT "are these values O.K.?", Q$
120 IF Q$="" THEN 110
130 Q$=LEFT$(Q$,1)
140 IF Q$="N" THEN 20
150 IF Q$="n" THEN 20
160 '
170 'COMPUTE PARAMETER VALUES
180 P1=LOG(P1): P2=LOG(P2)
190 A=(P1-P2)/(SLI1-SLI2)
200 B=P1-A*SLI1
210 '
220 PRINT:PRINT'-----------------------------------------------------------
------------":
230 PRINT:PRINT "Please give the name of a course of action for which you
require a SLI value :"
240 LINE INPUT ">", R$:PRINT
250 PRINT :what is its SLI value";:INPUT SLI
260 P=A*SLI+B: P=EXP(P)
270 PRINT "Assessed probability is ";P
280 LINE INPUT "MORE?";Q$
290 IF Q$="" THEN 280
300 Q$=LEFT$(Q$,1)
310 IF Q$="Y" THEN 230
320 IF Q$="y" THEN 230
330 END
```

Figure 3.3 Sample program to convert SLIs to assessed probabilities of failure (the code is written in microsoft BASIC).

The following is an example of the use of this program in assessing probabilities of failure for the course of action whose MAUD-computed SLI values are shown in Figure 3.2. In this example, NIS insert is the first reference course of action whose probability of failure is given as .001 and BLK OUT MALOP is the second reference course of action whose probabilit of failure is given as .10:

```
Please give the name of the first reference course of action:
 TASK 3:  NIS INSERT

What is the probability of failure?  .001
What is its SLI value?  .84

Please give the name of the second reference course of action:
 TASK 10:  BLK OUT MALOP

What is the probability of failure?  .10
What is its SLI value?  .13

Are these values OK?  Y

--------------------------------------------------------------------------------

Please give the name of a course of action for which you require a SLI value:
 TASK 1:  EDS MAN ACT

What is its SLI value?  .33
Assessed probability of failure is 2.732873E-02
MORE?  Y

Please give the name of a course of action for which you require a SLI value:
 TASK 2:  RCIC MAN

What is its SLI value?  .59
Assessed probability of failure is 5.060871E-03
MORE?  Y

Please give the name of a course of action for which you require a SLI value:
 TASK 3:  NIS INSERT

What is its SLI value?  .84
Assessed probability of failure is .001
MORE?  Y

Please give the name of a course of action for which you require a SLI value:
 TASK 10:  BLK OUT MALOP

What is its SLI value?  .13
Assessed probability of failure is 9.999996E-02
MORE?  N
OK
```

4. PHASE IV RESEARCH: TEST PLAN

At the time of the preparation of this report, the first three phases of the SLIM research program had been completed and Phase IV was in progress. This chapter describes the Test Plan for Phase IV research.

4.1 Overall Aims of the Test Plan

The overall aim of the Test Plan is to evaluate the MAUD implementation of SLIM--i.e., SLIM-MAUD--on the basis of its practicality, acceptability, and usefulness. The Test Plan will focus on one particular implementation of SLIM, through the use of the interactive computer system MAUD5 (Humphreys and Wisudha, 1984). Nevertheless, the findings and conclusions will contribute to a fuller understanding of general issues related to the various implementations of SLIM. In addition, a detailed user guide designed to maximize the effecient application of SLIM-MAUD will be included in the Test Plan report.

The utility of the MAUD-based implementation of SLIM will 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 specific issues, methods for implementing the Test Plan, expected data, and types of analyses to be performed were previously summarized in Table 8.1 of Volume I. For convenience, that table is reproduced here as Table 4.1.

4.2 Practicality

The practicality of the SLIM implementation will be evaluated in terms of the eight criteria listed in Table 4.1 and discussed separately below.

4.2.1 Cost

Costs for MAUD-based implementations of SLIM were summarized in Section 1.14.1. The accuracy of these costs will be confirmed or modified on the basis of the actual experience gained during the Test Plan. A careful record will be kept of costs incurred. Tables itemizing cost estimates will be provided for alternative implementations of SLIM, along with recommendations for choosing among alternative in particular contexts according to cost-benefit principles--e.g., using a format similar to that adopted by Kneppreth et al. (1979).

Table 4.1 SLIM-MAUD Test Plan: Issues and Procedures.

Issues	Methods/Data	Analysis
Practicality:		
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.	Multidimensional Scaling (MDS) of user responses.
Support Requirements	Enumeration 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.
Acceptability:		
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.
Usefulness:		
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.

4.2.2 Subject Matter Experts

Experience gained in Phases I and II of the SLIM research program underscored the need to systematically investigate the appropriate expertise of judges to be used in SLIM assessments. The Test Plan will undertake this investigation by comparing the inter- and intra-group agreement on the classification of the tasks to be assessed in the Test Plan.

4.2.3 Support Requirements

- Hardware. MAUD implementations of SLIM require the availability of microcomputer support. Minimum requirements of the microcomputer include a Z80 CPU or IBM/PC compatiblity, 64K bytes of memory, two

floppy disk drives, a CP/M or IBM/PC DOS operating system, VDU, and printer. For MAUD "group sessions," the microcomputer system must also have a video output to a TV monitor that can be viewed by all judges simultaneously. A wide variety of low cost microcomputer systems, such as the IBM/PC, Epson QX10, or North Star Advantage adequately meet these requirements.

The analysis of SLIM-MAUD sessions would also be greatly facilitated if a video recorder (without camera) were available to record speech and VDU contents during the interactive sessions with MAUD.

- Software. Procedures for obtaining MAUD5, the software needed to run a SLIM-MAUD session, are presented in Appendix D.

- Office Space. A room of sufficient size is needed to accommodate up to about 12 people comfortably, preferably one with a blackboard.

4.2.4 Transportability

If a predefined set of PSFs is available, SLIM can be implemented using "scoring sheets" completed by hand, ensuring its transportability to almost any setting. However, it is wise to caution, for reasons discussed in Section 1.11 of this volume, that certain theoretical suboptimalities accompany this form of implementation. Where microcomputer support is required, SLIM-MAUD can be implemented on virtually any microcomputer system meeting the requirements discussed in Section 4.2.3 above. Thus, SLIM-MAUD can be used in a wide variety of settings.

4.2.5 Expandability

The scope of the Test Plan will be limited to the selection of 27 tasks from the 35 tasks in the list of risk analysis tasks developed by the U.S. NRC and Sandia National Laboratories (SNL, 1983). The full list of tasks is reproduced in Appendix C. These tasks are wide ranging in scope and cover two levels of complexity. The empirical results generated from the analysis of material from the MAUD logs from Stage 1 of the Test Plan (see Section 4.5) will indicate the extent to which SLIM is appropriate across the whole range of tasks at each level in the test set. Given acceptable results for tasks located at any particular level, the MAUD implementation of SLIM may be considered expandable to the assessment of other tasks at that level, beyond those in the test set.

4.2.6 Time Requirements

Thus far, only a few pilot applications of SLIM through MAUD have been conducted. The average time taken per pilot session was approximately 45 minutes. Time requirements will vary with the number of tasks estimated per session, the number of sessions required to complete the assessment of all the tasks under consideration, and the experience of the judges. Implementing the Test Plan will provide additional data for a more accurate estimate of the time needed to run a SLIM-MAUD session.

4.2.7 Interface with Human Reliability Data Bank

The results of this Test Plan will provide a set of weighted PSFs (subsequently referred to as a "frame") for each subset of all the 27 tasks assessed, as the core for a future library of such frames. It is anticipated that a simple procedure will be specified for developing other frames in this library which will provide an important output across the interface with the Human Reliability Data Bank described in Section 1.14.2.

4.2.8 Implementability

It is assumed that facilitators of MAUD-based SLIM sessions should possess some knowledge of nuclear power plants and be familiar with the MAUD procedures. They need not, however, have a full understanding of the theoretical or technical underpinnings of MAUD (the MAUD program itself is a coherent product of these), nor do they need to have expertise in psychological scaling methods. These assumptions will be examined in the Test Plan by using facilitators with different types of expertise.

4.3 Acceptability

The ultimate test of the viability of SLIM as a method for estimating human error is its acceptability. The acceptability of the SLIM implementation will be evaluated in terms of the five criteria listed in Table 4.1 and discussed separately below.

4.3.1 Scientific Community

Multi-Attribute Utility Theory, the theoretical basis for SLIM-MAUD, has received considerable support in the scientific literature. The development of SLIM has been well documented and compared with other methodologies (Embrey, 1981, 1983a; Seaver and Stillwell, 1983). MAUD has been the subject of a number of validation studies in six countries (the UK, USA, the Netherlands, Sweden, Greece, and the Federal Republic of Germany), and has been widely distributed as an implementation tool for use in the scientific study of decision making, and as a decision aiding system with a wide number of applications. Nevertheless, SLIM is still in need of a full empirical validation study to ensure its acceptability. The aim of the Test Plan is to provide this empirical validation. The results will form the basis for a NRC report and will subsequently be submitted for publication in appropriate journals to ensure the proper dissemination of evidence supporting the validity of the methodology.

4.3.2 Expert Participants

The acceptability of SLIM to the judges (expert participants) taking part in SLIM-MAUD assessments will be examined by analyzing the debriefing interviews with all judges, as well as all the logs from MAUD sessions completed within the Test Plan. In addition, a formal survey questionnaire will be administered to judges. The results from the survey will be analyzed with respect to the method's acceptability.

4.3.3 Potential Users

The fundamental objective of SLIM is to produce HEPs that can be used in HRA segments of PRA. Thus, a crucial test of SLIM's viability is the extent to which experts in the PRA community (e.g., HRA and human factors experts) adopt it as part of PRA assessments. A preliminary and informal assessment of potential users of SLIM will be undertaken to estimate future acceptability. During the Test Plan, PRA specialists will be asked to give their opinions on the SLIM-MAUD Test Plan results.

4.3.4 U.S. Nuclear Regulatory Commission (NRC)

SLIM is a methodology in the public domain, documented in NUREG publications (Seaver and Stillwell, 1983; Embrey, 1983) and in the present report. Past and ongoing support by the NRC in developing SLIM is prima facie evidence of that agency's support. Continued NRC support can be assessed by the degree to which SLIM-MAUD is adopted by the NRC as a recommended methodology in future work.

4.3.5 Nuclear Utilities

Whether SLIM will be acceptable to nuclear utilities remains an unanswered question. Acceptability can only be determined by the extent to which utilities actually adopt SLIM in PRA assessments.

4.4 Usefulness

The basis for assessing the usefulness of any subjective estimation technique is the reliability and validity of the results provided.

4.4.1 Reliability

The reliability of SLIM-MAUD will be assessed by examining the consistency in the SLI estimates produced across judges and groups.

4.4.2 Face Validity

Face validity, also known as content validity, will be assessed by analyzing the results of the survey where judges and PRA experts have been asked to comment on the reasonableness of SLIM.

4.4.3 Convergent Validity

Convergent validity will be assessed by comparing the SLIM-MAUD HEP estimates with those produced by other techniques being tested (i.e., paired-comparisons and direct estimation [SNL, 1983]).

4.5 Procedures to Be Followed in Implementing the Test Plan

The Test Plan is divided into the following stages:

Stage 0 - selection of tasks for assessment by SLIM (already accomplished, see Appendix C).

Stage 1 - classification of tasks into subsets for simultaneous assessment within SLIM-MAUD in Stage 2.

Stage 2 - selection of the members of subject matter expert groups for Stage 3.

Stage 3 - implementation of SLIM-MAUD by each subject matter expert group for each subset of tasks.

Stage 4 - analysis and interpretation of results from SLIM-MAUD sessions.

Stage 5 - preparation and review of report of SLIM-MAUD validation study.

4.5.1 Stage 1: Classification of Tasks into Subsets for Simultaneous Assessments Within SLIM-MAUD in Stage 3

The 15 Level A and 12 Level B tasks identified in Stage 0 must meet the homogeneity requirements of SLIM-MAUD described in Section 1.13 of this volume. The first requirement, therefore, is to classify these tasks into subsets that meet these homogeneity requirements. The classification of tasks into subsets will be performed separately for Level A and B tasks.

Two groups of judges will be presented with descriptions of both levels of tasks that appear in Appendix C of this volume. They will be asked to make wholistic ratings of the interrelatedness of tasks using a paired comparison procedure. Ratings of interrelatedness will be based upon judgments of the relative importance of PSFs in determining the likelihood of success for each pair of tasks. Group consensus procedures will be used to obtain the wholistic ratings (see Nemiroff and King, 1975; Gustafson et al., 1983).

Written consensus ratings will be collected from each group for formal analysis. The analysis will be conducted in two steps as follows:

- Step 1 - Multidimensional scaling analysis of interrelatedness matrices. The data resulting from the above procedures will be "consensus" interrelatedness matrices for both Level A and B tasks. These four matrices (two groups x two levels) will be analyzed using multidimensional scaling procedures (implemented within KYST, Kruskal and Wish, 1978) to produce "interrelatedness maps" of clusters of tasks. Techniques are available for testing whether matrices produced by different groups, or at different levels, may be scaled within the

same hypothetical space (e.g., the SPLIT BY ROWS with BLOCKDIAGONAL = NO option of KYST). If this test fails across groups, then further groups of assessors will be given this task within a design aimed at locating the source of the instability. If the test fails across levels, then this will indicate that the levels distinction will form a major factor in the classification.

- Step 2 - Identification of clusters. This will be done by visually partioning the interrelatedness space(s) identified in step 1 in a way that yields clearly defined subsets of tasks. Details of this procedure may be found in Humphreys (1983). Each cluster and the rationale for its identification will then describe a particular cell in the task classification.

4.5.2 Stage 2: Selection of the Members of Subject Matter Expert Groups for Stage 3

Five six-member groups of judges will be used to implement the SLIM-MAUD procedures. The composition of each group will be as follows:

Group A - Three human factors experts and three judges with plant operating experience.

Group B - Two human factors experts, two judges with plant operating experience, and two judges with plant design experience.

Group C - Two human factors experts, two judges with plant operating experience, and two PRA experts.

Group D - A group of six judges comprising the four types of expertise described above.

These four groups will perform SLIM-MAUD assessments on the subsets of tasks identified in Stage 1.

Group E - Group composition the same as Group D above. However, these judges will classify the tasks into subsets themselves, before making their SLIM-MAUD assessments.

4.5.3 Stage 3: Implementation of SLIM-MAUD by each Subject Matter Expert Group for Each Subset of Tasks

Each group of subject matter experts will use SLIM-MAUD to assess SLIs for each subset of Level A and B tasks. A facilitator will be present at each session to monitor its progress and identify any critical incidents. A video recorder whose audio recording channel is connected to a microphone recording the facilitator's and the experts' comments should be used to keep a record of each session. The video recording channel will be connected to the microcomputer's video output socket, thus recording the contents of the VDU screen

in synchronization with the audio comments about these contents. After the sessions are completed, the facilitator will conduct a debriefing interview and formal survey to gain information from the judges on the issues of acceptability and usefulness discussed in Sections 4.3.4 and 4.4 of this volume.

Procedures to Be Followed in the SLIM-MAUD Sessions

- Assemble the members of each subset matter expert group (A, B, C, D, and E).
- The facilitator explains SLIM to each group (A, B, C, D, and E).
- The subsets of tasks, classified in Stage 1, are presented to Groups A, B, C, and D.
- Group E is given the list of Level A and B tasks used in Stage 1. The facilitator explains the need to classify them into homogeneous subsets consisting of 4 to 10 tasks. The group will be asked to make wholistic ratings of interrelatedness of tasks using a paired comparison procedure. Ratings of interrelatedness will be based upon judgments of the relative importance of PSFs in determining the likelihood of success for each pair of tasks. Group consensus procedures identical to those outlined for Stage 1 of Section 4.5.1 of this volume will be followed.
- All groups will implement SLIM-MAUD through consensus interaction procedures similar to those described in Section 3.4.5 of this volume. At the end of the implementation, the resultant SLI values will be converted into HEPs using the computer program described in Section 3.5 of this volume.

4.5.4 Stage 4: Analysis and Interpretation of Results from SLIM-MAUD Sessions

- MAUD logs will be analyzed to compare (1) differences in elicited PSFs between groups of judges, (2) differences in SLIs produced by Groups A, B, C, and D vs Group E, and (3) interactions between the above comparisons.
- A number of additional analyses of MAUD logs which may be carried out are described in Humphreys and McFadden (1980).
- Content analyses of the critical incident records and debriefing interviews will be conducted to pinpoint limitations and successes of the system with regard to practicality, usefulness, and acceptability (see Humphreys and Wooler, 1981). Statistical analyses will be conducted on the data collected in the formal survey of judges.

- Convergent validity will be assessed by examining the HEPs produced by SLIM-MAUD with those produced by the paired-comparison and direct estimation techniques employed in the Sandia National Laboratories study (SNL, 1983).

4.5.5 Stage 5: Preparation and Review of Report of SLIM-MAUD Study

The Test Plan report will summarize the findings of Stage 4. It will also discuss the advantages and limitations of alternative implementations of SLIM. Particular attention will be directed to issues of practicality, acceptability, and usefulness. A detailed user guide designed to maximize efficient application of SLIM will be included in the Test Plan report.

4.6 Test Plan Schedule

Figure 4.1 gives the schedule for the stages of the Test Plan identified in Section 4.5.

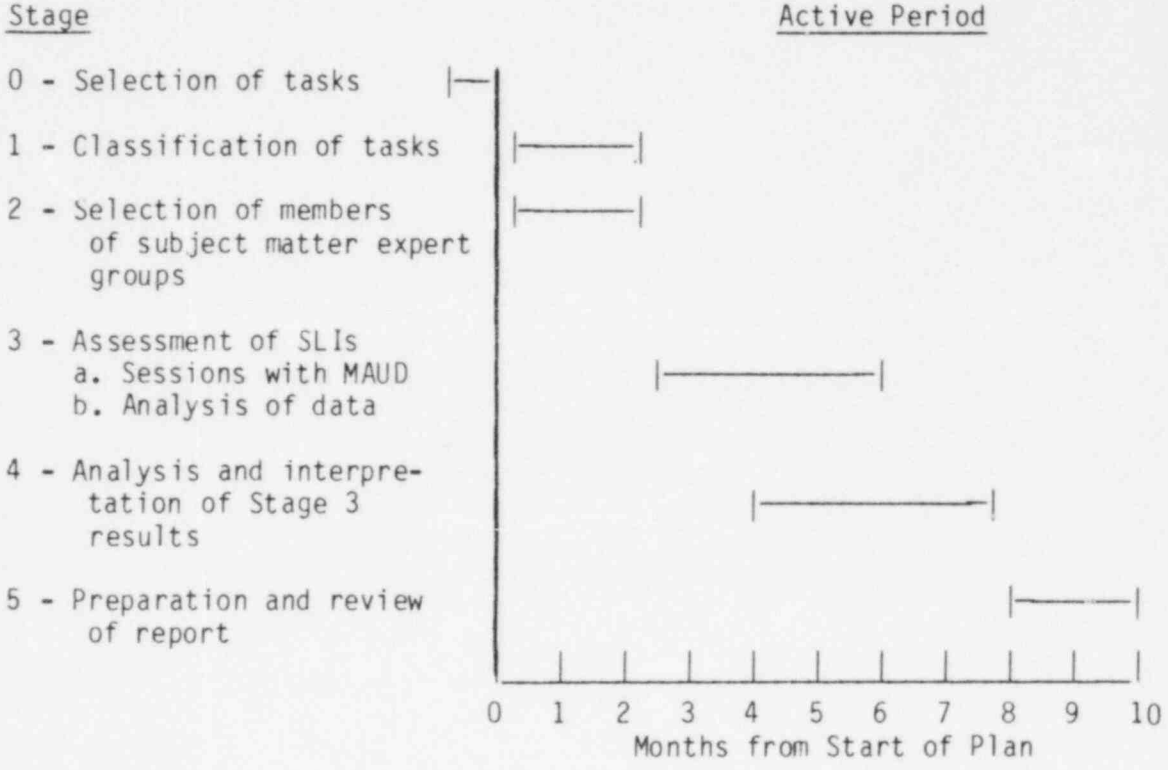


Figure 4.1 Schedule for the six stages of the Test Plan.

REFERENCES

Adams, S., Sabri, Z., Husseiny, A. (1980), "Maintenance and Testing Errors in Nuclear Power Plants: a Preliminary Assessment," in Proceedings of the 24th Human Factors Society Annual Conference, Los Angeles, California.

Altman, J. W. (1964a), "A Central Store of Human Performance Data," in Proceedings of the Symposium on Quantification of Human Performance Albuquerque, New Mexico, August 17-19.

Altman, J. W. (1964b), "Improvements Needed in a Central Store of Human Performance Data," Human Factors 6, 681-686.

Altman, J. W. (1967), "Classification of Human Error," In W. B. Askren (Ed.), Symposium on the Reliability of Human Performance in Work, AMRL Tech Report 67-88 May.

Alluisi, E. A. (1967), "Methodology in the Use of Synthetic Tasks to Assess Complex Performance," Human Factors 9, 375-384.

Arrow, K. J. (1952), Social Choice and Individual Values. New Haven: Yale University Press.

Beals, R., Krantz, D. H. and Tversky, A. (1968), "Foundations of Multidimensional Scaling," Psychological Review, 75, 127-142.

Berliner, C., Angel, D., and Shearer, J. W. (1964), "Behaviors, Measures and Instruments for Performance Evaluation in Simulated Environments," Symposium and Workshop on the Quantification of Human Performance, Albuquerque, New Mexico.

Bernouilli, D. (1954), "Exposition of a New Theory on the Measurement of Risk," Translated by L. Sommer, Econometrika, 22, 25-26.

Comer, M. K., Kozinsky, E. J., Eckel, J. S., and Miller, D. P. (1983), "Human Reliability Data Bank for Nuclear Power Plant Operations, Volume 2: Data Bank Concept and System Description." NUREG/CR-2744, Sandia National Laboratories.

Coombs, C. H. (1964), A Theory of Data. New York: Wiley.

Cronbach, L. J. and Meehl, P. E. (1955), "Construct Validity in Psychological Tests," Psychological Bulletin 52, 281-302.

Cronbach, L. J. (1964), Essentials of Psychological Testing. New York: Harper and Row.

Dawes, R. M. (1972), Fundamentals of Attitude Measurement. New York: Wiley

Dawes, R. M. and Corrigan, B. (1974), "Linear Models in Decision Making," Psychological Bulletin, 8, 95-106.

Edwards, A. L. (1957), Techniques of Attitude Scale Construction. New York: Appleton Century Crofts.

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

Edwards, W. (1981), "Human Factors and Safety," paper presented at Safety and Management Symposium, Trondheim, Norway.

Edwards, W., and Newman, R. (1982), "Multi-Attribute Evaluation," Sage University paper series on Quantitative Applications in the Social Sciences, 07-026, Beverly Hills and London: Sage Publications.

Einhorn, H. J. and Hogarth, R. M. (1975), "Unit Weighting Schemes for Decision Making," Organizational Behavior and Human Performance, 13, 171-192.

Embrey, D. E. (1979), "Human Reliability in the Process Industries," Tero technica 1, 109-117.

Embrey, D. E. (1981), "A New Approach to the Evaluation and Quantification of Human Reliability in Systems Assessment," in The Proceedings of the 3rd National Reliability Conference, Birmingham, UK.

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, Upton, New York.

Embrey, D. E. (1983b), "Modelling and Quantifying Human Reliability in Abnormal Conditions." in The Proceedings of the 4th National Reliability Conference, Birmingham, UK.

Epting, F. R., Suchman, D. I., and Nickeson, C. J. (1971), "An Evaluation of Elicitation Procedures for Personal Constructs," British Journal of Psychology, 62, 513-517.

Farina, A. J., and Wheaton, G. R. (1971), "Development of a Taxonomy of Human Performance: The Task Characteristics Approach to Performance Prediction," Report 7, American Institute of Research, Washington DC.

Fischer, G. W. (1972), "Multidimensional Value Assesment for Decision Making," Technical Report 037230-2-T. Ann Arbor: Engineering Psychology Laboratory, University of Michigan.

Fishburn, P. C. (1970), Utility Theory for Decision Making. New York: Wiley.

Fleishman, E. A. (1967), "Performance Assessment Based on an Empirically Based Task Taxonomy," Human Factors, 9, 349-366.

Fleishman, E. A. (1975), "Towards a Taxonomy of Human Performance," American Psychologist, 20, 1127-1149.

Fransella, F., and Bannister, D. (1977), A Manual for Repertory Grid Technique. London: Academic Press.

Galanter, E. (1962). "The Direct Measurement of Utility and Subjective Probability," American Journal of Psychology, 75, 208-220.

Gustafson, D., Peterson, P., Koetsky, E., von Koningsveld, R., Macco, A., Casper, S., and Rossmeissl, J. (1983), "A Decision Analystic System for Regulating the Quality of Care in Nursing Homes: System Design and Evaluation," in Humphreys, P. C., Svenson, O., and Vari, A., (Eds.), Analysing and Aiding Decision Processes, Amsterdam: North Holland.

Hollnagel, E. (1981), "On the Validity of Simulator Studies: Problems and Preliminary Precepts," Report no. N-39-80, Electronics Department, RISO National Laboratory, Denmark.

Hollnagel, G., Pedersen, O. M., and Rasmussen, J. (1981), "Notes on Human Performance Analysis," RISO-M-2285, RISO National Laboratory, Denmark.

Huber, G. P. (1974), "Methods for Quantifying Subjective Probabilities and Multiattribute Utilities," Decision Sciences, 5, 430-458.

Humphreys, P. C. (1977), "Applications of Multiattribute Utility Theory," in Jungermann, H., and de Zeeuw, G., Decision Making and Change in Human Affairs, Dordrecht: Reidel.

Humphreys, P. C., and McFadden, W. (1980), "Experiences with MAUD: Aiding Decision Structuring Versus Bootstrapping the Decision Makers," Acta Psychologica, 45, 51-69.

Humphreys, P. C. (1983), "Use of Problem Structuring Techniques for Option Generation: A Computer Choice Case Study," in P. C. Humphreys, Svenson, O., and Vari, A., (eds.), Analysing and Aiding Decision Processes. Amsterdam: North Holland.

Humphreys, P. C., and Wisudha, A. (1983), MAUD4, Technical Report 83-5, Decision Analysis Unit, London School of Economics and Political Science.

Humphreys, P. C., Wisudha, A. (1984), MAUD5, Technical Report 84-3, Decision Analysis Unit, London School of Economics and Political Science.

Humphreys, P. C., and Wooler, S. (1981), "Development of MAUD4 Through a Feasibility Study on the use of Decision Aiding Software in Vocational Guidance," Technical Report 81-4, Decision Analysis Unit, London School of Economics and Political Science.

Hunns, D. M. (1982), "Discussions Around a Human Factors Data-Base. An Interim Solution: The Method of Paired Comparisons," in Green, A. E., (Ed.), High Risk Safety Technology, London: John Wiley and Sons Ltd.

John, R. S., von Winterfeldt, D., and Edwards, W. (1983), "The Quality and User Acceptance of Multiattribute Utility Analysis Performed by Computer and Analyst," in Humphreys, P. C., Svenson, O., and Vari, A., (Eds.), Analyzing and Aiding Decision Processes, Amsterdam: North Holland.

Johnson, E. M., and Huber, G. P. (1977), "The Technology of Utility Measurement," in IEEE Transactions on Systems, Man and Cybernetics, SMC-7,5.

Keeney, R. C. and Raiffa, H. (1976), Decision With Multiple Objectives, Preferences and Value Tradeoffs. New York: Wiley.

Kelly, G. A. (1955), The Psychology of Personal Constructs. New York: Norton.

Klecka, W. R. (1980), "Discriminant Analysis," Sage University paper series on Quantitative Applications in the Social Sciences, 07-19, Beverly Hills and London: Sage Publications.

Kneppreth, N. P., Hoessel, W., Gustafson, D. J., and Johnson, E. M. (1978), "A Strategy for Selecting a Worth Assessment Technique," Technical Paper 280 (ADA055345), Arlington, Virginia: U. S. Army Research Institute for the Behavioral and Social Sciences.

Kobashi, Y. (1983), "Tables-Oriented Reconstruction of MAUD," paper presented at the 9th Research Conference on Subjective Probability, Utility and Decision Making, Groningen, Netherlands.

Krantz, D. H., Luce, R. D., Suppes, P., and Tversky, A. (1971), Foundations of Measurement, New York: Academic Press.

Kruskal, J. B. and Wish, M. (1978), "Multidimensional Scaling," Sage University paper series on Quantitative Applications in the Social Sciences, 07-011, Beverly Hills and London: Sage Publications.

Lee, W. (1971), Decision Theory and Human Behavior. New York: Wiley.

Levine, J. and Teichner, W. (1971), "Development of a Taxonomy of Human Performance: an Information Theoretic Approach," Report 9, American Institute of Research, Washington, DC.

Levine, J., Romashko, T. and Fleishman, E. A. (1971), "Evaluation of Abilities Classification System for Integrating and Generalizing Research Findings," Report 12, American Institute of Research, Washington, DC.

Lichtenstein, S., Fischhoff, B. and Phillips, L. D. (1981), "Calibration of Probabilities: The State of the Art to 1980," in Kahnemann, D., Slovic, P., Tversky, A., (Eds.), Judgment Under Uncertainty: Heuristics and Biases. London: Cambridge University Press.

Meister, D. (1964), "Methods of Predicting Human Reliability in Man-Machine Systems," Human Factors, 6, 621-646.

Meister, D. and Rabideau, G. F., (1965), Human Factors Evaluation in System Development. New York: Wiley.

Metwally, A., Sabri, Z., Adams, S., and Husseiny, A. (1982), "A Data Bank for Human Related Events in Nuclear Power Plants," in Proceedings of the 26th Human Factors Society Annual Meeting, Seattle, Washington.

Miller, R. B. (1967), "Task Taxonomy: Science or Technology?" in Singleton, W. T., Easterby, R. and Whitfield, D. J., (Eds.), The Human Operator in Complex Systems, London: Taylor and Francis.

Miller, R. B. (1971), "Development of a Taxonomy of Human Performance: Design of a Systems Task Vocabulary," Technical Report no. 11, American Institute for Research, Silver Spring, Md.

Nemiroff, P. M., and King, D. C. (1975), "Group Decision Making Performance as Influenced by Consensus and Self Orientation," Human Relations, 28, 1.

Norman, D. A. (1981), "Categorization of Action Slips," Psychological Review, 88, 1-15.

O'Brien, J. N., Rosa, E. A., and Stengrevics, J. M. (1983), "A Feasibility Assessment of Utilizing Organizational and Sociological Research in Understanding, Assessing, and Improving Control Room Operations," BNL-NUREG-33067, Brookhaven National Laboratory, Upton, NY.

Phillips, L. D., Humphreys, P. C., and Embrey, D. E. (1983), "A Socio-technical Approach to Assessing Human Reliability," Technical Report 83-4, Decision Analysis Unit, London School of Economics and Political Science.

Pontecorvo, A.B. (1966), "A Method of Predicting Human Reliability," Annals of Reliability and Maintenance, 4, 337-342.

President's Commission on the Accident at Three Mile Island (1979), The Need for Change: The Legacy of TMI. Washington, D. C.:U.S. Government Printing Office.

Raiffa, H. (1969), "Preferences for Multiattribute Alternatives," Memorandum RM-5868-DOT/RC, Santa Monica, CA: The Rand Corporation.

Rasmussen, J. (1980), "What Can Be Learned From Human Error Reports?" in Duncan, K. D., Gruneberg, M., and Wallis, D., (Eds.), Changes in Working Life. New York: Wiley.

Rasmussen, J., Pedersen, O. M., Carnino, A., Griffon, M., Mancini, G., and Gagnolet, P. (1981), "Classification System for Reporting Events Involving Human Malfunctions," RISO-M-2240, RISO National Laboratory, Denmark.

Reason, J. T. (1979), "Skill and Error in Everyday Life," in Adult Learning, Howe M. (Ed.), London: Wiley.

Rook, L. W. (1962), "Reduction of Human Error in Industrial Production," Report SCTM-65-135 Sandia Corporation, Albuquerque, New Mexico.

Savage, L. J. (1954), The Foundations of Statistics. New York, Wiley.

Sayeki, Y. (1972), "Allocation of Importance: an Axiom System," Journal of Mathematical Psychology, 9, 55-65.

Seaver, D. A., and Stillwell, W. G. (1983), "Procedures for Using Expert Judgment to Extract Human Error Probabilities in Nuclear Power Plant Operations," NUREG/CR-2473, Sandia National Laboratories, Albuquerque, New Mexico.

Selvidge, J. (1975), "A Three Step Procedure for Assigning Probabilities to Rare Events," in Wendt, D., and Vleck, C., (Eds.), Utility, Probability and Human Decision Making, Dordrecht, Holland: D. Reidel.

Smith, R. L., Westland, R. A., and Blanchard, R. E. (1969), "Technique for Establishing Personnel Performance Standards (TEPPS) Results of Navy User Test," Report PTB-70-5, Vol. III, Personnel Research Division, Bureau of Navy Personnel, Washington, DC.

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

Stael von Holstein, C-A.S. and Matheson, J. (1979), "A Manual for Encoding Probability Distributions," SRI Project 7028, SRI International.

Swain, A. D., and Guttmann, H. E. (1983) "Handbook of Human Reliability Analysis with Emphasis on Nuclear Power Plant Applications," NUREG/CR-1278, Sandia National Laboratories, Albuquerque, New Mexico.

Theologus, G. C., and Fleishman, G. A. (1971), "Development of a Taxonomy of Human Performance: Validation Study of Ability Scales for Classifying Human Tasks," Report 10, American Institute of Research, Washington, DC.

Thorndike, R. L. (1949), Personnel Selection. New York: Wiley.

von Winterfeldt, D. (1980), "Structuring Decision Problems for Decision Analysis," Acta Psychologica, 45, 71-94.

von Winterfeldt, D., and Edwards, W. (1973), "Evaluation of Complex Stimuli Using Multiattribute Utility Procedures," Technical Report 011313-2-T, Ann Arbor: Engineering Psychology Laboratory, University of Michigan.

von Winterfeldt, D., and Fischer, G. W. (1975), "Multiattribute Utility Theory: Models and Assessment Procedures," in Wendt, D., and Vlek, C., (Eds.), Utility, Probability and Human Decision Making. Dordrecht, Holland: Reidel.

von Winterfeldt, D., Barron, F. H., and Fischer, G. W. (1980), "Theoretical and Empirical Relationships Between Risky and Riskless Utility Functions," Research Report 80-3, Los Angeles: Social Science Research Institute, University of Southern California.

Wish, M., and Carroll, J. D. (1974), "Applications of Individual Differences Scaling to Studies of Human Perception and Judgment," in La Carterette, E. C. and Friedman, M. P. (Eds.), Handbook of Perception, Volume II, New York: Academic Press.

Wooler, S., and Erlich, A. (1982), "Interdependence Between Problem Structuring and Attribute Weighting in Transitional Decision Problems," in Humphreys, P. C., Svenson, O., and Vari, A., (Eds), Analyzing and Aiding Decision Processes, Amsterdam: North Holland.

APPENDIX A

INFORMATION USED IN EXPERIMENTAL EVALUATION OF SLIM

1. Introduction

In this chapter, the task descriptions and other information used by the subjects in Phase I of the research program, the SLIM experiment, are presented. A number of illustrations which accompanied the task descriptions have not been included. This is partly for reasons of space, and partly because the illustrations are available in the source references cited in Section 4 of Volume I. The illustrations which accompanied the tasks taken from consultancy studies are confidential and are not available for general release.

As will be discussed in the next section, the amount of information available from published accounts of experiments is far from adequate from the point of view of their use in human reliablility evaluation. However, to satisfy other criteria which were applied when selecting tasks the choice of tasks which could be included in the study was very small.

In addition to the task descriptions, subjects were also given explanatory information on the task categorization scheme (Section 5.1) and PSFs (Section 5.3.2), the weighting and the rating procedures (Sections 5.3.3 and 5.3.4). Section 5, therefore, represents the information given to the subjects.

2. Content Analysis of Task Descriptions

During the experiment, it became apparent that some of the task descriptions were inadequate in the information they contained, so that judges found it difficult to perform the rating part of SLIM. After the experiment, a content analysis was performed to determine which task descriptions were likely to have introduced a large random error component into expert assessments, causing inaccurate predictions.

The content analysis was carried out by considering the level of information content in each task description for each PSF. The level of information was assessed on a 4-point scale, 0-3, "0" representing no information on that PSF and "3" representing high information content.

Table A.1 shows the results of this analysis, and a simple additive procedure (adding the ratings for 6 PSFs in each task) gives a measure of the overall level of information for each task. At the right hand side of the table, the tasks are ranked from 1 to 21 denoting the best described task to the worst, respectively.

From Table A.1, several tasks appear to have inadequate task descriptions, particularly tasks 4, 5, 8, and 14. It is desirable to minimize the elimination of tasks from the task set, as this in turn reduces the number of degrees of freedom usable in statistical analysis of the data. Furthermore, task 14 was of particular importance in representing the type of task that might require quantification in a PRA, i.e., it was concerned with an omission

Table A.1 Content analysis of task description.

	Rating of Information of PSFs in Task Descriptions							
		A	B	C	D	E	F	∑R	Rank
	01	3	0	3	3	0	0	09	15
	02	0	1	3	2	2	3	11	10
	03	2	3	2	2	1	2	12	07
	04	0	2	3	3	0	0	08	18
	05	0	3	0	2	2	1	08	18
	06	0	0	3	2	2	3	10	13
	07	2	3	0	3	0	1	09	15
	08	0	3	0	2	2	0	07	20
	09	3	3	3	3	2	1	15	02
Task	10	1	2	0	3	3	2	11	10
Number	11	1	2	1	3	3	2	12	07
	12	1	2	0	3	3	2	11	10
	13	2	2	0	3	3	2	12	07
	14	3	2	0	0	0	1	05	21
	15	1	3	2	3	0	1	09	15
	16	2	3	2	3	1	3	14	05
	17	2	3	3	3	3	3	17	01
	18	3	3	3	3	2	1	15	02
	19	3	3	1	3	3	0	13	06
	20	2	3	3	3	1	3	15	02
	21	0	2	3	3	0	2	10	13

error in a BWR simulator. Despite the poor description of this task, it could be argued that the experts in the group, especially the operators, would have experience in this type of situation. Information on tasks 4 and 5 was equally poor, and task 5 created particular problems during the experiment. Several of the operators were reluctant to accept that an inoperative switch would not be removed, even in a simulator. Task 8 was confusing in that judges were making single judgments on what effectively constituted two error rates (calculating errors), those of experienced and less experienced subjects. In other words, the judges were being asked to give judgments on an "average" score. Also, the study from which this task had been taken suggested that subjects were timed, but it was uncertain whether they had to complete the calculations within a time limit or whether they were timed for another reason. Task 12 was adequately described in terms of its score on the content analysis measure, but caused assessment problems during the experiment, because the judges were unable to take into account certain important charactertistics of the control panel which were not given in the task description.

In summary, because of inadequate or misleading task descriptions and information, tasks 5, 8 and 12 were eliminated. This seemed to be the best compromise that could be achieved between eliminating tasks which had created difficulties for the judges and retaining those which were technically significant. Another consideration was the desire to retain as many tasks as possible in order to perform meaningful statistical analyses.

3. Probabilities of Failure for the Tasks Assessed

Failure probabilities were calculated by dividing the number of errors by the total number of trials reported in the literature from which the tasks were drawn.

1.	0.032	12.	0.0023
2.	0.195	13.	0.0042
3.	0.0095	14.	0.029
4.	0.064	15.	0.102
5.	0.001	16.	0.064
6.	0.00005	17.	0.073
7.	0.0293	18.	0.34
8.	0.27	19.	0.30
9.	0.163	20.	0.65
10.	0.0065	21.	0.20
11.	0.003		

4. Sources of Task Descriptions

The task descriptions were obtained from the following sources:

Task	Source
1.	Hull, A. J. (1976), "Human Performance with Homogeneous Patterned and Random Alphanumeric Displays," Ergonomics, 19 (6), pp. 741-750.

Task	Source
2.	Jacobson, H. J. (1953), "A Study of Inspector Accuracy," Engineering Inspection, 17 (2-10).
3.	Long, J. (1976), "Visual Feedback and Skilled Keeping: Differential Effects on Making the Printed Copy and the Keyboard," Ergonomics, 19 (1), pp. 93-110.
4.	Stammers, R., and Bird, J. (1980), "Controller Evaluation of Touch Input Air Traffic Data System: An Indelicate Experiment," Human factors, 22 (5), pp. 581-589.
5.	Kozinsky, E. J. (1981), "Human Factors Research on Power Plant Simulators," Proceedings of the 25th Annual Human Factors Society Conference.
6.	Swain, A. D. (1982), Personal Communication.
7.	Telecommunications Human Factors Proceedings.
8.	Agate, D. and Drury, C. G. (1980), "Electronic Calculators: Which Notation is Better?" Applied Ergonomics, 11.1, pp. 2-6.
9.	Marshall, E. C., Duncan, K. D., and Baker, S. M. (1981), "The Role of Withheld Information in the Training of Process Plant Fault Diagnosis," Ergonomics 24 (9), pp. 711-724.
10-13	Consultancy (not available for distribution).
14.	Kozinsky, E. J. (working draft), General Physics Corp. GP-R-23006 Contract W7405.
15.	Rouse, W. (1979), "Problem Solving Performance of Maintenance Trainees in a Fault Diagnosis Task," Human Factors 21 (2), pp. 195-203.
16.	Brooke, J. B., and Duncan, K. D. (1981), "Experimental Studies of Flow Chart Use at Different Stages of Prgram De-bugging," Ergonomics, 23 (11), pp. 1057-1091.
17.	Brooke, J. B., and Duncan, K. D., Effects of System Display Format on Performance in a Fault Location Task," Ergonomics, 24 (3), pp. 175-189.
18.	Marshal, E. C., Duncan, K. D., and Baker, S. M. (1981) "The Role of Withheld Information in the Training of Process Plant Fault Diagnosis," Ergonomics, 24 (9).
19.	Shephard, A., Marshall, E. C., Turner, A., and Duncan, K. D. (1977), "Diagnosis of Plant Failures from a Control Panel: A Comparison of Three Training Methods," Ergonomics, 20 (4).

Task	Source
20.	Brooke, J. B., and Duncan, K. D. (1980), "An Experimental Study of Flowcharts as an Aid to Identification of Procedural Faults," Ergonomics, 23 (4), pp. 287-399.
21.	Lees, F. P., and Sayers, L. B. (1976), "The Behavior of Process Operators under Emergency Conditions," in: Sheridan, T. B., and Johanssen, G., Monitoring Behavior and Supervisory Control, Nato Conference Series, Plenum Press, London.

5. Materials Provided for Subjects Participating in The Phase I Experiment.

The remainder of this appendix reproduces the material given to the subjects participating in the Phase I experiment described in Section 2.1 of the main text.

5.1 Categorization of Tasks

The 21 tasks used in this experiment have been divided up into three categories; Skill-, Rule- and Knowledge-based behavior. Each category contains 7 tasks. A brief description of these categories is given below:

5.1.1 Skill-Based Behavior

Skill-based behavior is an automated or subconscious pattern of behavior in a routine situation. An operator simply recognizes that a particular situation requires a normal response and then executes a skilled act which is more or less automatic. Errors may occur because of "manual variability," e.g., in Task 5 an operator may attempt to operate one control, miss it, and operate another. Alternatively, he may fail to discriminate correctly and operate the wrong control.

For the seven tasks in this category, all involve simple routine behaviors, and simple errors. Errors may be due to, e.g., low alertness, absent-mindedness, etc.

5.1.2 Rule-Based Behavior

This category deals with behavior in routine or familiar situations, where learned rules or procedures can be used. Using Rule based behavior requires recognition that the task requires a response, then associating the task with a previously encountered situation for which rules exist. The stored rules may then be used to execute the task.

Errors that may occur are responding to a familiar "cue" which is an incomplete part of the available information, leading to the use of inadequate/inappropriate rules; failing to recall the procedures correctly and/or totally (e.g., memory slip, forgetting an isolated act, a mistake among alternatives).

An example of the first type of error, caused by using incomplete information, is in Task 9. In this task, an operator using rules to diagnose a failure may make an error by believing he has recognized the failure before he has applied all the rules. Another error he could make is forgetting one of the rules. In Task 14, an operator reacting to a malfunc tion in a control room may omit an essential procedural step.

The rules themselves may be learned from trial and error, formed by causal reasoning, or prescribed as formal procedures (verbal or written).

5.1.3 Knowledge-Based Behavior

This level of behavior occurs in situations outside the individual's normal experience. It calls for intelligent problem solving, analysis of the situation, and planning. The individual must analyze the situation and decide on or deduce a task plan. Once the plan has been made, he or she can use Rule based behavior to achieve the goal of the plan.

Errors may occur at the information-collecting stage, so that the person can fail to collect enough information, he or she may make invalid assumptions, or even misinterpret information. For example, in Task 16, inadequate information may lead to incorrect or "premature" diagnosis of the malfunction.

Therefore, Knowledge based behavior is an evaluation of the situation and the planning of a proper sequence of actions to pursue the required goal. It depends upon fundamental knowledge of the processes, functions, and structure of the system in question. Thus, for example, in Task 19, the behavior of the subjects in this diagnostic task depends heavily on their knowledge of the plant.

5.2 Task Descriptions

Skill-Based Behavior

- Task 1

In an experiment investigating human recall performance with alphanumeric displays, a control group of 19 female subjects carried out the following task: A sequence of six digits (1-9, no zeros; no digit was repeated in any sequence) was displayed on a screen in front of the subjects from a slide projector. The digits appeared as white figures on a black background. Each digit sequence was visible for 2.6 seconds. Each subject then responded immediately by writing down the sequence during the 9.6 second interval before the next slide. Before each new slide, a verbal "Ready" signal was given.

Seventy-two slides were shown, taking roughly twelve minutes. The subjects' responses were scored correct if they remembered the digits in the correct order. The subjects were volunteers from a subject bank.

The error under consideration is an incorrectly ordered sequence of digits.

- Task 2

In a study of Quality Control inspector accuracy, a small wired unit was built, very similar to the units which the industrial inspectors routinely examined every day. The unit had approximately 1500 wires soldered to terminals. Thirty defects were placed throughout the unit in soldering, wiring and appearance, etc. Sixteen of the defects were soldering defects which in this particular context were highly important defects, i.e., with serious consequences for production if frequent. Seventeen inspectors were given a unit each, and allowed 3 hours to find as many defects as they could.

In their normal everyday work, the inspectors were of varying grades and responsibility, but they ranked higher than the shop inspectors and were in constant contact with supervisors in the shop and inspection department. In this way, they would have feedback about their work. No hard and fast selection procedure was in operation for hiring the inspectors at the particular factory under investigation.

The error under consideration is failing to find all of the defects.

- Task 3

In an experiment on typing performance, twelve female subjects (touch typists) each sat in a quiet cubicle with a typewriter. They were instructed, by means of an intercom, to type out a copy of a visually presented piece of text, 1000 characters in length. Each was instructed to type as fast as possible without exceeding an error rate of 1% (i.e. the maximum number of mistakes allowed was one mistake per hundred characters). Subjects were paid for their participation. The required error rate was that which was actually achieved by the typists.

- Task 4

In an evaluation of a new system for data transfer and display for airport traffic controllers (ATC's), fifteen controllers took part in exercises over three days, totalling just over five hours per subject. The controllers worked in groups of three. The study used a simulation of normal operations at Heathrow Airport.

The system displayed the controllers' data on a single cathode ray tube (CRT) screen. A radio enabled the subject to keep in contact with his colleagues and simulated aircraft. Data transfer and modification were enabled by a touch sensitive surface over the display screen itself. Thus, a controller could send, obtain, and update information by touching particular parts of the screen which were marked. For example, he might wish to update his information on the weather, approaching air traffic, etc.

The CRT Screen itself was 40 cm in diameter. The touching areas were 2 to 3 cm in diameter on one quadrant of the CRT screen.

Three types of errors could occur: miss touches, error touches (touching the wrong label), and illegal touches (touching a label out of sequence). The average touch rate was six touches per minute.

Exercises were of 1 1/4 hours duration, with a break between ranging from 1/4 hour to 1 1/2 hours. The main part of the experiment took place over two days.

The controllers were not experienced with the touch method described here, but the average ATC experience was ten years.

The error under consideration is a "miss" touch, where the controller misses the area to be touched.

- Task 5

In an extensive series of simulator exercises, performance of operators in reactor start-up and malfunctions was monitored. The simulator was of a Pressurized Water Reactor (PWR). The diagram below shows two controls, differing only by label and function. The two controls are 20 cm apart, but the partial rod control (right hand knob) was in fact inoperative, and this fact was known by operators.

The operators were experienced and worked in teams. A senior operator would be with the team but could obviously not monitor all the operators' actions at the same time.

The rod control (left hand knob) would be operated 50-100 times an hour in the start-up condition, and approximately 10 times during a malfunction situation. If the wrong knob was operated this had no deleterious effects of its own (as it was inoperative), although it meant that the correct switch had not been operated. The error probability required is the operation of the wrong control.

- Task 6

In a large organization (8-10,000 employees), one particular task involved soldering with a tensile type soldering iron, with the work piece firmly positioned in a holder. Soldering was performed manually and the task involved 10-100 solder joints depending upon the assembly in question. The task was carried out under 10x magnification.

The workers were female, seated operators, working in clean, well-lit conditions. The units were sub-assemblies for nuclear weapons. Quality was emphasized strongly, and there was no time limit for the task, i.e., they could take as long as they liked to achieve the best possible result.

The workers were in a stable environment with low stress except that the company would lose business if a bad product left the plant, and verbal punishment would then be given to the operator in question in front of others.

The staff turnover was very much lower than the normal commerical environment. Staff were told that the recording of errors was very important, but anonymous, in terms of records kept by personnel at the plants.

The error under consideration is a solder joint omission.

- Task 7

In an experiment looking at keying performance on a twelve-button key phone (see diagrams 1 and 2), seventy subjects who were due to have keyphones installed were tested over a two-year period. At testing sessions, each subject was required to "dial" 40 seven-digit numbers on a keyphone. A total of eleven tests were performed over a two-year period, with intervals between tests varying between one and twenty-two weeks. The tests were carried out at a center near the subject's offices.

Subjects were between 20 and 60 years of age. In the actual test each subject would search a list of telephone numbers which included 40 "tagged" for them to use (this was to simulate natural use: i.e., looking up a number in a directory and then phoning).

During the last six tests, learning effects were deemed to have become insignificant. Errors that could occur were: incorrect digit, transposed digit, missing digit, extra digit, double button depression, and premature handset replacement.

All subjects worked in offices making an average of approximately twelve phone calls per day. The error under consideration is making a keying error in the last six tests, i.e., when the user is fully experienced with the key phone.

5.2.2 Rule Based Behavior

- Task 8

In an experiment looking at performance with an electronic calculator (see diagram below), ten subjects were chosen from a large American company. Each was trained with feedback to use the calculator.

The ten subjects consisted of two subgroups, both of which had daily experience with calculators; one group coming from the accounting department and the other group from the engineering staff. The former were used to simple calculations, and the latter to more complex ones.

Each subject was given a test which was timed. The test consisting of ten questions ranging from very simple ones such as:

$(3,3 + 4,5)\ (5,2 + 6,1)\ (7,3 + 8,4) =$?

to the more complex ones such as:

$3,14 \times 435 \times 3,5\ (1,24)4 - (0,8)4/2 =$?

Each question was selected from the manual accompanying the calculator.

The error under consideration is the failure to obtain a correct answer.

- Task 9

In a simulated chemical process plant, a trainee in a diagnostic task faced with a panel with 33 instruments and 15 alarms (see diagram) which was actually a projection of a slide. The process plant simulated in this experiment included reaction, distillation, and the controls and features associated with chemical process plants.

Discriminating between simulated process failures entailed considering a number of indications, depending on the fault in question.

Trainees were instructed in the functioning of the instrumentation, the functions of control loops, etc. At the end of this description, trainees had to demonstrate that they had a good grasp of this knowledge. The trainees then learned the following rules to aid them in diagnosis:

1. Scan the panel to locate the general area of failure, i.e., feed, reactor/heat-exchange complex, column A or B.

2. Check all control loops in the area affected. Are there any anomalous valve positions?
3. High level in a vessel and low flow in associated take-off line indicates either a pump failure or valve failed "closed." If the valves OK then pump failure is probable diagnosis.

4. High temperature and pressure in column head associated with low level in reflux drum indicates overhead condenser failure - provided all pumps and valves are working correctly (rules (2) and (3)).

5. If the failure is in the reactor heat-exchange complex, determine whether it is in the reactor or the heat exchange system. A failure in the heat-exchange will produce symptoms in column A but not in B. A failure in the reactor will produce symptoms in both columns.

6. If the failure is in the feed system, check whether it is in stream X or stream Y. Because of the nature of the control system, a failure in the Y steam will produce associated symptoms in both the X and Y streams. A failure in the X stream will show symptoms in the X stream only.

After having learned the rules, eight trainees underwent a pretest in which they were given a list of eight failures and a plant circuit diagram. They were subsequently tested on these faults in a random order (i.e., the eight slides with faults on them were projected onto the screen). They had no feed back, and a limit of five minutes per problem.

Trainees then practiced with the eight failures but now with feedback on their performance, and an instructor who encouraged the use of the rules.

The next day, the trainees underwent the test in which ten failures were presented, these failures being the ones they had practiced on, two of the failures being repeated. No feedback or supervision was given here, and once again they had five munutes per problem.

The error under consideration is incorrect diagnosis (including failing to diagnose) in this final test.

- Task 10

In a batch chemical process plant the normal sequence of plant operations, e.g., filling vessels, opening and closing valves, etc., are control led by computer. When malfunctions occur (e.g., a valve sticks), the operator is required to intervene using the operator's control panel (OCP) and the call routine and display panel (CDP). Intervention requires the following steps:

1. Select call routine corresponding to the operation required by pressing the appropriate button on the CDP.

2. Press button A (top left, OCP).

3. Enter the address corresponding to the plant item to be manipulated, (addresses are of the form 342D, etc., and are entered via the keyboard on the OCP).

4. Enter one of the COM or M buttons if it is desired to change the status of the plant item from computer control to manual or vice-versa. (A plant item can only be manipulated from the panel if it is on manual control. After an operation or series of operations it is normal to return the item or items to computer control).

5. Press the O button on the OCP if it is desired to open (e.g., a valve) or stop (e.g., a pump). Or, press the CL button on the OCP if the valve is to be closed, or the pump started.

6. Press ENTER.

A mimic provides the operator of the state of the plant, and gives feedback when a manual control action is made such as opening or closing a valve. The mimic diagram is very large and has a high density of information. The address keyed in by the operator is also displayed on the CDP.

The nature of the operator's task is that of supervisory control, i.e., monitoring the state of the various batches and intervening where the computer cannot proceed. In general, the operators are highly autonomous and are not closely supervised. They have considerable responsiblity and reasonable pay. All the operators in this study were highly experienced.

Two types of errors are possible when entering an address. The operator may enter an invalid address, i.e., one for which no plant item exists. Alternatively the address entered may refer to a plant item, but not the one intended. In the former case he will always receive direct feedback in the form of an error message from the computer. In the latter, he will only receive feedback if he is monitoring the display on the CDP, and/or the mimic diagram which represents the plant item which has been erroneously operated.

The error of interest is the keyboard operation which could give rise to either of these failures.

- Task 11

This task is identical to Task 10, except that the operator is using the computer to change the value of a controller setpoint. In this case, after the enter button is pressed (step 6 in Task 10) the old and the new setpoints alternate in the display at the top of the CDP. If the operator agrees that the new value is correct, he presses the enter button a second time.

The error of interest is entering a setpoint which is outside the allowable range of the controller. In this case the computer will not accept the data and the operator will therefore obtain immediate feedback that he has made an error.

- Task 12

In the same situation as in Task 10 and Task 11 the operator may select a call routine button for which no routine is currently available in the computer (step 1 in the procedure described in Task 10). In this case the computer will display a "no routine" error message and hence the operator will always receive feedback after an error of this type.

The error of interest is that of initially selecting an invalid call routine button.

- Task 13

In the same situation as Task 10 and 12, a "control error" may occur when an illegal operation is attempted by certain call routines, e.g., an operator may try to open or close a valve that is still on computer control (and not on manual, as it must be for such an operation to work). It is expected that an operator should know the state of an item he is trying to manipulate from the mimic.

The error in this task is any illegal operation of the type described above.

- Task 14

In a study of operator's performance in a BWR simulator (see diagram 1), operators worked in teams. Ten malfunctions could occur. One of these ten was "loss of shutdown cooling." The malfunction, and the correct operator responses, are listed below:

Loss of Shutdown Cooling:

Cause: loss of operating residual heat removal (RHR) pump(s).

Required Response: Shift to alternate RHR loop for reactor decay heat removal capability.

Operator Responses:

1. Secure failed pump(s)
2. Attempt to restore operating loop to service.
3. Shift to standby loop.

Twenty-five teams from three separate utilities performed each of the ten malfunctions, including the one above. The mean control room experience was 5.21 years, and the mean age 35-38 years.

In the malfunction above, all three procedural steps had to be carried out in order to rectify the malfunction. The error under consideration is omitting a single procedural step. One out of sixteen operators thought that the procedures were inadequate.

5.2.3 Knowledge Based Behavior

- Task 15

Twenty maintenance trainees took part in an experimental study of "trouble shooting" of geographically displayed networks of logic "AND" units (see diagram 1).

All units (numbered 1-49 in the diagram) are connected to other units in the network, by 1,2,3, or 4 lines. If all these lines or inputs to a unit are 1 (the value of an input can only be either 1 or 0), then the respective unit's output will be one, unless the unit is faulty in which case its output will be 0. If a unit is faulty and gives out a 0 value, then the unit that receives the faulty unit's output will also give outputs of 0. In this way, a zero value from a faulty unit will propagate to other units in the network, the extent of which depends on the connections in the network, and the position of the faulty unit. In the diagram, the propagation will be in the direction from left to right towards the final outputs (43-49).

The task of the trainee was to solve problems given the network as shown, with the final outputs also shown. The trainee would therefore have to "back-track" from the final outputs and see where the single fault could have originated from (i.e. which unit was faulty). The trainee on a VDU, could input the numbers of two units and check the line between th see if it carried a "1" or a "0". In this way, he could locate the faul n this experiment, only one unit failed per problem. The experiment usec ee sizes of network; 9 units; 25 units; 49 units.

Twenty subjects performed four trials of ten problems each, all tasks in one session. Trials 2 and 3, however, utilized computer aiding as outlined below:

A computer aiding algorithm acted as a sophisticated bookkeeper using the topology of the network, and known outputs to eliminate units that could not possibly be faulty. Also, it iteratively used the tests run by the trainee to further eliminate units from consideration by drawing an "X" through them. The time the algorithm took to do this could be up to 20s per problem, and this was"charged" to the subject in that the clock timer was left running. During this time, the trainee could study the problem. Of the four trials, trials 2 and 3 utilized computer aiding.

The error under consideration is failing to achieve the correct solution on a problem with 49 nodes while in the aiding condition (trials 2 and 3).

- Task 16

In an experiment investigating program "debugging", three programs represented instructions to a computerized warehouse, and consisted of sequences of instruction to a computer-controlled truck to "turn left" or "turn right" or "go forward". The figure below is a map of the warehouse which is organized on a grid system. The four boxes at the corners represent loading bays, and the programs described alternative paths from one loading bay to another, and two such paths are marked on the figure.

Subjects (first year psychology undergraduates) were shown the diagram of the warehouse and its functioning explained. A program listing (see Diagram 2 - if-then-else structure) was then shown to them and it was explained how the program described two paths: one short, with many turns; one long with few turns. One path or other could be taken but not both, depending on the outcome of the conditional statement.

Subjects were then told that one of the alleyways had become blocked and consequently the truck had not reached its destination. They were required to find the blockage by finding the instruction which had failed to be executed. They did not know which path had been taken, but they could cause the program to be rerun under the same conditions (thus taking the same path) as many times as they liked. On each occasion that they reran the program they were to ask the computer whether particular program statements had been executed or not. However, they could only request information about one program statement

per run. Each subject used a visual display unit, and having entered the label of the statement they wanted to test, (e.g. B, C, E, etc), would be informed by the computer whether the statement had been executed properly or had failed, or had not been reached. After each test, the computer asked the subject if he knew which unit had failed. If he typed "N", the subject would continue testing. If diagnosis was incorrect, the subject was informed and allowed to continue. If diagnosis was correct, the subject could state a new problem.

Each subject solved five practice problems using the program shown. Another program was then demonstrated and the subject solved fourteen problems, this routine being repeated with a third program. Thus, each of five subjects had five practice trials, followed by twenty eight test problems.

Each subject was paid for participating and received extra rewards for completing problems within a certain time limit. Incorrect identification of the fault carried a time penalty.

There were four possible diagnostic errors:

1. Performing a test on a path that had already been eliminated.

2. A wrong guess at the location of the fault.

3. Performing a test above or below the established bracket

4. A correct identification of the location when the bracket enclosed more than one statement (i.e. incorrect inference).

Each of the two test programs contained two conditional statements, no feedback loops and fourteen statements. Each statement was labelled by a letter.

In considering the twenty eight problems, the error under consideration is that of a wrong guess at the location of the fault (i.e. diagnostic error 2).

- Task 17

In a fault finding experiment with ten subjects from an applied psychology department, the task consisted of finding a faulty component in a network of logic "AND" units (see Diagram 1). A total of six networks were used, three with twenty-four units. For these two levels of complexity of network (i.e. number of units), there were three types of representation, but each had the same information content. These units were displayed on a VDU.

In each problem one of the units was randomly designated (by a computer) as faulty, but the output units were always assumed to function correctly

(units 14, 24, 34, 44, in the 16 unit network, and 16, 26, 36, 46 in the 24 unit network). A colored bar by each output indicated its status: a green bar (light in diagram) indicated it was working, and a black bar that it was not. A logic unit only works if all inputs are 1; otherwise an output of zero is generated. If a unit is faulty, then even if all inputs to it were 1, it would still generate a zero output (see Diagram). Thus, if a unit was faulty, it would propagate a zero output to one or more of the final outputs. The subjects' task was to locate the faulty unit, by interrogating the computer. A VT100 visual display unit was used to enter the connection to be tested and to display the signal carried by the connection. When no further tests were desired, the subject pressed the "enter" key in response to the "TEST?" prompt from the computer and diagnosed the faulty unit. Appropriate feedback was given after each diagnosis, with the correct solution being displayed after a wrong diagnosis, and a new problem initiated. The VT100 VDU could contain up to 10 tests.

Subjects were tested individually. They were introduced to the task and the functioning of the networks was explained. The subject solved the first problem with the experimenter there to clarify any procedural points. No help was given to the subject in solving the problem. The subject then solved a further five problems, one on each network, with the experimenter present. The experimenter then told the subject that he would do another 30 problems (five on each network) and then left. Subjects were told to concentrate on solving the problems correctly but not to spend an excessive amount of time on any one problem.

There were four types of diagnostic error, as below:

1. Redundant tests.

2. Failure to utilize positive information.

3. Inadequate information at diagnosis (i.e. premature diagnosis).

4. Tests on unit not connected to known zero line.

Error number three is the one most likely to lead to incorrect diagnosis, and in these terms is likely to be the most costly.

Subjects were paid for taking part in this experiment, and took between 40 and 120 seconds on each of the 30 problems.

Error number 3 is the one under consideration.

- Task 18

In the same situation as Task 9 (simulated process plant - refer to Task 9), the same trainees took part in a second test after they had finished the one outlined in Task 9. In this second test, they were presented with sixteen

of the earlier faults, and eight new faults they had never seen before. The same testing conditions applied (i.e. 5 minutes maximum diagnosis time). The error under consideration is once again incorrect diagnosis (including failure to diagnose), for the new problems only.

- Task 19

In an experiment looking at three methods of training for process plant fault diagnosis, one group was given training in the form of a technical "story", as typically given by conscientious training officers in the chemical industry. This training consisted of tracing the flow of product through the plant, describing the functions of the various items of equipment (see Diagrams 1 & 2) and discussing the effects of the different control groups. The simulated plant was typical of that used in chemical plants. It incorporated 33 instruments and 15 alarms (similar to that used in Tasks 9 and 18). Basically, the process was as follows: two feeds are reacted and the gas from the reactor is compressed, and then distilled in column B. The liquid from the reactor is distilled in column A where the component is re-cycled back to one of the feed tanks.

"Symptom arrays" were recorded on slides which were then back projected onto a screen using a random access projector. Diagram 3 shows the plant failures used in this experiment. To discriminate between panel arrays entailed consideration of a number of features, the number depending on the fault in question.

Subjects in the "theory" group described above (those given the technical story introduction), were required to demonstrate their knowledge by tracing the flow of materials through the plant, and by specifying the effects throughout the plant of two faults (which did not appear in the test under consideration). They were given a short test on eight faults.

The ten subjects were next trained to recognize eight faults on an adaptive cumulative training method, i.e., each subject was trained to recognize two faults until he knew them well. Then, he was trained to recognize another fault, and subsequently tested on all three faults. This procedure continued until each subject could reliably recognize all eight faults.

After the introduction to the plant, the subjects had a short test on eight faults and subsequently learned those faults. They were then again tested on the eight failures resulting from the faults, each failure being presented four times in a randomized block of 32 trials.

Finally, each subject was given a test on sixteen failures, eight of which had been learned and eight which had never been seen before. The failures were presented randomly. The subjects were given a list of the 16 failures from which to select their responses.

The error under consideration is incorrect diagnosis in the final test.

- Task 20

In an experimental study of flowcharts as an aid to the identification of procedural faults, the following scenario was constructed:

A multigrade petrol (gasoline) pump was controlled by a hypothetical simple computer. The computer had eight input switches, through which signals from the outside could be received, and eight output switches to switch the external equipment on or off. It also had an information processing unit to obey programs of instruction; 20 "read only" memories, and 20 registers with read and write facilities (see Diagram 1).

Each subject read short descriptions of the hardware involved, with examples to follow. The subjects were then tested on their knowledge. The procedure the computer was to follow was shown and explained to each subject by means of a flowchart (see Diagram 2). Subjects were then told that when the flow-chart had been translated into a program for the computer, six attempts had been made and six faults had been made in the translation (e.g. such as a step being missed out, wrong registers being used, or the order of two steps being reversed). The subject was then shown the list of the six faults. The total presentation time so far was 15-25 minutes.

The task consisted of three separate problems (three of the six faults), e.g. "Correct petrol grade was given, but the wrong price was charged;

2* was charged at the price of 3*
3* was charged at the price of 3
4* was charged at the price of zero."

Each of the faults lay in a different part of the program, and the symptoms were easily distinguishable. For each fault, however, another fault in the list lay in the same area of the program and had similar symptoms (two faults related to the setting of valves by the computer; two faults related to the checking of inputs during the delivery of petrol; two faults related to the calculation of the price).

Therefore, for each problem it was possible to choose the actual fault; the fault lying in the same area but with different symptom detail; the four faults neither in the same area nor with the same symptoms.

Ten minutes were allowed to solve each problem (a one minute warning was given at 9 minutes). Subjects had to write down their diagnoses.

Twelve subjects were used; 6 from a subject bank (each paid one British Pound); 6 from the research staff. The error under consideration is incorrect diagnosis.

- Task 21

Rise in Inlet Gas Temperature in a Gas Cooled Nuclear-Reactor Simulator

This fault caused a slow rise in inlet gas temperature, increasing to a "trip" level in 20-30 seconds. An indication that this fault had occurred was shown on a three-point indicator/reader. The time to print the values of points 1, 2, and then 3, in succession, followed by point 1 again, took 10 seconds. This meant that some six seconds could elapse with the operator having no indication of the actual value of the inlet gas temperature.

Another indication of this fault was given by a change in indicated steam pressure on the desk mounted steam pressure gauge.

The operator had to respond within 28 seconds, and any response after this time was considered a failure. The error of interest is failing to respond (by pressing the "trip" button) within 28s of the onset of the fault.

5.3 Factors Giving Rise To Errors in Process Plants

5.3.1 Introduction

Human beings are involved in a number of areas of plant operation where errors can adversely affect system safety or system availability. Many forms of erroneous action exist. For example. steps can be missed out in a procedure, or carried out in the wrong order. The operator may do the right actions on the wrong piece of equipment, or may "improve" procedures such as that they no longer achieve the required objectives. For the purpose of this document, an error is defined as action (or lack of action) by an individual which reduces the safety, reliability or availability of a process or power plant. In most cases, this will correspond to a failure to carry out a specific task correctly.

The likelihood that an error will occur is largely determined by particular combinations of factors which either enhance or degrade the individual's ability to perform the task. These factors are known as Performance Shaping Factors or PSF.

It seems likely that a relatively small number of PSFs account for most of the variation in error rates observed in plant tasks. Examples of such factors are the time available to perform the task, quality of procedures and the level of training of the operator.

Because it is difficult to determine the relative importance of the various PSFs on task failure using analytical approaches, we are interested in obtaining the opinions of "experts" who either have direct experience of plant operations or who have some knowledge of human factors. Co-operation is therefore requested in carrying out the judgements to be described in subsequent sections, in order to assist us in obtaining as wide a consensus as possible.

5.3.2 Performance Shaping Factors

It is assumed that the following factors affect the likelihood that a task will be successfully carried out.

The factors are not presented in any particular order of importance.

A. Quality of Procedures. The existence of clear procedures and guidelines regarding the actions that need to be taken to achieve specified objectives given certain conditions. In addition, the procedures will also define the areas of responsibility to be undertaken by each operator, particularly in operations involving several individuals.

B. Relevance and Comprehensiveness of Training. The extent to which the operator has received relevant training necessary to carry out the task, together with on-the-job experience which may assist in coping with unexpected contingencies.

C. Time Available to Perform Task. The time constraints within which the task, or any of its constituent parts, must be performed.

D. Quality of the Information Available to the Operator Regarding the System State. The extent to which the operator is effectively provided with information regarding the state of the system. This information may be general, e.g. the current state of various temperatures, pressures, flows etc, or may be feedback indicating the results of specific control actions. The quality of the information available may be measured in terms of its comprehensiveness, and the extent to which it is presented to facilitate understanding and to minimize the possibility of information overload.

E. Quality of Supervision and Checking. The extent to which the operator's actions are monitored and checked by an independent individual, e.g. a supervisor or inspector.

F. Motivation. This factor is influenced by the degree to which the operator's personal needs are satisfied by the work that he carries out. Typical task related factors which are important are the variety of skills exercised, the degree to which a task is perceived to be meaningful, the degree of autonomy of the operator and the knowledge of results or feedback that is provided.

More general job related factors are the possibilities for advancement and recognition. Poor working conditions, and problems of interpersonal relationships are examples of demotivating factors.

5.3.3 Weighting Procedure

You are asked to carry out the following procedure for each of the three broad categories of task which have already been described in a separate note, i.e., Skill, Rule and Knowledge based tasks.

1. From your knowledge and experience, rank the PSF provided in the order of their importance in affecting the probability of success for the general category being considered (i.e., Knowledge, Rule and Skill based), starting with the most important factor.

2. Enter the ranks, together with the appropriate PSF letter (A to F) on the results sheets provided. The highest rank (i.e., the most important factor) should be given the number 6, and the lowest (i.e., least import ant) 1. Thus the result sheet for a task category might look like this for example:

PSF	RANK	
B	6	<—— most important
A	5	
C	4	
F	3	
E	2	
D	1	<—— least important

Do not be afraid to revise the rankings until you are satisfied they are in the correct order.

3. Take the lowest ranked factor (e.g., D in the example above), and assign an arbitrary weight of 10 to represent the effect of variations in that factor in determining the overall likelihood of task success. Take the most important factor (E in the example) and assign a weight which indicates how much more important this factor is than the lowest factor in determining task success. For example, if factor E was 3 times more important than the lowest factor it would be assigned a weight of 30. If 3 1/2 times more effective, the weight would be 35.

This process of successive comparison with the lowest factor is then repeated with all the remaining factors.

Thus, if the result below were obtained, it would indicate that the factors E, F, C, A, B, were respectively 3, 7, 9 1/2 and 20 times more important than D in determining the likelihood of success for the task category being considered.

PSF	Weight
B	200
A	100
C	95
F	70
E	30
D	10

4. Repeat steps 1 - 3 for Skill-, Rule- and Knowledge-based task categories. You should therefore have 3 sets of weights by this stage.

N.B. Check for consistency within weights and rankings. If you feel that an important factor has been omitted, state it and give its weight relative to the lowest factor.

5.3.4 Scales for Assessing Performance Shaping Factors

Using the scales shown below, assign a rating of the "quality" of these factors for the task under consideration, in the form of a number between Zero and 100. In this context, "quality" means the degree to which the factor concerned either degrades the likelihood of success (less than 50 on the scale), or enhances the likelihood of success (more than 50 on the scale). A scale value of 50 therefore represents the midpoint between the extremes where a particular PSF is as good as it can be in terms of its effect in maximizing success (100 on the scale) or as bad as it can be in reducing the likelihood of success (Zero on the scale).

Example:

A task might have the following PSF profile:

Quality of procedures	50
Relevance and comprehensiveness of training	80
Time available to perform task	30
Quality of information regarding system state	70
Quality of supervision and checking	15
Motivation	60

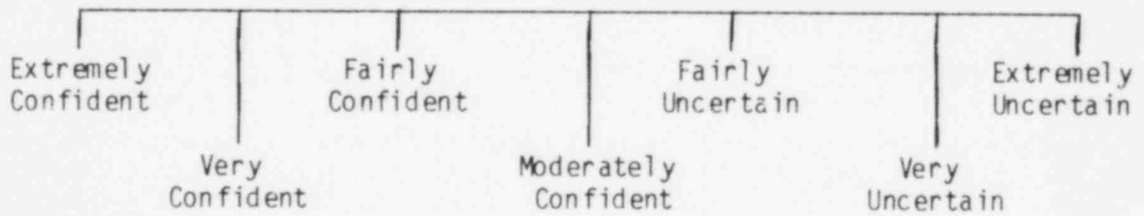
This would indicate that the procedures are acceptable but not out standing, the operators well trained, but there is a limited time available to perform the task. The operator has good indications of the state of the system, e.g., via a control panel, but there is an inadequate level of super vision and checking. The operators are somewhat more motivated than average.

Notes:

1. Indicate your rating for the task of interest by writing a number in the space designated.

2. Even if there is no information in the task descriptions to act as a guide, you should still attempt a rating.

3. Indicate your degree of confidence on the following scale for each rating:



N.B. Moderately confident is equivalent to moderately uncertain.

APPENDIX B

DEFINITIONS OF PSFs USED IN THE FIELD STUDY

DEFINITION/DESCRIPTION OF PERFORMANCE SHAPING FACTORS

1) Design

	Good	Poor
i) Displays	Easy to read and understand and accessible; Make sense; Easy to relate to controls; Alarms discriminable, relevant, coded; Mimic display.	Hard to read, difficult to interpret; inaccessible; Confusing; Not directly related to controls; Alarms confusing, irrelevant, not coded; Nonrepresentational display.
ii) Operator Involvement	Operators have say in modifications.	Little or no say.
iii) Automation of Routine Functions	Highly automated - operators act as systems managers.	Low level of automation - operators perform many routine functions.

2) Meaningfulness of Procedures

	Meaningful	Not Meaningful
i) Realism	Realistic; especially the way things are done.	Unrealistic; not the way things are done.
ii) Location Aids	Location aids provided.	Few or no location aids.
iii) Scrutibility	Procedures keep operators in touch with the plant.	Procedures do not.
iv) Operator Involvement	Operators involved in developing procedures.	Not involved.

3) Role of Operations

	Primary	Not Primary
i) Accountability	All other functions report to operations supervisor.	Only operations staff report to operations supervisor.
ii) Relationship to Maintenance and Other Functions	Good relations.	Antagonism.
iii) Paperwork	About right.	Excessive.
iv) Operator Involvement	Operators have a say in how the place is run.	No say.

4) Teams

	Present	Absent
i) Shifts	Allow teams to say together.	Prohibit team formation.
ii) Roles	Well defined.	Poorly defined.

5) Stress

	Functional	Level Not Functional
i) Shifts	No jet lag.	"Permanent jet lag."
ii) Time Available	Adequate.	Too little.
iii) Operating Objectives	No conflict.	Conflict.

6) Morale/Motivation

	Good	Poor
i) Status of Operators	Treated as professionals.	Treated as laborers.
ii) Career Structure	Operators can find best level in organization.	Peter Principle operates.

7) Competence

	High	Low
i) Training	Operators generally well trained in emergency procedures.	Poorly trained in emergency procedures.
ii) Certification	Peer review is used.	No peer review.

APPENDIX C

Task Definitions and Descriptions for Data Collection Sessions

This appendix contains two separate groupings of tasks, level A and level B tasks. Level A tasks pertain to system level tasks. Level B tasks include components, displays, instruments, and control tasks.

Assumption to be used by judges in assessing Level A tasks:

You are to assume the following for the tasks below (Level A 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 their 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 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 procedures 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 diese 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 acident (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?

Assumptions to be used by judges in assessing Level B tasks:

You are to assume the following for the tasks below (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 their current job position for at least six months.
- No one involved in performing these tasks is wearing any type of protective clothing.

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 10 annuciators 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.

APPENDIX D

Procedure for Ordering MAUD5 from the Decision Analysis Unit

The Decision Analysis Unit, London School of Economics, has supplied the following information on how to order MAUD5.

If you wish to use only one copy of MAUD5 on a single computer installation, under the terms of the End User Licence Agreement, the price of a non-exclusive licence is 500 pounds sterling* (or 334 pounds sterling* for bona-fide educational users who wish to use MAUD5 for teaching and research purposes only). These prices are current as of February 1984, and may be subject to change. The price includes supply of MAUD5 on any medium specified on the order form.

If you wish to have in use more than one copy of MAUD5, you can compute the cost of the licence by referring to the figure on the next page. Enter the number of copies you wish to have in use on the horizontal axis and read off the licence price on the vertical axis.

Ordering Information

To order MAUD5, enter on the End User Licence form reproduced at the back of this appendix, the number of copies of MAUD5 you wish to have in use and the appropriate price paid. (The Decision Analysis Unit will supply the serial number when MAUD5 is sent to you.)

You will find a blank End User Licence form at the end of this section. Send the form to the Decision Analysis Unit signed by you (or the authorized representative of your institution) together with:

1. The order form reproduced on the next two pages of this appendix.
2. Payment, or an official order form from your institution.

The Decision Analysis Unit will then dispatch MAUD5 and a copy of the completed End User Licence to you.

---

*Approximately equal to $750.00 and $500.00 (U.S), respectively, as of this date.

To: The Unit Secretary, Decision Analysis Unit, London School of Economics and Political Science, Houghton Street, London Wc2A 2AE, England.

From

NAME: DATE:

ORGANISATION: TELEPHONE:

ADDRESS:

- -

Please send me a copy of MAUD5.

☐ I enclose payment/an official order form and a signed End User License Agreement stating the number of copies I wish to have in use.

☐ I would like to discuss with you/you to quote for a special version of MAUD5. Details of my requirements are appended.

MAUD5 is currently available in the following forms. Please indicate the one you are interested in:

☐ IBM personal computer, requires DOS 2.0, BASIC, 64K RAM and a printer. Indicate which medium you require:

☐ 5-1/4-inch disk ☐ 8-inch disk

*

☐ CP/M, screen-oriented.
Requires CP/M operating system, 56K RAM and a 132-character printer. Versions readily available for North Star HORIZON and ADVANTAGE, SUPERBRAIN, TELEVIDEO and APPLE II (which must have a Z80 board and a 80-column card).

---

*Please specify the VDU you will be using:

Fill in the ASCII values for the following control functions on your VDU:

Cursor Right: Cursor Up:

Cursor Left: Cursor Down:

Clear Screen and Home Cursor:

If you require a CP/M version for another Z80-based computer or any of its family, a transfer fee may be charged.

Please specify the make and type of your computer:
Indicate which medium you require:

☐ 5-inch (hard-sectored) ☐ 5-inch (soft-sectored)

☐ 8-inch

☐ single density ☐ double density

Please give some very general idea of the type of application in which you think you may wish to use MAUD5:

- -

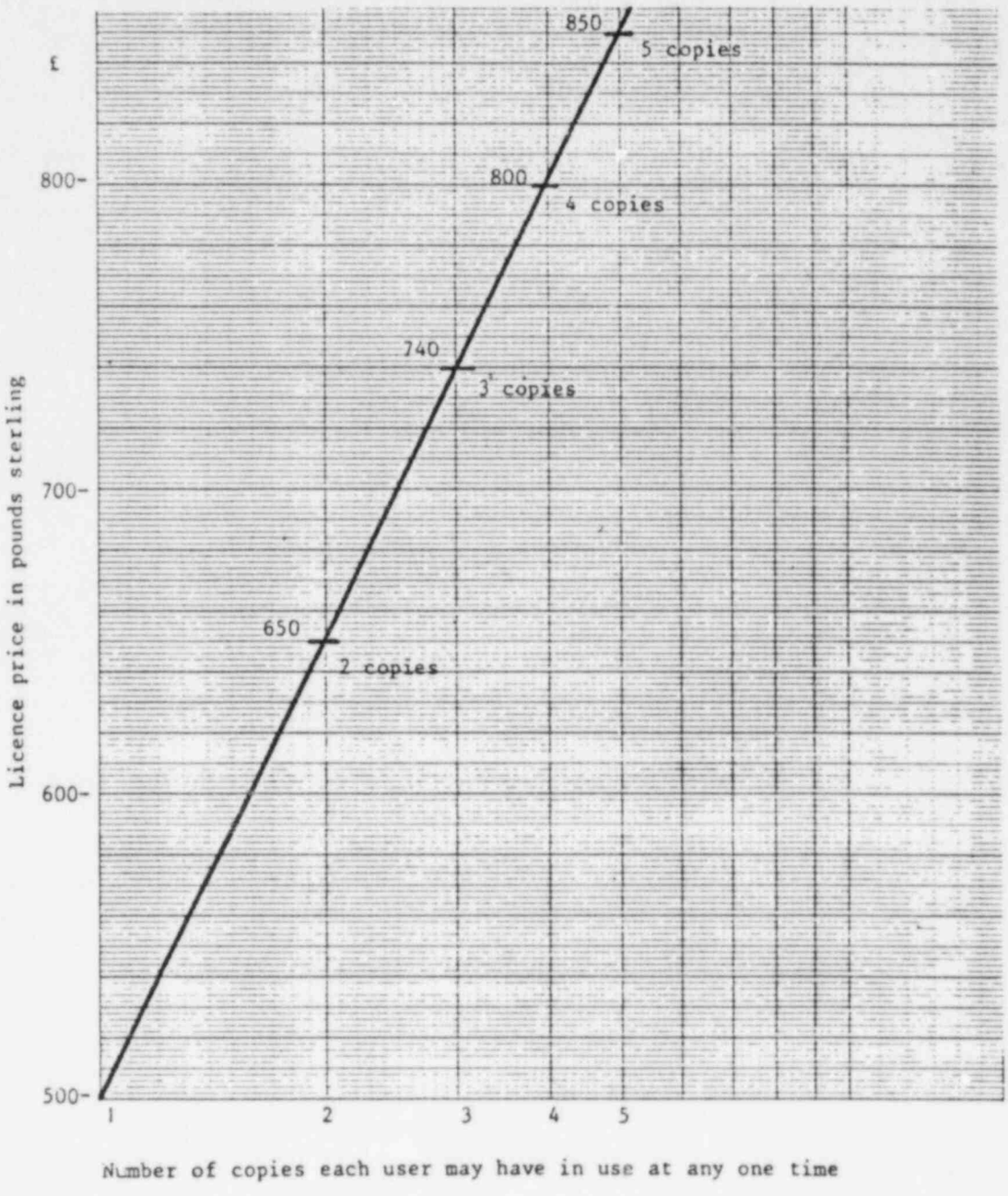


Figure D.1 MAUD5 end-user licence prices (February 1984).

- Prices for number of copies greater than five are prorated.
- Bona fide educational users are given a 33% discount on the prices shown above.
- All prices exclude VAT (currently at 15%) and are subject to change without prior notice.

The London School of Economics and Political Science
(University of London)

Decision Analysis Unit
Houghton Street
London WC2A 2AE

Telephone : 01-405-7686
Telex : 24655 BLPES G

END-USER LICENCE AGREEMENT FOR SOFTWARE PRODUCT

WHEREAS , the Decision Analysis Unit, London School of Economics and Political Sciences, Houghton Street, London WC2A 2AE, hereinafter referred to as DAU develops proprietary computer programs and licenses the use of such proprietary computer programs, together with or a part from accompanying copy-righted media material and documentation and;

WHEREAS, ...
of
City of
County of
State of
Country of

hereinafter referred to as End User, desires to obtain a license from DAU of the type aforesaid and in return is willing to abide by the obligations and fee agreements pertaining thereto.

NOW THIS AGREEMENT WITNESSETH AS FOLLOWS:

ARTICLE I. EXCLUSIVE SOURCE. The End User shall obtain DAU product materials covered by this agreement as set out in the schedule below through DAU and no other source. DAU product materials include, but are not limited to, manuals, license agreements, and media upon which DAU proprietary computer programs are recorded. Except for archival copies, ad defined in ARTICLE III of this agreement, End User shall make no copies, of any kind, of any of the materials furnished by DAU, unless specifically authorized to do so in writing signed by the director or deputy director of DAU.

ARTICLE II. NONDISCLOSURE. The End User agrees not to transfer, dispose of, publish, display, assign, sublicense, disclose, or otherwise make available the Licenced Software including but not limited to program listings, object code and source code, in any form, to any person other than End User or DAU employees, without prior written consent from DAU, except with the DAU's permission for purposes specifically related to the End User's licensed use of the Licenced Software.

End User is responsible for and agrees to pay DAU for any damages or losses due to the unauthorized copying or disclosure of the Licenced Software. End User recognizes that unauthorized copying or disclosure of the Licenced Software will cause irreparable injury to DAU and that DAU shall be entitled to, among other things, enjoin End User from any such activities.

ARTICLE III. ARCHIVAL COPIES. The End User may make archival copies of those portions of DAU's product(s) covered by this agreement that are provided on machine readable media, provided such copies are for the End User's personal use and that no more than one such copy is in use at any time unless End User has paid for multiple copy use as described in ARTICLE IV of this agreement.

ARTICLE IV. MULTIPLE COPY USE. DAU End user licences are applicable to a single micro-computer installation. In the event End User intends to use DAU product or any part thereof on more than one micro-computer, the required fee for such multiple use must be paid and the number of copies which the End User may accordingly have in use at any one time be recorded at the end of this agreement.

ARTICLE V. LIMITED WARRANTY POLICY. DAU warrants that all materials furnished by DAU constitutes an accurate manufacture of DAU product and will replace any such DAU furnished material found to be thus defective, provided such defect is found within ninety days of purchase by End User. However, DAU makes NO express or implied warranty of any kind with regard to performance or fitness of purpose for any DAU product. Furthermore, DAU is NOT responsible for any loss or inaccuracy of data of any kind nor for any consequential damages resulting therefrom whether through DAU negligence or not. DAU will not honour any warranty where DAU product has been subjected to physical abuse or use in defective or non-compatible equipment.

ARTICLE VI. UPDATE POLICY. DAU may, from time to time, revise the performance of its products and, in so doing, incur NO obligation to furnish such revisions to any customer. At DAU's option, DAU may provide its End Users with a revision newsletter and/or update information from time to time.

ARTICLE VII. GOVERNING LAW. This Agreement shall be interpreted in accordance with English law. In the event that any part of this Agreement is invalidated by court or legislative action, the remainder of this Agreement shall remain in binding effect.

ARTICLE VIII. LEGAL FEES. In the event of legal action brought by either party, the prevailing party shall be entitled to reimbursement of legal fees.

ARTICLE IX. ENTIRE AGREEMENT. This Agreement consistutes the entire agreement between the parties and supersedes any prior agreements. This Agreement may only be changed by mutual written consent.

_____ _____

Date For Decision Analysis Unit, London School of Economics and Political Science

_____ _____

Date For End User

SCHEDULE. DAU product materials covered by this agreement:
DAU product name: MAUD5

Description of media on which DAU product is supplied:

Number of copies which End User may have in use at any one time:

Valid End User serial number:

Fee paid to DAU by End User:

NRC FORM 335 (6 83) U S NUCLEAR REGULATORY COMMISSION

BIBLIOGRAPHIC DATA SHEET

1 REPORT NUMBER (Assigned by TIDC and Vol No. if any)
NUREG/CR-3518
BNL-NUREG-51716
Vol. II

2 Leave blank

3 TITLE AND SUBTITLE
SLIM-MAUD: An Approach to Assessing Human Error Probabilities Using Structured Expert Judgment, Volume II: Detailed Analysis of the Technical Issues

4 RECIPIENT'S ACCESSION NUMBER

5 DATE REPORT COMPLETED
MONTH: May YEAR: 1984

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

7 DATE REPORT ISSUED
MONTH: July YEAR: 1984

8 PERFORMING ORGANIZATION NAME AND MAILING ADDRESS (Include Zip Code)
Brookhaven National Laboratory
Department of Nuclear Energy
Upton, NY 11973

9 PROJECT/TASK/WORK UNIT NUMBER

10 FIN NUMBER
FIN A-3219

11 SPONSORING ORGANIZATION NAME AND MAILING ADDRESS (Include Zip Code)
U.S. Nuclear Regulatory Commission
Human Factors and Safeguards Branch
Washington, DC 20555

12a TYPE OF REPORT
Formal

12b PERIOD COVERED (Inclusive dates)

13 SUPPLEMENTARY NOTES

14 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.

15a KEY WORDS AND DOCUMENT ANALYSIS

15b DESCRIPTORS

subjective expert judgment
performance shaping factors
human factors

16 AVAILABILITY STATEMENT
Unlimited

17 SECURITY CLASSIFICATION (This report)
Unclassified

18 NUMBER OF PAGES

19 SECURITY CLASSIFICATION (This page)
Unclassified

20 PRICE
$