NUREG/CR-4016
BNL-NUREG-51828
Vol. 1

APPLICATION OF SLIM-MAUD: A TEST OF AN INTERACTIVE COMPUTER-BASED METHOD FOR ORGANIZING EXPERT ASSESSMENT OF HUMAN PERFORMANCE AND RELIABILITY

VOLUME 1: MAIN REPORT

E.A. Rosa, P.C. Humphreys, C.M. Spettell,
and D.E. Embrey

Date Published — September 1985

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-4016
BNL-NUREG-51828
Vol. 1
RX, AN, RG, 1S

APPLICATION OF SLIM-MAUD: A TEST OF AN INTERACTIVE COMPUTER-BASED METHOD FOR ORGANIZING EXPERT ASSESSMENT OF HUMAN PERFORMANCE AND RELIABILITY

VOLUME 1: MAIN REPORT

E.A. Rosa,* P.C. Humphreys,+ C.M. Spettell,* and D.E. Embrey†

+Department of Social Psychology
London School of Economics and Political Science
Houghton Street
London WC2A, England

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

Manuscript Completed — March 1985
Date Published — September 1985

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

Available from
Superintendent of Documents
U.S. Government Printing Office
P.O. Box 37082
Washington, DC 20013-7982
and
National Technical Information Service
Springfield, Virginia 22161

ABSTRACT

The U.S. Nuclear Regulatory Commission has been conducting a multiyear research program to investigate different methods for using expert judgments to estimate human error probabilities in nuclear power plants. One of the methods investigated, derived from multi-attribute utility theory, is the Success Likelihood Index Methodology implemented through Multi-Attribute Utility Decomposition (SLIM-MAUD). This report describes a systematic test application of the SLIM-MAUD methodology. The test application is evaluated on the basis of three criteria: practicality, acceptability, and usefulness.

Volume I of this report presents an overview of SLIM-MAUD, describes the procedures followed in the test application, and provides a summary of the results obtained.

Volume II consists of technical appendices to support in detail the materials contained in Volume I, and the users' package of explicit procedures to be followed in implementing SLIM-MAUD.

The results obtained in the test application provide support for the application of SLIM-MAUD to a wide variety of applications requiring estimates of human errors.

Previous Reports in the SLIM-MAUD Research Program:

"The Use of Performance Shaping Factors and Quantified Expert Judgments in the Evaluation of Human Reliability: An Initial Appraisal," NUREG/CR-2986, 1983.

"SLIM-MAUD: An Approach to Assessing Human Error Probabilities Using Structured Expert Judgment, Volume I: Overview of SLIM-MAUD and Volume II: Detailed Analysis of the Technical Issues," NUREG/CR-3518, BNL-NUREG-51716, 1984.

CONTENTS

CONTENTS (CONTINUED)

TABLES

ACKNOWLEDGMENTS

We would like to acknowledge the contributions of a number of individuals and organizations* who provided assistance that was essential to the completion of the work presented in this report. William J. Luckas, Jr. of Brookhaven National Laboratory (BNL), contract monitor, and Dr. Thomas G. Ryan of the U.S. Nuclear Regulatory Commission (NRC), program manager, provided encouragement, support, and guidance. Technical review was provided by Dr. John N. O'Brien of BNL. Clerical assistance was provided by Sharon M. Moore.

Special thanks are due to the eight individuals in the Division of Nuclear Energy, BNL, the two General Physics (GP) instructors, and the six staff members of the NRC who, by participating in the SLIM-MAUD sessions, provided the data analyzed in this report. Recognition is also due to K. Rea of Human Reliability Associates for performing some of the reported analyses, L. Weston of Sandia National Laboratories for technical guidance on the task descriptions, and M. K. Comer of GP for arranging for the participation of the GP instructors.

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

ABBREVIATIONS

BNL	Brookhaven National Laboratory
DOE	U.S. Department of Energy
HEP	Human Error Probability
HRA	Human Reliability Analysis
NRC	U.S. Nuclear Regulatory Commission
PRA	Probabilistic Risk Assessment
PSF	Performance Shaping Factor
RAM	Random Access Memory
SLI	Success Likelihood Index
SLIM-MAUD	Success Likelihood Index Methodology - Multi-Attribute Utility Decomposition
STAHR	Socio-Technical Approach to Human Reliability

EXECUTIVE SUMMARY

In order to determine the impact of human reliability on nuclear power plant safety, accurate estimates of human error probabilities (HEPs) are needed for probabilistic risk assessments (PRAs). Frequency data for use in estimating HEPs are generally unavailable, or if available, apply to a very limited range of fairly simple actions. To overcome this dilemma, the U.S. Nuclear Regulatory Commission (NRC) has embarked upon a program of research devoted to obtaining estimates of HEPs indirectly, that is, by using expert judgments to arrive at error estimates.

Brookhaven National Laboratory (BNL) developed and evaluated one method of obtaining human reliability estimates from expert judges--the Success Likelihood Index Methodology (SLIM). SLIM comprises a set of procedures based on Multi-Attribute Utility Theory for eliciting and organizing estimates by experts of the probability of success or failure of specific human actions in nuclear power plants.

The feasibility and implementability of SLIM were evaluated in a multiphase investigation. In the first phase, the basic characteristics of SLIM were defined (Embrey, 1983). Phases 2 and 3 consisted of an experimental evaluation and field test of SLIM. In Phase 4, SLIM was linked to an interactive computer program based upon Multi-Attribute Utility Decomposition (MAUD), and procedures for applying the resultant SLIM-MAUD methodology were developed (Embrey, Humphreys, Rosa, Kirwan, and Rea, 1984a; 1984b).

Phase 5, the final phase of the SLIM-MAUD research program, is reported in this two-volume document. Phase 5 was devoted to a systematic test application of the SLIM-MAUD methodology in order to evaluate its practicality, acceptability, and usefulness and to refine the procedures for implementing it.

The practicality of the SLIM-MAUD methodology was evaluated in terms of the costs of implementation, hardware and software requirements, personnel and time requirements, the expandability, transportability and ease of implementation of the methodology, and its ability to interface with the Human Reliability Data Bank. The acceptability of SLIM-MAUD to the scientific community, experts participating in the SLIM-MAUD test application, potential users, NRC and nuclear facilities was evaluated. The usefulness of the SLIM-MAUD methodology was evaluated in terms of its reliability, face validity and convergent validity.

The test application of SLIM-MAUD was divided into the following stages:

Stage 1 - Selection of Tasks for Assessment in the Test

Thirty tasks were selected for assessment using the SLIM-MAUD methodology. The tasks selected were identical to those employed in Comer, Seaver, Stillwell, and Gaddy's (1984) evaluation of psychological scaling as a method of estimating HEPs for nuclear power plant tasks. Fifteen tasks were designated as Level A and combined BWR plant systems with human actions which represented control room operator duties. Fifteen tasks were designated as Level B and combined equipment components with human actions which represented control room or equipment operator task elements.

Stage 2 - Selection of the Members of the Four Subject Mater Expert Groups for Stage 4

Four groups of subject matter experts were formed composed of individuals with human factors, PRA, or plant operations experience.

Stage 3 - Classification of Tasks into Subsets

A requirement of the SLIM-MAUD methodology is that tasks be sorted into subsets of 4 to 10 tasks which are reasonably homogeneous with respect to the performance shaping factors (PSFs) presumed to affect task outcome.

Therefore, each set of Level A and B tasks were sorted into subsets by a group of experts composed of individuals with PRA expertise, human factors expertise and nuclear power plant operations experience.

Stage 4 - Use of SLIM-MAUD by Each Subject Matter Expert Group for Each Subset of Tasks, Followed by Direct Numerical Assessment of all Tasks in all Subsets by Each Group Member

Each group of experts implemented the SLIM-MAUD methodology with the aid of a facilitator for each subset of tasks resulting in six SLIM-MAUD sessions per group. Then, in order to evaluate the relative usefulness of the SLIM-MAUD methodology as a technique for estimating HEPs, each expert participant used the psychological scaling techniques employed by Comer et al. (1984) to make direct estimates of HEPs for the 15 Level A and 15 Level B tasks. Finally, each expert participant completed an questionnaire to evaluate the SLIM-MAUD methodology in terms of ease of use, ability of the methodology to elicit and organize the judgments of a group of experts and the meaningfulness of the results produced.

Stage 5 - Analysis and Interpretation of Results from SLIM-MAUD Sessions With Respect to the Issues of Practicality, Acceptability and Usefulness

The issues of practicality of the SLIM-MAUD methodology were addressed in qualitative fashion. Formal and informal analyses were carried out to evaluate the acceptability and face validity of the methodology. Correlational and nonmetric multidimensional scaling analyses were conducted to assess the inter-judge reliability of SLIM-MAUD results, and its convergent validity with other subjective techniques for estimating HEPs. Additional analyses were also performed to investigate potential sources of bias in the SLIM-MAUD methodology.

The principal conclusions of this study were:

1. The practicality of SLIM-MAUD was demonstrated with respect to implementation costs, hardware, software, personnel, and time requirements and transportability of the methodology. The expandability of the SLIM-MAUD methodology was supported by the successful assessment of both complex (Level A) and simple (Level B) tasks. SLIM-MAUD can be implemented by a group of subject matter experts after receiving a minimal amount of training.

2. The estimates produced by SLIM-MAUD attained acceptable levels of reliability and showed greater inter-judge consistency than direct estimates of HEPs using psychological scaling techniques.

   To ensure the reliability of SLIM-MAUD results, the following recommendations are made:

   - Tasks to be assessed should be defined as concretely and completely as possible

   - For generic applications, expert groups should first identify a specific plant to typify the range of plants to which the results will be generalized;

   - The expert group should consist of four members. For the assessment of complex tasks, individuals with plant operations experience should form a majority in the group; for simple tasks, individuals with human factors expertise should form a majority in the group.

3. Considerable support was found for the face validity and convergent validity of SLIM-MAUD.

4. Expert participants in the SLIM-MAUD sessions found the SLIM-MAUD methodology easy to use and understand, useful in eliciting and organizing their judgments, and able to produce meaningful results.

5. The SLIM-MAUD methodology enables users to identify which PSFs have the most effect on the SLIs produced, thereby supporting the methodology's acceptability to safety study applications. The HEPs produced by SLIM-MAUD can be used in PRA and in the Human Reliability Data Bank.

Overall, SLIM-MAUD met or exceeded each of the criteria of practicality, acceptability and usefulness. Therefore, it is recommended as a methodology for producing HEPs needed for PRA and for entry into the Human Reliability Data Bank.

1.0 INTRODUCTION AND BACKGROUND

In recent years, human reliability in nuclear power plants has been the subject of widespread and growing concern. Considerable effort has been expended on obtaining accurate estimates of HEPs for PRAs. Frequency data for use in estimating human error rates (HERs) are generally unavailable, or if available, apply to a very limited range of fairly simple actions. To overcome this dilemma, the NRC embarked upon a program of research devoted to obtaining estimates of human errors indirectly, that is, by using expert judgments to arrive at error estimates. The goal of this research has been to produce HEP estimates in support of Human Reliability Analysis (HRA) segments of PRAs.

1.1 Purpose

The principal purpose of this study was to conduct a systematic test application of the Success Likelihood Index Methodology-Multi-Attribute Utility Decomposition (SLIM-MAUD) for estimating HEPs in nuclear power plants. The evaluation of the SLIM-MAUD test was based on three broad criteria: practicality, acceptability, and usefulness. These criteria, issues, and the methods and data used to address each, are presented in Table 1.1. Listed in this table are eight practicality issues, five acceptability issues, and three issues pertaining to usefulness. In addition, the experience gained in conducting the test--that is, the actual application of SLIM-MAUD in a realistic setting--served as the basis for developing a users' manual containing instructions and recommended procedures for implementing SLIM-MAUD.

1.2 Technology: Summary

The research conducted by BNL investigated one method of obtaining human reliability estimates from expert judges--the Success Likelihood Index Methodology (SLIM). SLIM comprises a set of procedures based on Multi-Attribute Utility theory for eliciting and organizing estimates by experts of the probability of success or failure of specific human actions in nuclear power

Table 1.1 SLIM-MAUD Test: Issues and Methods

Issues	Methods/Data
Practicality:	
P1 Cost	Compilation of actual costs incurred.
P2 Subject Matter Experts	Test sessions conducted with groups composed of PRA and human factors experts and individuals with operating experience.
P3 Support Requirements	Specification of equipment and human resources needed.
P4 Transportability	Implementation of test in two locations.
P5 Expandability	Task level compatibility.
P6 Time Requirements	Time expended for each task level.
P7 Interface With Human Reliability Data Bank	Ensured by tasks chosen for evaluation.
P8 Implementability of Procedure	Implementation by minimally trained facilitator.
Acceptability:	
A1 Scientific Community	Submission to professional journals.
A2 Expert Participants	Survey results.
A3 Potential Users	Informal comparative evaluation.
A4 Nuclear Regulatory Commission (NRC)	Not addressed directly (indirect evidence from survey results).
A5 Nuclear Facilities	Not addressed directly (indirect evidence from survey results).
Usefulness:	
U1 Reliability	Consistency of SLI estimates produced.
U2 Face Validity	Survey results.
U3 Convergent Validity	Comparisons with estimates produced by other techniques.

plants. The feasibility and implementability of SLIM were evaluated in a multiphase investigation. In the first phase, the basic characteristics of SLIM were defined (Embrey, 1983). Phases 2 and 3 consisted of an experimental evaluation and field test of SLIM. In Phase 4, SLIM was linked to an

interactive computer program based upon MAUD, and procedures for applying the resultant SLIM-MAUD methodology were developed. A detailed discussion of the investigation's second through fourth phases is reported in Embrey, Humphreys, Rosa, Kirwan, and Rea (1984a; 1984b).

The final phase of the SLIM-MAUD research program, reported in this document, was devoted to a systematic test application of the methodology and to refining the procedures for implementing it. The test application consisted of an evaluation of the practicality, acceptability, and usefulness of the SLIM-MAUD approach by the methods presented in Table 1.1.

The practicality of using the SLIM-MAUD methodology to obtain estimates of human errors was evaluated in terms of the cost of software and equipment, personnel requirements, transportability, expandability, time requirements, interface with the Human Reliability Data Bank (Comer, Kozinsky, Eckel, and Miller, 1983), and ease of implementability of procedure.

The acceptability of the SLIM-MAUD methodology to the individual subject matter experts who participated in the test application was evaluated via a survey administered after SLIM-MAUD sessions. Survey respondents (four groups, two composed of BNL staff members and two of NRC staff members) rated the methodology on the criteria of usefulness to PRA, ease of use, and meaningfulness of results. The acceptability of SLIM-MAUD in comparison to four other methods for estimating human error rates was evaluated by potential users during an informal evaluation conducted at a workshop sponsored by the Department of Energy (DOE) held at the University of Maryland in the fall of 1984.

The usefulness of the SLIM-MAUD methodology was evaluated in terms of the reliability or consistency of SLI estimates, the face validity of the methodology to the subject matter experts who used it, and the convergent validity of SLI estimates with independent HEPs for the same tasks generated by other methods of subjective expert judgment. Correlation coefficients were computed for the reliability and convergent validity analyses. Analysis of

the face validity of SLIM-MAUD was done qualitatively based on the responses of subject matter experts to a survey.

1.3 Organization of Report

This document is organized into two volumes consisting of four major parts: Volume I contains the first part and Volume II contains parts two through four--labeled Appendix A: SLIM-MAUD Users' Manual, Appendix B: Detailed Methods and Results of Test Application of SLIM-MAUD; and Appendix C: Human Reliability Estimates for the Tasks included in the SLIM-MAUD Test Application.

Volume I, the first part of this document, presents an overview of methodologies for systematizing the subjective judgments of experts, a general discussion of MAUD procedures for implementing SLIM, and an overview of the results of the SLIM-MAUD test evaluation.

Appendix A of Volume II presents detailed instructions for implementing SLIM-MAUD. The required resources and procedures to be followed in conducting each step in a SLIM-MAUD application are described.

Appendix B of Volume II contains a detailed description of the test evaluation of SLIM-MAUD. Definitions of the tasks assessed, a discussion of the procedures followed in the classification of tasks, and reliability and validity analyses are presented. In addition, results from a questionnaire designed to assess the acceptability of SLIM-MAUD and analyses comparing the SLIM-MAUD results with other methods of expert judgment are presented.

Appendix C of Volume II presents the human reliability estimates produced during SLIM-MAUD test application via MAUD and direct estimation procedures. Uncertainty bounds associated with the latter estimates are also presented.

2.0 METHODOLOGIES FOR SYSTEMATIZING THE SUBJECTIVE JUDGMENTS OF EXPERTS ON HUMAN PERFORMANCE AND RELIABILITY

In this section we briefly discuss some current methodologies designed to obtain expert judgments regarding the likelihood of human success or failure which may be transformed into probabilities appropriate for use in plant reliability analyses.

2.1 Wholistic Versus Decomposed Judgments

Two major categories of methods are: (1) those which rely upon the wholistic judgments of experts on task performance without decomposing these judgments into aspects, elements, or factors of the task, and (2) those which systematize the decomposed judgments of experts' on a comprehensive set of the aspects, elements, or factors which together define the task. In each case, the judgments made by experts are subjective and hence a psychological scaling technique must be used to present them in appropriate numerical order. In the case of decomposed judgments, the technique must also provide an appropriate composition rule, which defines how the set of numbers obtained from the decomposed judgments of task aspects, elements, or factors are to be combined to produce the single-number indexing of the experts' overall judgment of the probability of success or failure of operator performance on the task being assessed.

All methodologies employing wholistic judgments assume that judges can immediately and directly bring their experience to bear on rating the operator's likelihood of success (or failure) in carrying out the task in question. Four methodologies for obtaining wholistic subjective judgments were reviewed in Seaver and Stillwell (1983): (1) the paired comparison technique, (2) rating/ranking techniques, (3) the direct numerical estimation technique, and (4) the indirect numerical estimation technique. Comer et al. (1984) used the paired comparison technique and a direct numerical estimation technique to derive HEPs for the same tasks used for the test of SLIM-MAUD described here. Comer et al. found that both these psychological scaling techniques produced

reasonably consistent estimates of HEPs; essentially no differences were found in the measures of consistency and convergent validity obtained from the two techniques. However, the paired comparison technique is limited in that the paired comparison scale must be calibrated into a probability scale, it is more time consuming and requires many more judgments than does the direct numerical estimation technique, and it is appropriate only when operator actions on a number of comparable tasks are to be judged as a set. For these reasons, this test application of SLIM-MAUD utilized the direct numerical estimation technique to provide a baseline against which SLIM-MAUD's performance could be assessed.

Methodologies which use decomposed judgment techniques assume that it is much easier to assess elements of the task, aspects of the task situation, or factors affecting task performance than the overall probability of task success or failure. It is further assumed that a methodology employing a formal mathematical "composition rule" which takes these judgments as its input, and provides an overall (i.e., wholistic) judgment as output will produce results superior to those of judges who intuitively try to combine their decomposed judgments into an overall judgment.

Two decomposition methodologies have been proposed for use in PRA context: SLIM-MAUD and Socio-Technical Analysis of Human Reliability (STAHR). SLIM-MAUD, the subject of this report, is reviewed in Section 2.2.1. STAHR requires that a group of experts construct an influence diagram showing the influences of possible conditioning events and factors on the operator's likelihood of error in carrying out a designated task. Details of the STAHR approach are provided in Phillips, Humphreys, and Embrey (1985) and Phillips and Embrey (1985).

2.2 Assumptions and Capabilities of SLIM-MAUD

SLIM-MAUD is the name given to a methodology whereby an interactive computer-based procedure, MAUD, is used to elicit and organize the assessments of

experts within the framework of SLIM, which was first described by Embrey and Hall (1981).

The approach is based on the assumption that the likelihood of successfully accomplishing an action or task is a function of various characteristics of the individual, the situation, and the task itself. These factors are known as Performance Shaping Factors (PSFs) and are presumed to combine together to determine the probability of success. The success likelihood is based upon a rating of how good or bad these factors are in a particular situation, weighted by their relative importance in affecting success.

When MAUD is used to implement SLIM, 4 to 10 tasks must be assessed simultaneously on the same set of PSFs. The judges use MAUD to identify which PSFs are relevant in assessing the set of tasks and assign relative importance weights to these PSFs. This procedure is based upon the assumption that the same set of PSFs is appropriate for assessing all the tasks under consideration in any single session with MAUD. When a large number of tasks is to be assessed, the full set of tasks under consideration must first be sorted into subsets comprising 4 to 10 tasks. Each subset must be reasonably homogeneous with regard to the PSFs which are important in discriminating between correct and erroneous performance. In most cases the experts are able to sort the tasks directly, using a simple card-sorting procedure which is described in Appendix A of Volume II.

Explicit consideration of PSFs is the basic underpinning of SLIM. Each PSF may denote some particular human trait or a condition of the work setting that is perceived by subject matter experts to have a major influence on the success likelihood in the scenario being evaluated. The precise set of PSFs identified will vary from session to session; however, the following PSFs are often identified by experts when assessing tasks: (1) training, (2) time available, (3) task-relevant information available, (4) procedures, (5) complexity of the task, (6) level of stress, (7) personnel competence/skill, (8) equipment design characteristics, (9) characteristics of the work environment.

A full description of the process by which MAUD works in interaction with a group of experts in a SLIM-MAUD session is given in Sections 1.11 through 1.13 in Embrey et al. (1984b), together with an account of its theoretical foundations and assumptions which lie within multi-attribute utility theory.

3.0 IMPLEMENTATION OF SLIM-MAUD

The steps required to obtain HEP assessments using SLIM-MAUD are described in this section. Detailed instructions for the implementation of each step are given in Volume II, Appendix A. Table 3.1 shows these steps in sequence.

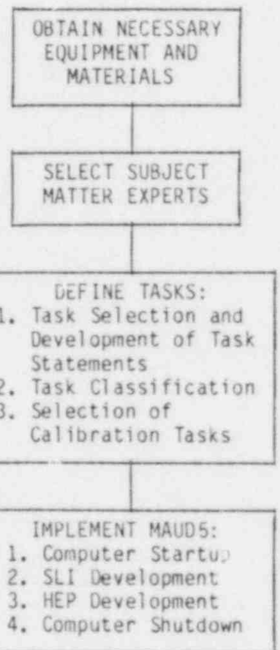

Figure 3.1 Steps for the implementation of SLIM-MAUD.

3.1 Obtain Necessary Equipment and Materials

3.1.1 Equipment

The equipment needed to implement SLIM-MAUD includes a personal computer with a minimum of 64K Random Access Memory (RAM) which runs under CP/M or IBM/PC DOS operating systems, two 360K disk drives, a monitor, and a printer.

3.1.2 Materials

1. The SLIM-MAUD software package is contained on one disk. An End User's License Agreement must be purchased from the Decision

Analysis Unit of the London School of Economics in order to implement SLIM-MAUD.

2. Two or more blank 5-1/4 inch diskettes (one to make a working copy of SLIM-MAUD programs and one to store data from the SLIM-MAUD session).
3. Index cards (any size).
4. Documentation booklet.

3.2 Select Subject Matter Experts

A key emphasis in SLIM-MAUD applications, and one which distinguishes it from other subjective techniques, is the requirement that SLIM-MAUD sessions be conducted within a group context. This group requirement is designed to achieve two objectives: (1) to arrive at a shared definition and understanding of the tasks to be assessed and (2) to facilitate the identification of the full range of PSFs presumed to affect task outcome. This latter objective is best achieved if the group is composed of four individuals with a broad range of expertise and experience including actual nuclear power plant operation, human factors, and PRA. Such groups are far more likely than narrowly specialized groups to recognize the importance of a wide range of PSFs. Ensuring that the PSFs selected are representative of the important influences on task outcome is especially crucial to optimum applications of SLIM-MAUD because PSF identification is the basic underpinning of the entire SLIM-MAUD methodology.

3.3 Task Development

Task development for a SLIM-MAUD application involves three steps: (1) task selection and the development of task statements, (2) task classification, and (3) the selection of calibration tasks.

3.3.1 Task Selection and the Development of Task Statements

The subject matter experts who implement SLIM-MAUD must select the specific nuclear power plant tasks for which HEPs are needed and develop task statements.

Comer et al. (1984) pointed out that "Probably the most critical requirement for the use of judgmental procedures to estimate HEPs is that the tasks to be judged be defined carefully and completely. The more fully the tasks are specified, the less likely they will be open to variable interpretation by the experts judging their likelihood." Because clarity of task definition is so critical, SLIM-MAUD works best in concrete applications such as in a PRA for a specific nuclear power plant. For more generic applications it is useful for user groups to first decide upon a specific plant that best typifies the range of plants being considered in the application. Such a decision assures the SLIM-MAUD participants that they will be using a shared concrete image to elicit their inputs.

3.3.2 Task Classification

Task statements must be classified into subsets of 2 to 8 tasks based on similarity of the factors which influence the probability of human error on each task. That is, each subset of tasks must be homogeneous with respect to the group of PSFs presumed to affect task outcome. This classification can be accomplished by asking the group of subject matter experts to sort the tasks into subsets based on their judgments of the interrelatedness of the tasks with respect to the PSFs influencing task outcome.

3.3.3 Selection of Calibration Reference Tasks

Two additional nuclear power plant tasks for which HEPs are known or have previously been estimated must be included in each subset of tasks to

serve as "calibration reference tasks." Thus, the minimum number of tasks which can be assessed in a SLIM-MAUD session is 4; the maximum number of tasks is 10. For each pair of calibration reference tasks, one task should have a relatively high probability of failure and one task should have a relatively low probability of failure. Also, each pair of calibration reference tasks should be similar to the other tasks in the subset in terms of the PSFs presumed to affect task outcome. SLIs will be generated for all tasks in the subset, including the calibration reference tasks. Later, the SLIs and HEPs for the calibration reference tasks will be used to calculate HEPs for the other tasks in the subset.

3.4 Implement SLIM-MAUD Programs

There are four steps in the implementation of the SLIM-MAUD programs: (1) computer startup, (2) SLI development, (3) HEP development, and (4) computer shutdown.

The SLIM-MAUD software diskette contains several programs. The MAUD5 program is run to generate SLIs. The SLIMHEP program is run to convert SLIs to HEPs. The M5CONFIG program can be run to customize the text in the MAUD5 program. Instructions for starting the computer, making backup copies of the programs, running the programs, and shutting down the computer are in Volume II, Appendix A.

The MAUD5 program is interactive and directs the group of subject matter experts to type in the names of the tasks which make up each subset, identify the PSFs which influence the performance of the set of tasks, rate the tasks on each PSF, and assign relative importance weights to each PSF. Instructions appear on the computer monitor. The subject matter experts using MAUD5 type their inputs to the program using the keyboard. Opportunities are provided for the subject matter experts to change or edit their inputs. The results of the SLIM-MAUD session can be stored as a file on the data diskette, and can be printed out in a summary report. The data diskette file can be used in a subsequent SLIM-MAUD session to reassess the tasks. The summary report includes

the names of the tasks in the subset, the PSFs identified, the experts' ratings of the PSF for each task, the relative weights of the PSFs used by the program, and the SLIs for the tasks in the subset. Each SLI represents the likelihood of success for a particular task relative to the other tasks in the subset.

The MAUD5 program must be run separately for each subset of tasks. When all the subsets of tasks have been assessed and relative SLIs for each task have been printed out, the SLIMHEP program is run to convert the SLIs into HEPs.

Uncertainty bounds for the HEPs can be derived by identifying the highest and lowest HEPs within a subset of tasks. Since subsets of tasks share similar characteristics but differ in quality or degree of PSFs, the range of HEPs for tasks within a subset define the upper and lower uncertainty bounds for a given HEP estimate. That is, for tasks of that nature, it is reasonable to expect that the HEP will fall within the range of variability defined by the set of HEPs. More information about selecting uncertainty bounds for tasks assessed using SLIM-MAUD is contained in Appendix A of Volume II.

The HEPs derived from a SLIM-MAUD session can be used to support the HRA segments of PRAs. The estimates can also be entered into the Human Reliability Data Bank (Comer et al., 1983) for reference by the human reliability and PRA community. SLIs, representing the relative position of tasks on the Success Likelihood scale, can be used prescriptively in safety studies.

4.0 TEST APPLICATION OF SLIM-MAUD

A plan for the test application of SLIM-MAUD was outlined by Embrey et al. (1984a, Section 8; 1984b, Section 4). This plan was implemented with certain revisions, and carried out under NRC contract via DOE to the Department of Nuclear Energy, Brookhaven National Laboratory, between June and December, 1984.

The issues addressed by the study, the evaluation methods used, and the results obtained are discussed in this section.

4.1 Issues Addressed

The test plan was designed to assess the utility of the MAUD-based implementation of SLIM 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 rigorously addressed within the Test Plan. These specific issues and methods for addressing them were summarized in Table 1.1 in Section 1.1 of this volume. The methods are described in Section 4.2 below. The findings with regard to the application of SLIM-MAUD on these issues are discussed in detail in Section 4.2.5.

4.2 Method of Evaluation

The actual test application of SLIM-MAUD outlined in Embrey et al. (1984a; 1984b) underwent several revisions. The principal revisions were:

1. Subject matter experts participating in each group evaluated the set of tasks under consideration using both SLIM-MAUD and the direct numerical estimation psychological scaling procedure described by Comer et al. (1984).

2. Four groups of subject matter experts were used, instead of the five in the original test plan.

3. The scope of the test plan analysis was considerably expanded in terms of the reliability and validity analyses performed.

The test was divided into the following stages:

Stage 1: Selection of tasks for assessment in the test.

Stage 2: Selection of the members of the four subject matter expert groups.

Stage 3: Classification of tasks into subsets for simultaneous assessment within SLIM-MAUD.

Stage 4: Use of SLIM-MAUD by each subject matter expert group for each subset of tasks, followed by direct numerical assessment of all tasks in all subsets by each group member.

Stage 5: Analysis and interpretation of results from SLIM-MAUD sessions with respect to the criteria and issues outlined in Table 1.1.

The procedures followed in each of these stages are described below.

4.2.1 Stage 1: Selection of Tasks for Assessment in the Test

For comparison purposes, the definitions of the tasks assessed in the SLIM-MAUD test were identical to those employed in the Comer et al. (1984)

test of psychological scaling methods employing wholistic judgment. Comer et al. describe the general characteristics of these task as follows:

> ".. The tasks correspond to Level 1 and Levels 2 and 3 as defined by the Human Reliability Data Bank (Comer et al., 1983). Level 1 of the Human Reliability Data Bank structure combined power plant systems with human actions that represented job duties. In this project the Level 1 tasks represented BWR systems and control room operator duties. Level 2 of the data bank structure combined equipment components with human actions defined as tasks. The tasks defined for this project included those associated with control room operators and equipment operators. Level 3 correspond to controls and displays and task elements..."

Comer et al. developed written descriptions of 15 Level A and 20 Level B tasks. Level A tasks correspond to Level 1 tasks within the Human Reliability Data Bank (Comer et al., 1983); Level B tasks correspond to Levels 2 and 3 tasks in the Human Reliability Data Bank. All 15 Level A tasks were assessed in the SLIM-MAUD test. However, to assure compatibility of task set size, the level B task set was reduced from 20 to 15 tasks for the SLIM-MAUD test. Criteria used to determine which 5 of the Level B tasks to exclude from the SLIM-MAUD test are described in Volume II, Appendix B.

4.2.2 Stage 2: Selection of the Members of the Four Subject Matter Expert Groups for Stage 4

Each group of experts selected to participate in the SLIM-MAUD test was composed of participants with expertise in the areas of human factors, PRA, and nuclear power plant operations.

The precise composition of each group and the venue at which it met were as follows:

> Group 1. One human factors specialist, two PRA experts, one expert with operations experience; Brookhaven National Laboratory.

Group 2. One human factors specialist, two PRA experts, one expert with operations experience; Brookhaven National Laboratory.

Group 3. One human factors specialist, one PRA experts, and two experts with operations and operator training experience; NRC in Maryland.

Group 4. Two human factors specialists, one PRA expert, and one expert with operations and operator training experience; meeting at the NRC in Maryland.

4.2.3 Stage 3: Classification of Tasks into Subtasks

The set of 15 Level A tasks and the set of 15 Level B tasks identified in Stage 1 each covers a wide range of tasks and thus it could not be assumed that the same set of PSFs would apply equally in assessing all tasks in each set. Hence, it was necessary to sort each set of tasks into subsets, each containing 4 to 10 tasks, which are reasonably homogeneous in each subset with regard to those PSFs which were likely to be considered important by the experts chosen to assess them in the SLIM-MAUD sessions in Stage 4 of the test plan.

Two groups, each comprised of four subject matter experts (with experience similar to those participating in the four groups described in Section 4.2.2) were presented with the descriptions of the 15 Level A and 15 Level B tasks.

Each group of experts was asked to make wholistic ratings of the interrelatedness of tasks using a paired comparison procedure. These ratings were based upon judgments of the relative importance of PSFs in determining the likelihood of success for each pair of tasks. Group consensus procedures were used to obtain the wholistic ratings (see Nemiroff and King, 1975; Gustafson et al., 1983).

Written consensus ratings were collected from each group for formal analysis. A clustering technique was applied to these ratings based on a comprehensive analysis of the pattern of relations between the tasks. This was achieved through the use of nonmetric multidimensional scaling implemented through the computer program KYST (Kruskal, Young, and Seery, 1973; Kruskal and Wish, 1978) which, although in the public domain, is available only on certain large mainframe computer installations. This analysis indicated that at each level three reasonably homogeneous subsets of tasks could be clearly identified.

The tasks comprising each subset were identified as follows:

Level A Subset 1: Tasks 2, 3, 4, 5, 7, 14 (six tasks)
Subset 2: Tasks 6, 8, 12, 13, 15 (five tasks)
Subset 3: Tasks 1, 9, 10, 11 (four tasks)

Level B Subset 1: Tasks 3, 7, 8, 12, 19 (five tasks)
Subset 2: Tasks 1, 2, 9, 10, 17, 18 (six tasks)
Subset 3: Tasks 5, 6, 14, 15 (four tasks)

The task numbers correspond to those given to the task descriptions in Comer et al. (1984, Appendix 3). The complete set of Level A and Level B task descriptions, arranged in these subsets, is given in Appendix B of Volume II.

4.2.4 Stage 4: Use of SLIM-MAUD by Each Subject Matter Expert Group for Each Subset of Tasks, Followed by Direct Numerical Assessment for All Tasks by Each Group Member

Each of four groups of subject matter experts met for one day at the venue described in Stage 3. Each group used SLIM-MAUD to assess the three Level A subsets of tasks, followed by the three Level B subsets (six MAUD sessions in all).

On arrival, each group member was handed a booklet containing the task descriptions and asked to study the task descriptions. After all members became familiar with the tasks, each group worked in direct interaction with SLIM-MAUD, with one member of the group (nominated by the group as a whole) being responsible for typing in the group's responses to the interaction with MAUD. Also present were a facilitator and a technical recorder.

The facilitator gave the group a brief nontechnical introduction to the SLIM-MAUD session (the text of this introduction is given in Volume II, Appendix B). Questions from group members were answered, and the group was then asked to commence the first SLIM-MAUD session. From here on, the facilitator's intervention in the group process was minimal; all the steps in the assessment procedure were controlled by MAUD in direct interaction with the group. The facilitator's few interventions were almost exclusively concerned with ensuring that the views of all the group members were fully considered in forming each judgment input to MAUD. In this way, consensus (or very occasionally an agreed compromise) was reached by the group on all aspects of the (decomposed) assessments of the tasks. In no case did a group member withdraw from the judgment process or indicate that his or her views were not represented in the interactions with MAUD.

The technical recorder's role was purely passive: keeping track of the group discussion during the sessions and noting the major points raised by group members in this discussion. Most of this discussion involved clarification of task statements or PSFs. Appendix B provides typical examples of these discussions as noted by the technical recorder.

Immediately upon completion of the SLIM-MAUD sessions, the four members in each group were assigned to separate rooms and asked to complete a direct estimate response booklet following the format described in Comer et al. (1984, Volume II, Section 3.2.1). The tasks assessed within this booklet were the same 15 Level A and Level B tasks in the SLIM-MAUD sessions but were arranged in numerical order rather than by subset (details are given in Volume II, Appendix B).

After completion of the six sessions with MAUD and the rating of the tasks using direct estimation, each participant in each group completed a questionnaire designed to assess acceptability of SLIM-MAUD to himself or herself, and to other potential users. This questionnaire is described in detail in Appendix B of Volume II.

The time schedule for the activities described above is shown in Table 4.1. Times given for each MAUD activity are the average of the times taken by the four groups. No group deviated more than 25% from the average time shown for each activity.

Table 4.1 Average Time for Completion of Each Activity in SLIM-MAUD Test Application

Activity	Time Taken (Minutes)
1. Individual participants familiarize themselves with Levels A and B tasks	30
2. Introduction to SLIM-MAUD (given by group facilitator)	30
3. Assessment of Level A subset 1 tasks (break for refreshments)	110
4. Assessment of Level A subset 2 tasks	85
5. Assessment of Level A subset 3 tasks	50
6. Assessment of Level B subset 1 tasks	20
7. Assessment of Level B subset 2 tasks	15
8. Assessment of Level B subset 3 tasks (break for refreshments)	15
9. Direct rating of Levels A and B tasks by experts (each working alone)	45
10. Completion of SLIM-MAUD acceptability questionnaire by experts (each working alone)	15

4.2.5 Stage 5: Analysis and Interpretation of Results From SLIM-MAUD Sessions With Respect to the Criteria and Issues Outlined in Table 1.1

The test application of SLIM-MAUD reported here provided the opportunity to evaluate the practicality, acceptability, and usefulness of the methodology. The results of the test application with respect to the criteria of practicality, acceptability, and usefulness (and the issues encompassed by each) are reported in this section. Details of this analysis and interpretation of results from Stage 5 of the test application are given in Appendix B of Volume II.

4.2.5.1 Practicality

For this evaluation, practicality was defined by cost, subject matter experts, support requirements, transportability, expandability, time requirements, ability to interface with the Human Reliability Data Bank, and implementability of the SLIM-MAUD procedures (issues P1-P8 in Table 1.1). This test evaluation demonstrated that SLIM-MAUD fulfills the criterion of practicality--it is relatively inexpensive in terms of cost, equipment, personnel, and time requirements, requires a minimum of training, is easy to implement in different locations, is applicable to a wide range of tasks, and is compatible with the Human Reliability Data Bank (Comer et al., 1983). Each of the issues that make up the criterion of practicality will be briefly discussed.

P1 - Cost. Two essential components--software and equipment--define the basic cost of implementing SLIM-MAUD. The proprietary SLIM-MAUD software programs can be obtained by purchasing a MAUD5 End User's License Agreement for approximately 200 pounds sterling. (See Appendix A of Volume II for procedures for obtaining a MAUD5 End User's License Agreement.) The program code for the nonproprietary SLIM-MAUD software (SLIMHEP) for converting SLIs to HEPs is contained in Appendix A of Volume II. The minimum equipment required include a personal computer with a minimum of 64K random access memory (RAM) which runs under CP/M or IBM/PC DOS operating systems, two floppy disk drives, monitor, and

printer. Other materials needed to implement SLIM-MAUD include floppy diskettes, index cards, and a SLIM-MAUD documentation booklet.

P2 - Subject Matter Experts. Four SLIM-MAUD users, the recommended number for implementing SLIM-MAUD, took part in each of the four test sessions constituting this evaluation. Each group should include at least one individual experienced in operating nuclear power plants of the specific type being assessed and one individual with human factors experience, as well as individuals with PRA experience. Although not recommended, SLIM-MAUD can be implemented with a two-member group if the areas of expertise represented include human factors experience and operating experience.

A multidimensional scaling analysis, reported in Volume II, Appendix B, showed that the composition of the group and the complexity of the tasks interacted to affect the inter-expert reliability coefficients for direct estimates of HEPs. The results suggested that for applications of SLIM-MAUD to complex tasks, such as Level A, group members with plant-operating experience should be well represented. For applications to simple tasks, such as Level B, human factors experts should be well represented within groups.

P3 - Support Requirements. When the group of subject matter experts is inexperienced with SLIM-MAUD, it is useful to have a facilitator present to provide an introduction to the SLIM-MAUD procedures and to guide group inputs. When the group does include an individual familiar with SLIM-MAUD, that person can assume the facilitator's role.

P4 - Transportability. SLIM-MAUD can be implemented in a wide variety of settings, provided the requirements enumerated above (P1-P3) are available at each location. The test application of SLIM-MAUD was implemented in two different locations using separate personal computers. The fact that SLIM-MAUD may be implemented on a variety of personal computers with compatible operating systems, and the popularity and

availability of personal computers in general, adds to the portability of the procedure. Personal computers can easily be transported from one location to another or rented for short periods of time.

P5 - Expandability. The assumptions underlying SLIM-MAUD and the SLIM-MAUD procedures themselves are sufficiently robust to be capable of assessing virtually any human task in nuclear power plants, although it is particularly useful for assessing complex tasks like the Level A tasks used in this test application (see Volume II, Appendix B).

P6 - Time Requirements. Three factors determine the time required to conduct a SLIM-MAUD session: user experience, number of tasks, and task complexity. A group of experienced SLIM-MAUD users can assess approximately 25 complex tasks (such as the Level A tasks) and as many as 60 simple tasks (such as the Level B tasks) in one working day. In this test application of SLIM-MAUD, the average total time taken by inexperienced groups to assess 15 complex tasks and 15 simple tasks was less than six hours.

P7 - Interface With Human Peliability Data Bank. The two levels of tasks, A and B, used in this test application correspond to Level 1 and Levels 2 and 3 as defined in the Human Reliability Data Bank (Comer et al., 1983), thereby ensuring compatible interface of the SLIM-MAUD methodology with industry-specific data.

P8 - Implementability of Procedure. SLIM-MAUD was successfully implemented by a facilitator who had not taken part in the development of the methodology, and who had minimal training in its application. Because its software is interactive, all interactions with, and data input to, SLIM-MAUD can be accomplished by user groups without previous training.

4.2.5.2 Acceptability

Acceptability of the SLIM-MAUD methodology was defined by its acceptance by the scientific community, expert participants, potential users, NRC, and

nuclear utilities (issues A1-A5 in Table 1.1). Each of these issues are discussed briefly below, and details of these evaluations are given in Volume II, Appendix B.

A1 - Scientific Community. The SLIM-MAUD test application was carried out with sufficient rigor to produce results that meet the standards of publication in reputable scientific journals. Journal publication of the test application findings will appear soon after the publication of this report.

A2 - Expert Participants. Results of the survey administered to participants in the SLIM-MAUD sessions indicate that they found the methodology relatively easy to use, thought the results to be meaningful, and generally believed SLIM-MAUD to be useful to PRAs. Furthermore, participants indicated it would be easy to identify which PSFs had the greatest effect on the SLIs produced, thus supporting SLIM-MAUD's diagnostic and prescriptive capabilities for safety applications.

A3 - Potential Users. A formal evaluation of the acceptability of SLIM-MAUD to potential users was not carried out as part of this test application. However, an informal evaluation of five methods of estimating HEPs, including SLIM-MAUD, took place at a DOE-sponsored workshop held at the University of Maryland in the autumn of 1984. The results suggested that potential users will find SLIM-MAUD to be moderately-to-highly acceptable as a technique for estimating HEPs. Because the rating procedure had been relatively informal, and hence unvalidated, the participants agreed that the numerical results would not be made public. Therefore, only the general nature of the results with respect to SLIM-MAUD will be presented here.

Each technique was rated on the basis of 12 criteria: (1) traceability of procedure, (2) reproducibility of results, (3) flexibility of technique, (4) freedom from judgmental biases, (5) training required, (6) resources required, (7) specificity, (8) completeness of modeling, (9)

sensitivity analysis capability, (10) data content, (11) completeness of procedures, and (12) degree of insight provided by application of technique. The ratings of SLIM-MAUD were generally favorable: high on traceability, flexibility, freedom from judgmental bias, specificity, completeness of modeling, sensitivity analysis capability, completeness of procedures; moderately high on reproducibility of results, training required, resources required, and insight provided by techniques: and low on the extent to which the technique contained a built-in data base.

A4 - Nuclear Regulatory Commission. Data directly addressing the issue of acceptability to NRC (i.e., the extent to which the NRC recommends the use of SLIM-MAUD in future PRA applications) were not collected as part of this evaluation. Indirect evidence relevant to SLIM-MAUD's acceptability, however, was gathered from the NRC subject matter experts who participated in the test application of SLIM-MAUD. The responses suggest that SLIM-MAUD will likely be acceptable for government PRA work.

A5 - Nuclear Utilities. Data directly addressing the issue of acceptability to nuclear utilities was not collected as part of this evaluation. Such acceptability can be determined by the extent to which utilities adopt SLIM-MAUD for future PRA applications. Indirect evidence (i.e., perceptions of session participants) suggests that SLIM-MAUD is moderately likely to be acceptable to utilities.

4.2.5.3 Usefulness

The three major issues associated with the usefulness of the SLIM-MAUD methodology are its reliability, face validity, and convergent validity (issues U1-U3 in Table 1.1). These are discussed below:

U1 - Reliability. The reliability of the SLIM-MAUD methodology was examined separately for Level A and Level B tasks. In a SLIM-MAUD session, the group as a whole produces a single set of SLIs for the subset

of tasks being assessed. Therefore, the reliability of SLIM-MAUD was examined by comparing the consistency of the SLI values generated across the four groups in the test plan. The reliability coefficients for overall intergroup consistency were +0.62 ($p < 0.01$) for Level A tasks (range = +0.47 - +0.63), and +0.65 ($p < 0.01$) for Level B tasks (range = +0.52 - +0.77). These results indicate a moderate degree of agreement between the SLI values generated by any two groups of experts.

The stability of the assessments made using SLIM-MAUD, one aspect of reliability, was compared to the stability of two psychological scaling techniques for obtaining direct estimates of HEPs. Table 4.2 summarizes the results of these comparisons. These reliability analyses indicated that SLIM-MAUD is a considerably more stable procedure than psychological scaling techniques used to form direct estimates of HEPs. This superior stability was particularly marked for Level A tasks. Appendix B of Volume II includes details of the reliability analyses of these techniques and a discussion of potential sources of bias which might lead to unreliability of the SLIM-MAUD procedure. The instructions provided in Appendix A of Volume II are designed to ensure maximum reliability of SLIM-MAUD results in practical applications. These describe the optimal procedures for defining and classifying tasks and for selecting subject matter experts.

Table 4.2 SLIM-MAUD Reliability Compared With Two Psychological Scaling Reliability Baselines*

Technique	Task Level A	Task Level B
SLIM-MAUD	0.62	0.65
Direct HEP estimates using psychological scaling, SLIM-MAUD subject matter experts	0.36	0.55
Direct HEP estimates using psychological scaling, Comer et al. (1984) subject matter experts	0.43	0.63

*$p<0.01$ for all correlation coefficients.

U2 - Face Validity. Face validity refers to whether the procedures appear relevant, appropriate, and valid to users of the methodology. It is considered an essential precursor to more rigorous types of validity, and is a desirable feature of any methodology. Although frequently assessed informally, face validity nevertheless provides some assurance that a methodology measures what it is supposed to be measuring.

The face validity of SLIM-MAUD was assessed with three items on the questionnaire administered to the session participants, where participants were asked to respond on a five-point Likert scale (ranging from strongly agree to strongly disagree). Results obtained for the three items are presented in Table 4.3.

Table 4.3 Results of Questionnaire Items Assessing Face Validity of SLIM-MAUD

	Percent of Respondents Who:				
	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagreee
The SLIM-MAUD procedures led to results that seemed meaningful	19 (N=3)	63 (N=10)	12 (N=2)	6 (N=1)	0 (N=0)
It would be easy to determine which PSFs had the greatest impact on HEP estimates by reviewing the SLIM-MAUD log of results	12 (N=2)	63 (N=10)	19 (N=3)	6 (N=1)	0 (N=0)
SLIM-MAUD can be useful to HRA segments of PRAs	12 (N=2)	75 (N=12)	6 (N=1)	6 (N=1)	0 (N=0)

A strong majority of participants expressed approval of the results produced by SLIM-MAUD: 82% said that the results seemed meaningful; 75% thought that the PSFs having the greatest impact on HEP estimates were easily traceable; and 87% thought SLIM-MAUD were useful to HRA segments of PRA. Only one respondent disagreed with each of the three statements

and three additional respondents expressed a neutral position to one or two face validity items. Thus, the evidence in general shows SLIM-MAUD to have an acceptable level of face validity.

U3 - Convergent Validity. It is difficult to validate the utility of subjective judgment techniques for estimating HEPs in nuclear power plant operations because of the low frequency of recorded operator failures. Thus, there are few objective criteria against which to measure the subjective probability estimates. Therefore, the criterion-related validity of the SLIM-MAUD methodology was assessed by "convergence" with other methods for estimating HEPs in nuclear power plants.

Convergent validity refers to the correlation between methodologies which are designed to measure the same construct. Thus, if the SLIM-MAUD methodology produces valid assessments, we would expect the SLI results to correlate highly with the results of other reliable subjective judgment techniques applied to the same tasks.

The SLI assessments were correlated with direct HEP estimates for the Levels A and B tasks produced by psychological scaling techniques from two groups of experts: (1) the 16 experts who participated in the SLIM-MAUD test application and (2) the 19 experts who participated in the evaluation by Comer et al. (1984) of psychological scaling techniques. In addition, the SLI assessments for the 15 Level B tasks were correlated with HEP estimates for these tasks given in the Handbook of Human Reliability Analysis (Swain and Guttmann, 1983). Table 4.4 gives the overall correlations between the SLIM-MAUD assessments and those of the other methods.

The degree of convergence between the results of two methodologies will be influenced by the reliability of each of the methodologies. The intergroup reliability coefficients for the results of SLIM-MAUD and direct psychological scaling of HEPs indicated that the methodologies possess different levels of reliability (see Table 4.2), ranging from

Table 4.4 Correlations Between SLI Assessments and Other Techniques for Human Reliability Analysis*

Source of Data for Comparison	Task Level A	Task Level B
Direct HEP estimates using psychological scaling, subject matter experts	0.48	0.66
Direct HEP estimates using psychological scaling, Comer et al. (1984) subject matter experts	0.52	0.69
HEP estimates from Handbook of Human Reliability Analysis	--	0.54

*$p<0.01$ for all correlation coefficients.

low to moderate. This measurement error will tend to decrease the size of the correlation coefficient between SLIM-MAUD and any other methodology. However, to the extent that SLIM-MAUD is capable of producing reasonably stable and appropriate estimates of HEPs, it should demonstrate acceptable levels of convergent validity with a range of similar methods.

Table 4.4 presents the correlation coefficients between SLIs and the results of other techniques for human reliability analysis. In general, the pattern of correlations in Table 4.4 is satisfactory. The fact that correlations at Level A are lower than at Level B may reflect the fact that psychological scaling techniques are less reliable for assessing Level A tasks than Level B tasks.

Multidimensional scaling analyses of assessments made by different groups of experts provided an additional indication of the convergence of SLIM-MAUD results with the results of other subjective judgment techniques (see Appendix B of Volume II for details). For Level A tasks, the SLIM-MAUD assessments were more consistent and occupied a separate

domain in the "reliability space" mapped by the multidimensional scaling analysis smaller than that occupied by the psychological scaling assessments. For Level B tasks, the domains mapped by each technique were similar in size with substantial overlap. This indicated greater convergent validity between SLIM-MAUD assessments and direct psychological scaling assessments for Level B than for Level A tasks. In addition, the direct scaling technique appeared to yield more "off-center" (i.e., method-idiosyncratic) results.

5.0 CONCLUSIONS AND RECOMMENDATIONS

This project, a rigorous test application of the SLIM-MAUD methodology, provided results from which several key conclusions and recommendations can be drawn. The principal conclusions are:

- The practicality of SLIM-MAUD was demonstrated with respect to implementation costs, required subject matter experts, time requirements, and transportability.

- The assumptions underlying SLIM-MAUD, presumed to ensure the methodology's expandability, were supported by the assessment of both complex (Level A) and simple (Level B) tasks.

- SLIM-MAUD can be implemented by a group of subject matter experts after receiving a minimal amount of training.

- The estimates produced by SLIM-MAUD attained acceptable levels of reliability and were shown to be more stable than direct numerical estimates.

- Negligible differences in the reliability of Levels A and B tasks were observed.

- Considerable support was found for the face validity and convergent validity of SLIM-MAUD.

- The convergent validity between the SLIM-MAUD estimates and the estimates produced by other procedures was greater for the Level B than for the Level A tasks; however, this difference is substantively negligible.

- Results of a questionnaire administered to experts participating in the SLIM-MAUD sessions indicate their general confidence in the

methodology; they found the SLIM-MAUD methodology easy to understand and use, useful in tapping and organizing their expertise, and able to produce results that appear meaningful.

- The SLIM-MAUD methodology enables users to identify which PSFs have the most effect on the SLIs produced, thereby supporting the methodology's applicability to safety study applications.

- The HEPs produced with SLIM-MAUD can be used in PRA and in the Human Reliability Data Bank.

Overall, SLIM-MAUD has met or exceeded the criteria of practicality, acceptability, and usefulness established for the test application carried out in this project. The methodology can be implemented in a cost-effective manner, can be readily understood and applied by experts with no previous experience, and produces results that appear reasonable and valid. Thus, SLIM-MAUD can be recommended as a methodology for use in producing HEPs needed for PRA and for entry into the Human Reliability Data Bank.

Experience gained in the test application provided a basis for several key recommendations in future applications of SLIM-MAUD.

The recommendations for future applications of SLIM-MAUD are:

- For specific PRA applications, tasks to be assessed should be defined as concretely and completely as possible to ensure the reliability of SLI estimates.

- For generic applications expert groups should first identify a specific plant that typifies the range of plants to which the results are presumed to generalize. Such a procedure helps the experts to arrive at a shared, concrete image for performing their assessments.

- The recommended group consists of four members with a wide range of expertise. For the assessment of complex tasks, the majority of group members should have nuclear power plant operating experience. For simple tasks, the majority should be human factors experts.

- Classification into homogeneous subsets of 4 to 10 tasks each should be accomplished through group consensus procedures of task interrelatedness. It is recommended that the resultant interrelatedness judgment be clustered into subsets using the procedures described in Appendix A of Volume II.

In addition, the test application pinpointed certain gaps in our understanding of the application of SLIM-MAUD to nuclear power plant tasks. These gaps provide the basis for recommendations for future research.

The recommendations for future research on the SLIM-MAUD methodology are:

- The extent to which individual background variables such as education, experience, and type of license or certification affect SLIM-MAUD assessments should be investigated.

- Further investigation of the effects of expert group composition on SLI estimates should be undertaken.

- A data base of HEPs for nuclear power plant tasks that can be used as anchors for converting SLI values to HEPs should be developed.

- Additional empirical data from actual nuclear power plant experience, simulator, and laboratory studies should be gathered to test the criterion validity of SLIM-MAUD.

- The SLIM-MAUD results produced in this test application were compared to results produced by wholistic methods. Additional research to compare SLIM-MAUD to other decomposition techniques (e.g., STAHR) is recommended.

In summary, the test application of SLIM-MAUD, the most recent phase in a multiphase research program, provided considerable support for the methodology's practicality, acceptability, and usefulness to the estimation of HEPs where actuarial data are unavailable. Thus, SLIM-MAUD can be a useful tool to support HRA segments of PRA work.

REFERENCES

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

Comer, M. K., Seaver, D. A., Stillwell, W. G., and Gaddy, C. D. (1984), "Generating Human Reliability Estimates Using Expert Judgment, Volume 1: Main Report," NUREG/CR-3688, Sandia National Laboratories, Alburquerque, NM.

Embrey, D. E. and Hall, R. E. (1981), "Quantification of Human Performance Shaping Factors," Proceedings of the ANS/ENS Optical Meeting on Probabilistic Risk Assessment, Portchester, NY, LaGrange Park, IL, American Nuclear Society.

Embrey, D. E (1983), "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, NY.

Embrey, D. E., Humphreys, P., Rosa, E. A., Kirwan, B., and Rea, K. (1984a) "SLIM-MAUD: An Approach to Assessing Human Error Probabilities Using Structured Expert Judgment, Volume 1: Overview of SLIM-MAUD," NUREG/CR-3518, Brookhaven National Laboratory, Upton, NY.

Embrey, D. E., Humphreys, P., Rosa, E. A., Kirwan, B., and Rea, K. (1984b), "SLIM-MAUD: An Approach to Assessing Human Error Probabilities Using Structured Expert Judgment, Volume 2: Detailed Analysis of the Technical Issues," NUREG/CR-3518, Brookhaven National Laboratory, Upton, NY.

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

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

Kruskal, J. B., Young, F. W., and Seery, J. B. (1973), "How to use KYST, a Very Flexible Program to do Multidimensional Scaling and Unfolding," Murry Hill, NJ: Bell Laboratories, unpublished.

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

Phillips, L. D. Humphreys, P., and Embrey, D. E. (1985), "Appendix D: A Socio-Technical Approach to Assessing Human Reliability (STAHR)" in Selby, D., Pressurized Thermal Shock Evaluation of the Calvert Cliffs Unit 1 Nuclear Power Plant, Research Report on DOE Contract 105840R21400, Oak Ridge National Laboratory, Oak Ridge, TN.

Phillips, L. D. and Embrey, D. E. (1985), "Appendix E: Quantification of Operator Actions by STAHR," in Selby, D., Pressurized Thermal Shock Evaluation of the Calvert Cliffs Unit 1 Nuclear Power Plant, Research Report on DOE Contract 105840R21400, Oak Ridge National Laboratory, Oak Ridge, TN.

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

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, Alburquerque, NM.

NRC FORM 335 (6-83)

U.S. NUCLEAR REGULATORY COMMISSION

BIBLIOGRAPHIC DATA SHEET

1. REPORT NUMBER (Assigned by TIDC add Vol. No., if any)

NUREG/CR-4016
BNL-NUREG-51828
Vol. I

2. Leave blank

3. TITLE AND SUBTITLE

Application of SLIM-MAUD: A Test of an Interactive Computer-Based Method for Organizing Expert Assessment of Human Performance and Reliability, Volume I: Main Report

4. RECIPIENT'S ACCESSION NUMBER

5. DATE REPORT COMPLETED

MONTH	YEAR
March	1985

6. AUTHOR(S)

E.A. Rosa, P.C. Humphreys, C.M. Spettell, and D.E. Embrey

7. DATE REPORT ISSUED

MONTH	YEAR

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

Brookhaven National Laboratory
Engineering Technology Division
Department of Nuclear Energy
Upton, NY 11973

9. PROJECT/TASK/WORK UNIT NUMBER

10. FIN NUMBER

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)

The U.S. Nuclear Regulatory Commission (NRC) has been conducting a multi-year research program to investigate different methods for using expert judgments to estimate human error probabilities (HEPs) in nuclear power plants. One of the methods investigated, derived from multi-attribute utility theory, is the Success Likelihood Index Methodology implemented through Multi-Attribute Utility Decomposition (SLIM-MAUD). This report describes a systematic test application of the SLIM-MAUD methodology. The test application is evaluated on the basis of three criteria: practicality, acceptability, and usefulness.

Volume I of this report presents an overview of SLIM-MAUD, describes the procedures followed in the test application, and provides a summary of the results obtained.

Volume II consists of technical appendices to support in detail the materials contained in Volume I, and the users' package of explicit procedures to be followed in implementing SLIM-MAUD.

The results obtained in the test application provide support for the application of SLIM-MAUD to a wide variety of applications requiring estimates of human errors.

15a. KEY WORDS AND DOCUMENT ANALYSIS

Human Factors
Subjective Expert Judgment
Performance Shaping Factors
Nuclear Power Plants Operation

15b. DESCRIPTORS

16. AVAILABILITY STATEMENT

17. SECURITY CLASSIFICATION (This report)

Unclassified

18. NUMBER OF PAGES

40

19. SECURITY CLASSIFICATION (This page)

Unclassified

20. PRICE

$