U.S.NRC

United States Nuclear Regulatory Commission

Protecting People and the Environment

NUREG/CR-6949
INL/EXT-06-11670

The Employment of Empirical Data and Bayesian Methods in Human Reliability Analysis: A Feasibility Study

Office of Nuclear Regulatory Research

AVAILABILITY OF REFERENCE MATERIALS IN NRC PUBLICATIONS

NRC Reference Material

As of November 1999, you may electronically access NUREG-series publications and other NRC records at NRC's Public Electronic Reading Room at http://www.nrc.gov/reading-rm.html. Publicly released records include, to name a few, NUREG-series publications; *Federal Register* notices; applicant, licensee, and vendor documents and correspondence; NRC correspondence and internal memoranda; bulletins and information notices; inspection and investigative reports; licensee event reports; and Commission papers and their attachments.

NRC publications in the NUREG series, NRC regulations, and *Title 10, Energy*, in the Code of *Federal Regulations* may also be purchased from one of these two sources.

1. The Superintendent of Documents
   U.S. Government Printing Office
   Mail Stop SSOP
   Washington, DC 20402–0001
   Internet: bookstore.gpo.gov
   Telephone: 202-512-1800
   Fax: 202-512-2250
2. The National Technical Information Service
   Springfield, VA 22161–0002
   www.ntis.gov
   1–800–553–6847 or, locally, 703–605–6000

A single copy of each NRC draft report for comment is available free, to the extent of supply, upon written request as follows:

Address: U.S. Nuclear Regulatory Commission
Office of Administration
Mail, Distribution and Messenger Team
Washington, DC 20555-0001
E-mail: DISTRIBUTION@nrc.gov
Facsimile: 301–415–2289

Some publications in the NUREG series that are posted at NRC's Web site address http://www.nrc.gov/reading-rm/doc-collections/nuregs are updated periodically and may differ from the last printed version. Although references to material found on a Web site bear the date the material was accessed, the material available on the date cited may subsequently be removed from the site.

Non-NRC Reference Material

Documents available from public and special technical libraries include all open literature items, such as books, journal articles, and transactions, *Federal Register* notices, Federal and State legislation, and congressional reports. Such documents as theses, dissertations, foreign reports and translations, and non-NRC conference proceedings may be purchased from their sponsoring organization.

Copies of industry codes and standards used in a substantive manner in the NRC regulatory process are maintained at—

The NRC Technical Library
Two White Flint North
11545 Rockville Pike
Rockville, MD 20852–2738

These standards are available in the library for reference use by the public. Codes and standards are usually copyrighted and may be purchased from the originating organization or, if they are American National Standards, from—

American National Standards Institute
11 West 42nd Street
New York, NY 10036–8002
www.ansi.org
212–642–4900

Legally binding regulatory requirements are stated only in laws; NRC regulations; licenses, including technical specifications; or orders, not in NUREG-series publications. The views expressed in contractor-prepared publications in this series are not necessarily those of the NRC.

The NUREG series comprises (1) technical and administrative reports and books prepared by the staff (NUREG–XXXX) or agency contractors (NUREG/CR–XXXX), (2) proceedings of conferences (NUREG/CP–XXXX), (3) reports resulting from international agreements (NUREG/IA–XXXX), (4) brochures (NUREG/BR–XXXX), and (5) compilations of legal decisions and orders of the Commission and Atomic and Safety Licensing Boards and of Directors' decisions under Section 2.206 of NRC's regulations (NUREG–0750).

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

U.S.NRC

United States Nuclear Regulatory Commission

Protecting People and the Environment

NUREG/CR-6949
INL/EXT-06-11670

The Employment of Empirical Data and Bayesian Methods in Human Reliability Analysis: A Feasibility Study

Manuscript Completed: September 2007
Date Published: December 2007

Edited by
B. Hallbert, INL
A. Kolaczkowski, SAIC

Idaho National Laboratory
Idaho Falls, ID 83415

Science Applications International Corporation
10260 Campus Point Drive
San Diego, CA 92121

E. Lois, NRC Project Manager

NRC Job Code Y6496

Office of Nuclear Regulatory Research

ABSTRACT

The U.S. Nuclear Regulatory Commission (NRC) is addressing issues related to the quality of Probabilistic Risk Assessment (PRA), including issues related to human reliability analysis (HRA) performed as part of PRA. Among the issues of concern is an inadequate use of human performance data in the estimation of human error probabilities (HEPs), as well as in testing or otherwise validating underlying models used in HRA to predict human performance under cognitively demanding conditions. In order to address issues related to the use of human performance data in HRA, the NRC is developing the Human Event Repository and Analysis (HERA) database (NUREG/CR-6903). In addition, in August 2005, the NRC hosted an expert workshop on the use of Bayesian and other quantitative formalisms in conjunction with empirical data, such as that available in HERA, to improve both the estimation of human error probabilities and the underlying assumptions and quantitative algorithms employed by different HRA methods.

This report contains a collection of papers that were produced as a result of the workshop. It also summarizes the peer review comments of a draft version of this report, includes conclusions about the feasibility of using empirical data and quantitative methods for HRA, and provides suggestions on how to proceed to address the issues under consideration.

Paperwork Reduction Act Statement

This NUREG does not contain information collection requirements and, therefore, is not subject to the requirements of the Paperwork Reduction Act of 1995 (44 U.S.C. 3501 et seq.).

Public Protection Notification

The NRC may not conduct or sponsor, and a person is not required to respond to, a request for information or an information collection requirement unless the requesting document displays a currently valid OMB control number.

FOREWORD

This report documents a collection of papers that were produced as a result of a workshop conducted by the U.S. Nuclear Regulatory Commission (NRC) on August 10-11, 2005. The objective of the workshop was to explore the possibility of developing quantitative approaches, including Bayesian-type approaches, that would allow the use of readily available human performance data for informing human reliability analysis (HRA). To date, readily available human performance data have not been used extensively in HRA because probabilistic risk assessments (PRAs) of human events are rare, and databases contain only a few events of interest. In the absence of objective data, considerable judgment is used in the estimation of the likelihood of human failure events in PRA. Similarly, the models of human performance that form the underlying technical basis of many HRA methods may benefit from data obtained from representative HRA contexts. However, with the increased use of HRA results in regulatory decisions, the need to limit the use of subjective judgment by utilizing existing experience has become a focus of NRC's PRA Quality Program.

The workshop participants were recognized experts in the areas of PRA, HRA, data analysis, and statistics. Technical presentations addressed incorporating empirical data in HRA, and participating experts provided feedback on the proposed approaches. Experts were focused on examining the degree to which proposed methods provide a theoretically valid framework for HRA, if the examples provided demonstrated a method's applicability and usefulness, and what needs to be done to further develop such methods to address HRA needs. The results of the workshop were documented in a draft report which was peer reviewed by nationally and internationally recognized experts.

This report summarizes the technical approaches proposed, as well as the peer review results. It also includes a discussion of the technical work needed to further demonstrate the usefulness of such techniques in HRA. The work supports addressing outstanding technical issues identified in "Plan for the implementation of the commission's phased approach to probabilistic risk assessment quality" [SECY-04-0118 and SECY-07-0042].

Christiana Lui, Director
Division of Risk Analysis
Office of Nuclear Regulatory Research
U.S. Nuclear Regulatory Commission

CONTENTS

Abstract iii

Foreword v

Acknowledgements xi

Acronyms xiii

1. MOTIVATION FOR THIS REPORT 1

1.1 Introduction 1

1.2 The PRA – HRA Relationship 2

1.3 The Availability and Usability of Human Performance Data 3

1.4 Workshop and Subsequent Activities 3

1.5 Purpose of This Report 4

2. THE FEASIBILITY OF EMPLOYING BAYESIAN TECHNIQUES AND OTHER MATHEMATICAL FORMALISMS IN HUMAN RELIABILITY ANALYSIS 5

2.1 Methodological Foundations and Applications of Bayesian Methods in HRA 5

2.1.1 Foundations 5
2.1.2 Advanced Bayesian Methods for Use in HRA 10
2.1.3 Concluding Remarks 15

2.2 Modeling Causal Mechanisms and Context in HRA 16

2.2.1 Background 16
2.2.2 Purpose 17
2.2.3 Method 18
2.2.4 Data Collected 19
2.2.5 Analyses Conducted 20
2.2.6 Summary 32

2.3 Bayesian Updating of PSF Effects and HRA Estimates 36

2.3.1 Introduction 36
2.3.2 Proposed Methodology 38
2.3.3 Numerical Example 43

2.4 The Use of HERA and Bayesian Analysis to Support Quantification in Context-Based HRA Methods 48

2.4.1 Introduction 48
2.4.2 Bayesian Approaches to Improve Quantification in ATHEANA: Description and Examples 49
2.4.3 Developing Generalized Contexts for Interpolation 51
2.4.4 Quantification Method Based on Data for Errors of Commission 53
2.4.5 A Caveat on Context: Plant-to-Plant Crew Variability 54

3. SUMMARY OF PEER REVIEW 59

3.1 Peer Review Team and Purpose 59

3.2 Summary of Peer Review Comments 59

3.3 Addressing the Peer Review Comments 60

4. SUMMARY AND CONCLUSIONS 61

4.1 Introduction 61

4.2 The Validity of Using Quantitative Techniques and Empirical Evidence for HRA Use 61

4.3 Examples of Informing Our Human Performance Assessments 63

4.4 Demonstrating the Feasibility of Using Empirical Evidence and Quantitative Techniques for HRA 64

4.5 Possible Steps for Implementing Suggestions to Improve the Use of Empirical Evidence and Quantitative Techniques in HRA 65

5. REFERENCES 67

FIGURES

Figure 1 Example Operator Response Reliability from a Limited Data Sample. 10

Figure 2 Aleatory distribution of failure rate. 13

Figure 3 Tossing a coin using all information. 14

Figure 4 Tossing a coin using partial information. 14

Figure 5 Beta Weights of PSFs in Conventional and Advanced Plant Settings. 24

Figure 6 Beta Weights of Performance Shaping Factors. 25

Figure 7 Factors and Factor Loadings of PSFs from Plant 1. 30

Figure 8 Factors and Factor Loadings of PSFs from Plant 2. 31

Figure 9 Factors and Factor Loadings from Plant 3. 31

Figure 10 Prior and updated PMF for factor 1 47

Figure 11 Prior and updated PMF for factor 2 47

Figure 12 Prior and updated PMF for factor 3 47

Figure 13 Prior and updated PMF for factor 4 47

Figure 14 Prior and updated PMF for factor 5 47

Figure 15 Prior and updated PMF for factor 6 47

Figure 16 Prior and updated PMF for factor 7 47

Figure 17 Prior and updated PMF for factor 8 47

Figure 18 Possible Models for Human "Error" Rate Variability. 49

Figure 19 The Multidisciplinary ATHEANA Framework. 50

Figure 20 Five CAP Distributions 52

Figure 21 The Five CAPs of Figure 3, Grouped as GCAPs 52

Figure 22 Shifting Strong-EFC Curve vs EFC 53

TABLE

Table 1 Table of Multiple Regression Results Using Data Collected From Different Plants and Scenarios. 21

Table 2 Beta Weights of PSFs from Regression Models. 22

Table 3 b-weights of PSFs from a U.S. plant 26

Table 4 Sensitivity Calculations of Transient Mitigation time Using Scenario-Specific Model Parameters. 27

Table 5 Factors and Factor Loadings of PSFs. 29

Table 6 PSF multipliers and assumed prior probabilities. 37

Table 7 Reported PSF frequency data. 44

Table 8 Examples of Data Including EOCs. 54

ACKNOWLEDGEMENTS

We gratefully acknowledge the following individuals who participated in the workshop and/or served as peer reviewers for this document: Mr. Bernhard Reer, Dr. Oliver Straeter, Dr. Curtis Smith, Mr. Dana Kelly, Mr. Andy Dykes, Dr. Gareth Parry, Dr. Nathan Siu, Dr. Yung Hsien Chang, Dr. Larry Blackwood, Dr. Dale Rasmussen., and Dr. Susan Cooper. The comments and suggestions they provided regarding the effort and the report are greatly appreciated. We also gratefully acknowledge the authors of the individual papers: Professor Ali Mosleh, Professor Sankaran Mahadevan, and Dr. Dennis Bley, whose efforts made this report a reality.

ACRONYMS

AO	Action Outcome
ATHEANA	A Technique for Human Events ANAlysis
BBN	Bayesian Belief Network
BOP	Balance of Plant (Operator)
CAP	Context Anchored Probability
CDF	Cumulative Density Function
EFC	Error-Forcing Context
EOC	Error of Commission
EOP	Emergency Operating Procedure
GCAP	Generalized Context Anchored Probability
HCLPF	High Confidence Low Probability of Failure
HEP	Human Error Probability
HERA	Human Event Repository and Analysis
HFE	Human Failure Events
HMI	Human-Machine Interface
HRA	Human Reliability Analysis
INL	Idaho National Laboratory
LER	Licensee Event Report
LOCA	Loss of Coolant Accident
LOFW	Loss of Feed Water
MLE	Maximum Likelihood Estimate
NRC	U.S. Nuclear Regulatory Commission
OECD/NEA	Organization for Economic Cooperation and Development / Nuclear Energy Agency
PDF	Probability Density Function
PIF	Performance Influencing Factor
PMF	Probability Mass Function
PRA	Probabilistic Risk Assessment
PSAM	Probabilistic Safety Assessment and Management
PSF	Performance Shaping Factors
P-T	Pressure – Temperature
PWR	Pressurized Water Reactor
RO	Reactor Operator
SLIM/MAUD	Success Likelihood Index Methodology / Multi-Attribute Utility Decomposition

SGOF	Steam Generator Overfill
SGTR	Steam Generator Tube Rupture
SPAR-H	Simplified Plant Analysis Risk Human Reliability Analysis
THERP	A Technique for Human Error Rate Prediction Method
UA	Unsafe Act
UOI	Unknown of Interest

1. MOTIVATION FOR THIS REPORT

Prepared by Bruce Hallbert and Alan Kolaczkowski

1.1 Introduction

In accordance with its policy statement on the use of probabilistic risk assessment (PRA) [60 FR 42622, 1995], the U.S. Nuclear Regulatory Commission (NRC) has been increasingly using PRA technology in "all regulatory matters to the extent supported by the state of the art in PRA methods and data." Therefore, it is crucial that decision makers have confidence in the results produced by PRAs in order to make appropriate risk-informed decisions.

To address PRA quality issues, the NRC has developed an "Action Plan–Stabilizing the PRA Quality Expectation and Requirements," [SECY-04-0118, 2004 and SECY-07-0042]. Among the issues of concern is the fact that predicting operator performance as reflected in human reliability analysis (HRA) results continues to be a source of uncertainty when decision makers attempt to use the findings of PRAs and similar risk-related studies (e.g., risks involving nuclear material usage in medical and other applications). In particular, the plan includes data collection for improving both the estimation of human error probabilities (HEPs) as well as for testing or otherwise validating underlying models used in HRA to predict human performance under accident conditions.

In order to address the need for using data from operational experience or other sources in HRA, the NRC is sponsoring the Human Event Repository and Analysis (HERA) project. The objective of HERA is to analyze and code human events reported in licensee event reports, inspection reports, and other sources, in a format and structure appropriate for HRA. The main objective of HERA is to provide empirical evidence about human performance so that HRA analysts can derive qualitative information regarding human failure under various conditions. Analysts can use qualitative information for understanding and questioning the assumptions used in HRAs as well as for directly testing the assumptions employed by the methods themselves by, for example, applying a method to evaluate historical events. The use of empirical data to directly support the estimation of HEPs is also an incentive for collecting empirical human performance data.

The NRC, in addition to funding Idaho National Laboratory (INL) to perform HERA [Hallbert, *et. al.*, 2006], is also supporting international efforts to obtain and organize human performance data relevant to NPP operations. In particular, the NRC supports the Halden Reactor Project in Norway where state-of-the-art nuclear power plant (NPP) simulators are used to design experiments to collect operator performance data in simulated conditions similar to those modeled in PRAs and the Organization for Economic Cooperation and Development Nuclear Energy Agency (OECD/NEA) efforts to develop a framework for collecting and sharing NPP events among member countries.

Analysts have historically avoided the use of observed data in HRA for many reasons, primarily because events modeled in a PRA are rare events, and hence the conditions under which humans must accomplish mitigation tasks are also rare. As a result, analysts have eschewed or limited their use of information gathered in Licensee Event Reports (LERs) and other observation-based sources because of inherent difficulties in employing such evidence as well as concerns related to it's direct relevance to events modeled in PRAs. However, although there

are few events in the NPP history that can be applied in PRA/HRA, there is a significant number of "risk-significant" events in NPP or other industries. If systematically mined, these events can be an important source of data for human performance under challenging conditions. Furthermore, operational experience related to less risk-significant events can also provide useful information with regards to how or why events occur and, more importantly, how events are recovered and do not become more risk-significant. Both types of information, the occurrence of human error and underlying causes and recovery from the errors, are modeled in a HRA and are very important aspects of HRA quality. In addition, mathematical frameworks, specifically the Bayesian framework, allows the use of "evidence" from various sources to allow estimation of probabilities in areas dealing with rare events. Other engineering areas (e.g., seismic) that also are dealing with rare events are using Bayesian techniques to estimate likelihood of events. The purpose of this workshop was to discuss the potential use of Bayesian techniques in HRA.

Before directly addressing the feasibility of using quantitative methods and empirical data for HRA, the role of HRA in today's risk-informed regulatory environment is discussed.

1.2 The PRA – HRA Relationship

To support the probabilistic models and calculations of PRA, HRA provides a means to identify, and estimate the probabilities of human failure events (HFEs) modeled in a PRA. The discipline of HRA, and particularly its use in NPP PRAs, includes formalized analytical techniques for examining the potential for operators to perform unsafe actions, to commit inadvertent errors, and the failure to act and estimate their likelihood. These techniques embody the use of task analysis, models, data, and considerable judgment to assess operator performance and its impact on the overall risk. This is done by assessing the potential for unsafe acts and errors during both routine operations (e.g., failures while performing equipment surveillances) and potential accidents including operator unsafe acts and errors or their failure to act when needed that may contribute to those accidents (e.g., failure to properly initiate safety system operations).

HRA technology has evolved over the past thirty years, in response to our needs to better model human performance in a PRA, better reflect design and operational features of a continually evolving industry, and improved understanding of human performance in the behavioral sciences. Simple modeling and quantitative techniques developed over twenty-five years ago continue to be used today. For instance, human failure events that are typically classified as pre-initiator events, involving failure to properly restore equipment after test or maintenance and miscalibrations during routine operation of the plant, are typically analyzed using early HRA methods that appear to remain adequate even for today's uses. However, methods have been developed to model and quantify post-initiator human events, i.e., human failure events that may occur during operator response to a plant upset. Newer methods try to depict those influencing factors that may be particularly relevant to the conditions under which human actions could be performed, e.g., the nature and speed of changing plant conditions and the availability and clarity of cues about the plant state.

As we have improved human-machine interfaces in NPPs, thus making operator implementation errors less likely, it has become increasingly important to understand and better model the cognitive aspects of human performance within the context of situations that operators may experience. This, along with the increasing use of PRA and HRA results to make risk-informed decisions, has required more complex and higher fidelity modeling as well as greater reliance on improved quantitative techniques. Thus to support the uses of this more sophisticated modeling, data is also needed to better support the resulting HEP estimates using these more complex models.

1.3 The Availability and Usability of Human Performance Data

Current human performance models and quantitative estimates provide useful and reasonable results. Nevertheless, HRA practitioners are still working to obtain and use sufficient real world experience to (a) gauge the appropriateness of models and the qualitative insights they provide as well as (b) gauge and improve the accuracy of our HEPs that are currently based on considerable judgment without a comparable level of supporting empirical evidence. The use of considerable judgment, along with inherent stochastic characteristics associated with human performance, contribute significantly to the uncertainties in HRA results. This is especially the case for the post-initiators, since serious challenges to operator performance in the form of plant upsets tend to be rare, such experience is slow in coming. Thus it is desirable in NPP applications to use that data that is available to validate and improve HRA methods, their associated predictive models, and quantification techniques.

The "recording" of human performance as well as the influencing factors important to human behavior can be found in licensee event reports (LERs), other incident reports, inspection reports, licensee operator qualification examinations, simulator training experiences, special design and validation studies (e.g., control room design reviews), behavioral science experiments and other controlled studies, similar international sources of data, and other (non-nuclear) experience. Much of this data could be used, to support the development and improvement of human performance models needed in HRA, and in fact such information has been used to develop HRA models (e.g., ATHEANA). However, such data have not been traditionally used to directly derive HEPs of interest in PRAs. As stated above, serious challenges to operator performance tend to be rare, and hence such data are not used to create probabilities in the classical form (i.e., *x failures ÷ n opportunities*).

Furthermore, experience strongly suggests that human failure types and rates change depending on the situation encountered. As a result, no human performance data has been created in a form useful to the frequentist approach. Because of the inherent difficulties to create databases for direct HEP estimation, HRA has relied on developing models for human performance using theories and understanding of human behavior at the time of their development in conjunction with some empirical data. The result is that all HRA models involve considerable judgment to predict HEPs and the factors that cause humans to fail in various situations.

So the question arises "what can we do with all these various and often incomplete data (i.e., empirical evidence) to validate or improve our HRA models and techniques, and the qualitative and quantitative results they produce so as to have greater confidence in those results?" To answer this question, it is recognized that HRA is not the only PRA area that is dealing with "sparse data" or data not easily useable for our methods and models, and that many other areas and applications (e.g., seismic risks) are dealing with this same issue. Their solution has been to utilize a variety of quantitative techniques, including Bayesian approaches. It is reasonable to look to other PRA areas and the approaches they have taken and to consider whether they provide an avenue that may be followed to address similar needs in HRA, as well as to consider other approaches that may not yet have been tried.

1.4 Workshop and Subsequent Activities

To this end, a workshop was held in Rockville, MD, August 10-11, 2005, in which meeting participants were invited to present and discuss quantitative approaches suitable for using human performance data from various sources in HRA. Meeting attendees were national and international experts in the areas of PRA/HRA, Bayesian, and classical statistics. Five of the

experts presented or proposed approaches for utilizing "evidence," like that contained in the NRC's Human Event Repository and Analysis (HERA) system, in HRA. The focus of the meeting was the identification of (new or old) "promising" approaches for using available information to better identify the factors that influence human behavior in PRA-relevant contexts and the kinds of human failure events these factors could result in, and then to estimate the likelihood of their occurrence.

The results of the workshop were documented in a draft report that was submitted for peer review. The draft report was reviewed by those workshop participants whose role was to provide feedback to the presenters as well as by additional recognized experts in PRA, data analysis, and statistics. As a result, technical papers were produced that documented four technical quantitative approaches on the use of empirical data in support of HRA. Each technical paper addresses some aspect(s) of the workshop questions that are presented in Section 1.5. In preparing this report, two workshop proposals were structurally consolidated (those from Drs. Mosleh and Smith in Section 2.1). This consolidated paper and the remaining papers are included in Section 2 of this report.

1.5 Purpose of This Report

The purpose of this document is to summarize presentations and discussions led by individual workshop participants addressing the feasibility as well as the associated issues relevant to using quantitative methods and formal frameworks to employ evidence for improving our human performance models and gaining more confidence in the qualitative and quantitative results produced by HRA methods. This includes the potential for using such methods to validate aspects of the HRA methods to the degree that may be supported with such methods. In particular, these specific questions were discussed in the workshop and are addressed in this report in order to be responsive to the overall purpose:

1. Do quantitative techniques offer a theoretically valid framework for using empirical evidence to inform our current HRA methods?

2. What are some examples of ways we could inform current HRA methods (i.e., provide illustrations)?

3. What more needs to be done to demonstrate the feasibility of using these methods and empirical evidence to inform current HRA methods?

Section 2 consists of four technical proposals for employing empirical information with quantitative methods. Each paper provides a discussion of the theoretical bases for its suitability in HRA in general, and in particular to the specific application proposed, practical issues associated with its development and use, the potential of its applicability in the particular area, expected results, and thoughts and recommendations for future work. Section 3 summarizes the results of a peer review of the draft workshop summary report. Section 4 provides a summary and preliminary conclusions relative to the overall purpose of this report, including specific observations about the three questions above.

2. THE FEASIBILITY OF EMPLOYING BAYESIAN TECHNIQUES AND OTHER MATHEMATICAL FORMALISMS IN HUMAN RELIABILITY ANALYSIS

2.1 Methodological Foundations and Applications of Bayesian Methods in HRA

Prepared by Ali Mosleh and Curtis Smith

2.1.1 Foundations

People have long recognized that some events are imperfectly predictable (e.g., human failures). Probability theory arose, in part, to deal with these types of problems. In the twentieth century, probability theory provided a good model for treating a broad class of physical, social, and other problems. Using a well-known nuclear example, while we may not be able to precisely predict which neutron will hit a U-235 nucleus during the nuclear fission process in a way that extends the chain reaction, our ability to estimate the probability of such an occurrence, on average, allows us to design and operate nuclear reactors in a way that is safe and predictable.

When interpreting probabilities, a *frequentist* believes probability is an objective property in the real world and applies only to events generated by a random process. A *subjectivist* believes probability is an expression of a rational person's degree of belief about an uncertain proposition, and, with feedback, assessed probabilities will, in the limit, converge to observed frequencies.

Using the subjectivist approach, in the Bayesian setting, probability is a measure of uncertainty, a quantification of degree of belief. It treats "degree of belief" in a logical and rational way, not merely as personal opinion. In this methodology, each unknown parameter is assigned an initial prior probability and distribution modeling our belief concerning the true value of the parameter. Then based on evidence, our prior belief about the parameter is updated, using Bayes Theorem, to produce a posterior belief. The final inference, that is the posterior belief, makes use of and is, in fact, conditional on the evidence, as this statement illustrates:

> For billions of years, the sun has risen after it has set. The sun has set tonight. With very high probability (or I strongly believe that, or it is true that) the sun will rise tomorrow. With very low probability (or I do not at all believe that, or it is false that) the sun will not rise tomorrow.

In human performance issues, for instance, the initial belief may be the extent that increasing complexity leads to a greater human error rate, or it may, for instance, be an initial HEP estimate. The evidence may be both qualitative and quantitative human performance data collected from observed events. The posterior or updated belief could be a more confident prediction as to how increasing complexity leads to a greater human error rate, or an updated value for the HEP. The theory and implementation of Bayesian techniques, as well as the use of empirical evidence, are covered extensively in the *Handbook of Parameter Estimation for Probabilistic Risk Assessment* [Atwood, *et al.*, 2003]. However, as stated in the Handbook's foreword, the information provided does not specifically apply to human error probabilities, but instead, to other modeled events, such as component failures. Nevertheless, many of the same

principles, theory, and illustrations seem relevant, which suggests there may be ways to use available empirical evidence specifically for HRA.

Bayesian techniques have been successfully used in many disciplines, and their use continues to grow. The recovery of the USS Scorpion illustrates the Bayesian potential. In May 1968, the US nuclear submarine USS Scorpion failed to arrive as expected at her home port of Norfolk, Virginia. The US Navy was convinced that the vessel had been lost off the Eastern seaboard, but an extensive search failed to discover the wreck. A US Navy deep water expert believed that it was elsewhere, and he organized a search south-west of the Azores based on a controversial approximate triangulation by hydrophones. He was allocated only a single ship to perform the search, and he took advice from a firm of consultant mathematicians in order to maximize his resources.

A Bayesian search methodology was adopted. Experienced submarine commanders were interviewed to construct hypotheses about what could have caused the loss of the Scorpion. The sea area was divided up into grid squares and a probability assigned to each square, under each of the hypotheses, to give a number of probability grids, one for each hypothesis. These were then added together to produce an overall probability grid. The probability attached to each square was then the probability that the wreck was in that square. A second grid was constructed with probabilities that represented the probability of successfully finding the wreck if that square were to be searched and the wreck were to be actually there. This was a known function of water depth. The result of combining this grid with the previous grid is a grid which gives the probability of finding the wreck in each grid square of the sea if it were to be searched.

This sea grid was systematically searched in a manner which started with the high probability regions first and worked down to the low probability regions last. Each time a grid square was searched and found to be empty its probability was reassessed using Bayes Theorem. This then forced the probabilities of all the other grid squares to be reassessed (upwards), also by Bayes Theorem. The use of this approach was a major computational challenge for the time, but it was successful, and the Scorpion was found in October of that year.

Recently, Bayes filtering has become a popular mechanism to distinguish illegitimate spam e-mail from legitimate e-mail. Many modern mail programs implement Bayesian spam filtering. Further, some server email filters make use of Bayesian spam filtering techniques, and the functionality is sometimes embedded within mail server software itself.

More relevant to nuclear power plant applications, seismic risks are typically estimated based on information about the level of robustness believed to exist for various types of hardware found in nuclear plants (e.g., motor control center cabinets, individual relays, cable trays). Based on evidence gained during walkdowns of actual installations of such equipment in a plant, estimates are made about how seismically robust a particular equipment item is in a particular plant. Each estimate provides a so-called HCLPF (High Confidence that there is a Low Probability of Failure) for that equipment given a seismic event.

Illustrations of using Bayesian techniques in current PRAs are also covered in the *PRA Parameter Estimation Handbook* [Atwood, *et al.,* 2003].

Ideally, human error probabilities should be estimated with direct evidence (e.g., N_E, the number of error of a specific type, and N_O, the number of similar opportunities to make such errors). Currently however, in the vast majority of cases, such direct evidence is not available. With some exceptions, all HEPs are estimated judgmentally or based on models, with parameters that are also estimated subjectively or with soft evidence (e.g., assessment of performance

shaping factors, (PSFs). Bayesian methods are particularly appealing since HEP estimation requires use of a variety of sources and types of evidence. This presentation reviews some of the methodological foundations for proper use of Bayesian methods and, through a characterization of the nature of the evidence in HRA, explores the various advanced Bayesian inference methods that can use such evidence to estimate HEPs.

In discussing the need for and feasibility of using Bayesian inference methods in HRA, we must first clearly define the relevant unknowns of interest. In their most general forms, these are defined as follows:

- AO ≡ Human Action Outcome (e.g., Failure, Success)
- P ≡ Probability that AO = Failure
- $\pi(p)$ ≡ Probability Distribution of p (which could be aleatory or epistemic in nature).

Two questions arise:

1. Why are we uncertain about AO? (this uncertainty is captured by p)
2. Why are we uncertain about p?

The answer is not independent of the quantitative model one uses. Examples of HEP estimation frameworks include the following:

- Model A (Direct Estimation, maximum likelihood estimate (MLE)):

  $p = N_E/N_O$

  where N_E = Number of errors observed

  N_O = Number of observations or opportunities

- Model B (e.g., A Technique for Human Event ANAlysis (ATHEANA))

$$p = \sum_i P(response \mid condition\ i)\, P(condition\ i)$$

- Model C (e.g., Success Likelihood Index Methodology / Multi-Attribute Utility Decomposition (SLIM-MAUD)) [Embrey, *et al.*, 1984]

$$p = f(PIF)$$

  where PIF are performance influencing factors (i.e., PSFs).

  Some examples of the "function" f are

- Tables
- Mathematical Function
- Expert Judgment.

One possible interpretation of "p" that fits all of the above models is that, given a very specific condition (external and internal), the operator response would be predictable as either success or failure. However, in reality we can only specify a class of similar but not identical conditions, a fraction of which lead to failure, and the rest result in success. That fraction is "p;" therefore, "p" is the product of our grouping of a spectrum of conditions (external and internal) as one context. Uncertainty about the first assumption (a part of model uncertainty) is also a source of "p." This is sometimes referred to as residual randomness of human response.

A possible interpretation of $\pi(p)$ is that it represents one or more of the following sources of variability and uncertainty:

- Variability of p from one subclass of context to another, all within a "context super class" (e.g., generic context). Some examples are
  - Crew characteristics variability
  - Stochastic (aleatory) variability in factors (e.g., PSFs) that characterize the context (e.g., variability in "time pressure" due to variability in time and sequence of events)
- Uncertainty about the assessed values or states of PSFs for the specific context of interest
- Model Uncertainty, arising, for example, from incompleteness of PSFs (or factors used to characterize the condition or context) to represent the condition class.

When estimating p and $\pi(p)$, it is essential to be clear about what sources and types of uncertainty each represents.

Introducing how Bayesian principles can be applied to estimation of these parameters, it is appropriate to introduce Bayesian inference, which is based on the following elements:

- The unknown of interest (UOI)
- What we know about the UOI (Prior)
- Other evidence (e.g., data or observations)
- Model of the process generating the evidence (Likelihood)
- Combined state of knowledge about the UOI (Posterior)

The mathematical expression of the Bayesian engine of inference is

$$\pi(p|E) = \frac{L(E|p)\pi_0(p)}{\int L(E|p)\pi_0(p)dp}$$

(Eq. 1)

where p is the UOI, and E is the evidence. This mathematical expression provides a mechanism for updating the prior state of knowledge $\pi_0(p)$ based on the likelihood of the evidence $L(E \mid p)$ to arrive at the posterior (updated) distribution $\pi(p \mid E)$ of p given E.

The key element of this process is the model of the evidence, such as the likelihood function, $L(E \mid p)$. The form of the likelihood function is directly tied to the nature of the evidence. For instance, for direct evidence, such as E= {N_E errors in N_O opportunities}, a binomial likelihood function would be appropriate, assuming that "p" does not change from trial to trial.

In order to use this Bayesian engine of influence, it is important to explore the types of evidence we encounter in relation to estimating human error probabilities. Some of the characteristics of the available information are:

- Different forms and types of information
  - Expert Estimates
  - HEPs generated by applying HRA Models
  - Actual counts of failure and success
- Non-homogenous evidence (different pieces of information from multiple sources)
- Incomplete, indirect, or partially relevant observations. Examples include:
  - HEP estimates based on data from situations other than the error of interest, for example, a different plant state, a different set of performance shaping factors, or different system and technology altogether
  - Human performance data from simulator experiments
  - Incomplete information on "success counts" (NO) or exposure space
  - Uncertainty and ambiguity in the classification of observed events, their contexts, and their causes.

For the simplest case among these (when we have the actual counts of failure and success), the Bayesian formulation will take the conventional form (Eq.1). Even in this simplest form, the Bayesian approach provides significant advantages. For example, considering the number of human failure events observed in the database, even when the number of trials (success data) is small, the impact of the data on the resulting posterior distribution might be visible. The actual effect is a function of the additional piece of information embodied in the prior distribution. To illustrate this point, we utilized Halden research data from one of several activities that operators were asked to perform during a series of experiments in 2002. The 2002 Halden experiment involved eight crews and eight scenarios (hence 64 trials), with no failures. The posterior probability of action failure is plotted in the following figure. This data was selected for several reasons, including demonstrating that Bayesian methods and limited data can make strong statements for even low-probability events (less than 1E-2 per activity). This low-probability application contradicts assertions that operator actions cannot be simulated if they are at the 1E-2 or lower probability level. It should be noted that using just half of the information from a single series of experiments (2002) of Halden data could improve our knowledge of events that are postulated to occur in this low frequency and probability (i.e., 1E-2) range.

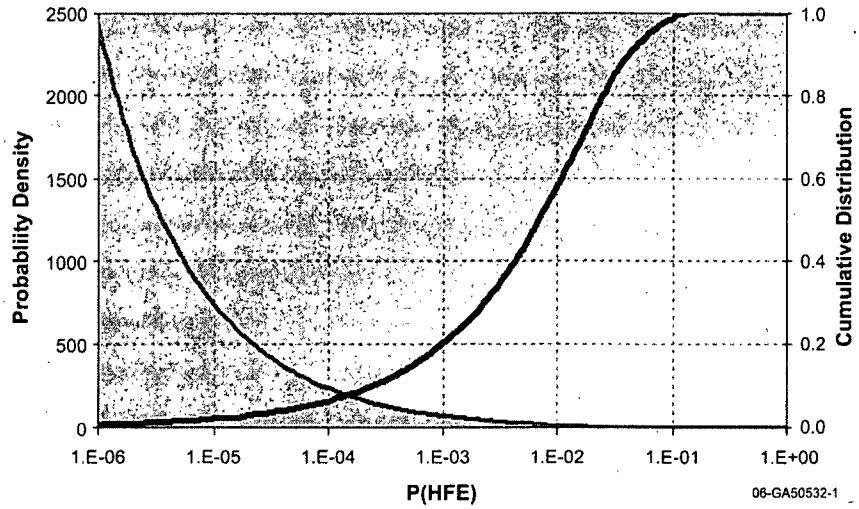

Figure 1 Example Operator Response Reliability from a Limited Data Sample.

For other types and grades of data listed above, the Bayesian formulation needs to go beyond the conventional approaches. Fortunately, while it is clear that the specific solutions would have to be formulated for HEP estimation, when dealing with similar situations, some of the techniques used in estimating hardware failure probabilities can serve as a good starting point. We summarize some these techniques in Section 2.1.2.

2.1.2 Advanced Bayesian Methods for Use in HRA

This section discusses a few examples of the more advanced Bayesian inference models that correspond to some of the characteristics of the HRA data mentioned earlier. The key element in all such inference models is the form of the likelihood function, which should be constructed from the following factors:

- Number and types of information
- Dependence (of information sources)
- Credibility (of data from experts and models)
- Applicability (to the HEP of interest)
- Homogeneity (of data points)
- Uncertainty (of evidence).

The methods described in the following are examples of how one or more of the above features of the evidence can be accommodated.

2.1.2.1 HEP Estimation with Multiple Types or Sources of Information

The formulation in case of multiple sources of information E = {I_1, I_2, ..., I_n} is standard and straightforward:

$$\pi(p|E) = \frac{L(E|p)\pi_0(p)}{\int L(E|p)\pi_0(p)dp}$$

(Eq. 2)

The second equation applies when sources of information are independent. Examples of I are:

- I_1 = Actual Event Counts
- I_2 = Expert Estimates
- I_3 = Estimates Based on HRA Models

The specific mathematical form of the individual likelihood functions varies again. depending on the specific type of information. These include the binomial distribution for Type I_1, and lognormal (multiplicative error model) [Mosleh, 1992] for I_2 and perhaps I_3.

For further reading see Mosleh and Apostolakis, 1985; Mosleh, 1992; Mosleh and Apostolakis, 1984.

2.1.2.2 Estimating Aleatory Uncertainty of "p"

In this case, the unknown of interest is an entire distribution, f(p), representing "inherent" variability of p due to any reason. One can assume a parametric form for f(p | $\underline{\theta}$), with a set of parameters $\underline{\theta}$ to be estimated based on the available evidence:

$$\pi(\underline{\theta}|E) = \frac{L(E|\underline{\theta})\pi_0(\underline{\theta})}{\int L(E|\underline{\theta})\pi_0(\underline{\theta})d\underline{\theta}}$$

(Eq. 3)

We then use $\pi(\underline{\theta} \mid E)$ to estimate f(p)

$$f(p) = \int_{\underline{\theta}} f(p|\underline{\theta})\pi(\underline{\theta}|E)d\underline{\theta}$$

(Eq. 4)

An example application is the case where the evidence is HEP estimates from different sources that reflect different context or conditions) [Mosleh, 1992]

E = {p_1, p_2, p_3, ..., p_n}

We can assume that the aleatory distribution of p is a beta distribution with parameters α and β

$$f(p|\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\,p^{\alpha-1}(1-p)^{\beta-1} \qquad \text{(Eq. 5)}$$

and then use E to estimate the epistemic distribution of α and β:

$$\pi(\alpha,\beta|E)=\frac{L(E|\alpha,\beta)\pi_0(\alpha,\beta)}{\int_\alpha\int_\beta L(E|\alpha,\beta)\pi_0(\alpha,\beta)\,d\alpha\,d\beta} \qquad \text{(Eq. 6)}$$

Finally we estimate f(p) with:

$$f(p)=\int_\beta\int_\alpha f(p\,|\,\alpha,\beta)\pi(\alpha,\beta\,|\,E)\,d\alpha d\beta \qquad \text{(Eq. 7)}$$

This approach has been used in PRAs to develop "generic distributions" for component failure probabilities from various plants, thereby preserving the plant-to-plant aleatory variability.

As an example, consider the case where six estimates are available for the failure rate of pressure transmitters. These estimates, along with the assigned measure of confidence expressed as the range factor used in modifying the likelihood function, are listed below. Note, for example, a range factor of 3 implies that the estimated failures per hour could range from 3 times higher to 3 times lower than the specific value shown.

Expert	Estimates in failure per hour	Assigned range factor
1	3.0E-6	3
2	2.5E-5	3
3	1.0E-5	5
4	6.8E-6	5
5	2.0E-6	5
6	8.8E-7	10

By applying the above set of steps, the following aleatory distribution of the failure rate is obtained (Figure 2). The aleatory distribution was assumed to be lognormal, and the likelihood functions were based on the logarithmic error model of (Mosleh, 1992). The Bayesian computations were performed using R-DAT software [Reliability Data Analysis Tool, Prediction Technologies, Inc.]).

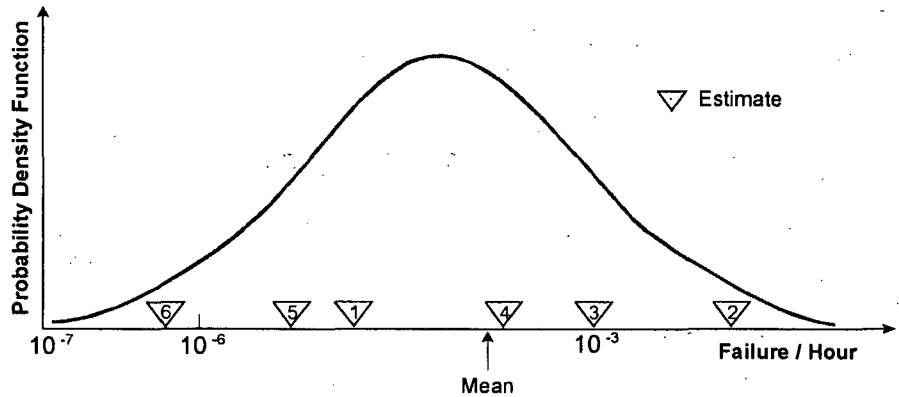

Figure 2 Aleatory distribution of failure rate.

2.1.2.3 Dealing with Evidence Uncertainty

The evidence E may be uncertain in nature due to the following factors:

- Uncertainty in event interpretation
- Partial Failures (quickly recovered error)
- Uncertainty about the success data (error opportunities)
- Uncertainty in cause classification (e.g., PSFs involved and their values and states).

In these cases, the content of E itself has an associated uncertainty which needs to be factored into the estimation of p. Assume that the uncertainty in E is expressed in the form of a probability distribution f(E). Three approaches have been suggested and used in the past (mostly in dealing with similar situations for hardware failure probabilities):

- **Weighted Posterior Method:** In this approach, each interpretation of the evidence is first used to obtain a posterior distribution for p. The final posterior distribution of p is obtained by averaging the individual posterior distributions (UE stands for Uncertain Evidence):

$$\pi(p|UE) = \int \pi(p|E) f(E) dp \qquad \text{(Eq. 8)}$$

- **Weighted Likelihood Method:** In this approach, each interpretation of the evidence is first used in a corresponding likelihood function, and the total likelihood function is then obtained by averaging the individual likelihoods:

$$\pi(p|UE) = \frac{\left[\int L(E|p) f(E)\right] \pi_0(p)}{\int \left[\int L(E|p) f(E)\right] \pi_0(p) dp} \qquad \text{(Eq. 9)}$$

- **Evidence Averaging:** In this approach, the expected value of the evidence is used to construct the likelihood function:

$$\pi(p|UE) = \frac{L(\overline{E}|p)\pi_0(p)}{\int L(\overline{E}|p)\pi_0(p)dp} \qquad \text{(Eq. 10)}$$
$$\overline{E} = \int E\, f(E)$$

One form of evidence uncertainty can be due to the fact that only partial evidence is available. To illustrate how partial information can be incorporated into Bayesian applications, we define a case where coin toss data are collected but filtered. In reality, the number of coin tosses may be known, but if we were provided with only the number of heads, we can still make an inference. (In the nuclear industry, Licensee Event Reports [LERs] are examples of missing data because we do not know how many trials [the "n" in the binomial likelihood model] were realized to provide the recorded failures). First, let us estimate the probability of tossing heads using the full data set so that we have a "known" condition to compare. To perform this analysis, a Bayesian analysis tool called WinBUGS [Open Source Software, 2003] was used.

Assuming that we use a uniform prior uncertainty distribution on the probability of heads, we find the results from a known case involving 10 tosses with 5 heads and 5 tails in the resulting probability density plot in Figure 3. Using partial information (we only know the number of heads and do not know the total number of tosses), we see the result shown in Figure 4, whereby it is assumed that there may have been as many as 30 tosses (there were actually only 10). When comparing the "all information" case, where the number of heads and number of tails are known, to the "partial information" case, where only the number of heads is known, one should note that the effect on the estimated probability of heads is noticeable but not dramatic. Even though a significant amount of information has been lost, the expected value decreases slightly from 0.501 to 0.454. The net result of having missing data was that the probabilistic information changed, but only slightly, even though we were somewhat uninformed on the missing information. However, this diffuse state of knowledge only leads to an 11% decrease in the mean estimate for the probability of tossing heads. *For further reading see Mosleh, 1986 and Groen, 2005.*

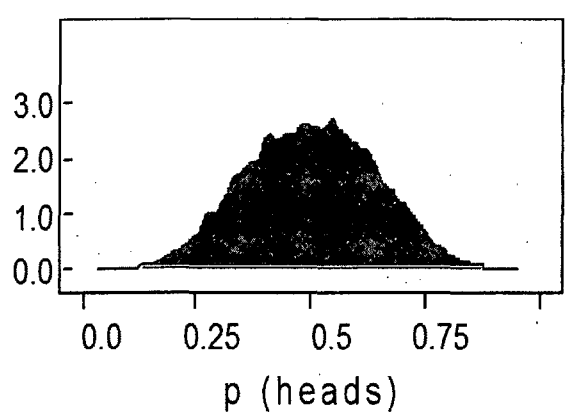

Figure 3 Tossing a coin using all information

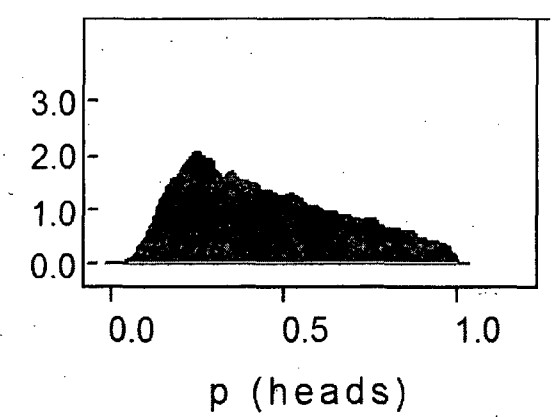

Figure 4 Tossing a coin using partial information

2.1.2.4 Dealing with Evidence Relevance

Methods suggested for situations where the evidence is partially relevant to estimation of p are simple extensions of the case where the evidence is totally applicable and relevant (Eq.1). More specifically, when E is believed to be partially relevant, the analyst can assign a relevance factor

of $0 \leq \omega \leq 1$ as a subjective measure of the degree of relevance of the data for the HEP of interest. One of a number of options to use such partially relevant data is to modify the likelihood function as follows:

$$\pi(p|E) = \frac{[L(E|p)]^{\omega} \pi_0(p)}{\int [L(E|p)]^{\omega} \pi_0(p) dp}, \qquad 0 \leq \omega \leq 1 \qquad \text{(Eq. 11)}$$

When $\omega=1$ (i.e., when the evidence is fully relevant), the evidence that is used is full strength, as in the conventional form of Bayes theorem. The other extreme is when $\omega= 0$ (when the evidence is totally irrelevant), and the likelihood function will become a constant (L=1) and will have no influence in forming the posterior distribution. For values of $\omega<1$, the likelihood function will be wider than the case for $\omega =1$, reducing the influence of data on the posterior distribution.

This formulation can be used, for example, to "discount" simulator data when used in estimation of HEPs for real accident conditions.

For further reading see Groen and Mosleh, 2005 and WINBUG, 2003.

2.1.3 Concluding Remarks

This discussion has highlighted many of the inherent flexibilities of Bayesian methods in dealing with the sparse, diverse, and uncertain data with which HEPs have to be estimated. An equally important message in presenting the range of available Bayesian techniques is that our view of what constitutes useful information in HRA data gathering and database development also needs to be broadened. This has implications for how we design databases and data classification schemes. For instance, the data classification could include assignment of applicability of causes (or PSFs) of a given human failure event when analyzing the data for another application. This enables some degree of generic data specialization for the application of interest. As another example, the success data, though only partially known, could be specified with an uncertainty range.

We also note that the methods discussed here are all sufficiently mature for near-term implementation. In fact, as stated above, many of them have already been applied for the hardware failure data assigned to PRAs. Advances in computing and numerical methods have made solving multi-dimensional and hierarchical Bayes formulations quite practical. The most immediate need or step is the identification of the sources of information, and mapping the identified types to the mathematical frameworks described above.

2.2 Modeling Causal Mechanisms and Context in HRA

Prepared by Bruce Hallbert

2.2.1 Background

Issues related to the quality of human reliability data and models date, at least, to the Probabilistic Safety Assessment and Management (PSAM) 3 conference in 1996. At that meeting, a vigorous discussion initiated in one of the HRA sessions between proponents for better human reliability data production and proponents for better human reliability methods or models. Those holding the view that what was most urgently needed at that time was more and better human reliability data argued that the uncertainties in error identification, modeling, and quantification could best be reduced through sources of evidence. Proponents for improved models cited the fact that many methods in use at that time (a.k.a., first generation methods) overlooked important cognitive abilities of the human operator and only implicitly treated such issues. As a consequence, data production efforts without the appropriate model(s) informing its generation or capture would be likely to replicate the inadequate treatment of cognitive and contextual factors prevalent at the time.

Since 1996, a number of advances have been made in HRA, particularly with respect to modeling and accounting for cognitive and contextual influences. However, less progress has been made in developing sources of information or data that can be used to assist analysts in identifying the appropriate human failure events to be included in their PRAs, how to best represent such failures, and how to better estimate the likelihood of their occurrence. Recently, several efforts have begun and are now underway to generate data specifically aimed at informing HRA activities [Hallbert, *et al.* 2004; Kirwan, *et al.,* 2004; Hallbert, *et al.*, 2006]. Concurrent with the development of sources of information, methods must also be developed and demonstrated to permit their use in risk-informed applications and, most specifically, PRAs. An advantage of Bayesian methods for using such information is that they are capable of using all available information. They can also be used to make predictions directly about the quantity or behavior of interest, and if employed properly, can account for the causal and conditional nature of context and performance.

In cases where we may have access to a source of human performance information that is complete in the sense that it provides information about outcomes of interest (i.e., success or failure of human actions) and the total number of opportunities or demands, classical statistical methods may be used to estimate the human reliability parameter of interest, which is typically assumed to be unknown or uncertain. For example, the number of correct responses to a given stimulus, the latency in response to the stimulus, and other characteristics may be estimated from simple response data. However, behavior is not only random and variable from time to time (i.e., random and uncertain), but is also causally linked to characteristics of the human as an organism as well as to the context in which behavior is elicited. Unless explicitly directed for by models, the analysis of response data is likely to overlook subtle and important determinants of behavior. Consequently, a classical statistical treatment of such response data may not reflect the response reliability of human action in contexts that were not represented in the environment that produced the source data.

This presents a quandary for human reliability modeling; classical statistical treatment of human performance data in a manner similar to the treatment of equipment performance data for purposes of estimating reliability parameters may not be appropriate, yet models that dictate

how to incorporate and analyze human performance data are approximations with uncertainty. So, even though Bayesian methods are capable of employing all of the available data, this is only true if the likelihood function employed in the particular form of the Bayesian formalism is structured to accept all of the available data. This is not trivial and may be as important as having good quality data available for use.

As an illustration of the issue, consider the manner in which information on context or performance shaping factors (PSFs) are typically accounted for in current HRA methods. In the Technique for Human Error Rate Prediction (THERP), for example, static multipliers are used to treat the effects of such important factors as stress and experience. Doubling or tripling the rate of error is recommended for some actions where stress is high and experience of performing an action is low. For a given action such treatment may be warranted, or even for some contexts, yet such treatment may be unwarranted for many others. The HEP *cum* PSF approach is prevalent in many HRA methods, and thus the accuracy of the resulting risk metrics that are derived from our analyses and the uncertainty in those results are dependent on knowing the relationships between the PSFs we model and the resulting human performance for different situations, or contexts.

What is needed is the ability to relate elements of the environment to performance using formal models that can express the relationships in terms of the quantities of interest to reliability analyses. This involves a number of related activities. First, data are needed that will allow us to test assumptions about the relationships between PSFs and operator performance, and to develop models of the relationship(s). This includes distinguishing important factors from irrelevant factors in contexts of interest. It also includes estimating the magnitude of effects of the relationships, and the degree to which these relationships vary in different contexts of interest to PRA. HRA methods currently treat the effects of PSFs on performance reliability through either individual assessments or by aggregating their effects. That is, an important underlying hypothesis of many HRA methods currently in use is that contextual factors can be addressed by modifying a nominal HEP by a multiplication factor that represents the effect of a particular PSF (or set of PSFs) on the nominal HEP. This approach overlooks potentially important interactions among PSFs and the way(s) they may affect performance. Thus, an additional application for PSF data concerns the study of their systematic interaction and associated implications for improving the manner of accounting for their influence in human reliability models and methods.

2.2.2 Purpose

This section presents research that demonstrates an approach to address PSF representation in human performance modeling by obtaining empirical information related to performance shaping and contextual factors. The approach employed here involved collecting information related to PSFs and objective operator performance across several studies of operator performance, and then relating the PSFs to performance through a linear mathematical model. This was done to (1) assess whether PSFs are predictive of important aspects of operator performance, (2) distinguish between more important and less important PSFs, (3) demonstrate a method to relate their influence to a general model of operator performance, and (4) demonstrate methods that may be used to illustrate the systematic interactions of PSFs in PRA-relevant scenarios.

The data and analyses reported here are the products of empirical research, and they are intended to motivate discussion and consideration of approaches to advance the use of data to support improvements in modeling human performance in PRA-relevant contexts. Because of the small sample sizes involved that were the result of opportunistically collecting these data

(i.e., as ancillary to the purposes for which the main studies were sponsored), these results should be viewed as preliminary and illustrative.

2.2.3 Method

Licensed commercial nuclear power plant operators who participated in several studies involving simulated accident (i.e., PRA-relevant) conditions were presented with scenarios in which their performance was measured. The objectives of these studies differed, as did some of the conditions in the studies. In one study, for example, crews performed simulated control room activities to mitigate design basis events in order to support resolution of regulatory issues associated with the operational safety of nuclear power plants [Hanson *et al.*, 1987]. In another study, operating crews performed in either a minimum or normal staffing complement per 10 CFR 50.54(m) in order to evaluate the effects of staffing changes on crew performance [Hallbert *et al.*, 2000]. Crews in the latter study also performed their activities in either a conventional control room or an advanced control room setting. Operators who participated in all of these studies performed as crews with others they were assigned to work with at the time. Although the data presented here were generated from the two referenced studies above, the data are not presented or discussed in the referenced reports as they are ancillary to the purposes for which the studies were commissioned.

In these studies, crews were from pressurized water reactors (PWRs). Several scenarios were replicated, including a sustained total loss of feed water, a steam generator overfill scenario, and a loss of coolant scenario (a steam generator tube rupture transient in two simulator settings; a 0.5 in.^{2} small break LOCA in the other). The scenarios thus represented a range of thermal-hydraulic challenges, including overheating, overcooling, and loss of coolant. In this way, scenarios were qualitatively matched across studies. Crews were free to interact with the plant simulation to effect control and stabilization as they were trained, and were further expected to perform activities as they normally would, were such events to actually occur. Role-play of external on-site personnel (e.g., auxiliary operators) and off-site personnel (e.g., regulatory and municipal authorities) was used to simulate important command, control, and communication activities that would also be present in events such as those simulated.

Each scenario included one or more critical mitigation activities, which, if performed correctly and in a timely manner, would help restore plant functions and prevent further degradation of simulated plant thermal-hydraulic conditions. In the case of the sustained loss of feed water (LOFW) transient, crews had to initiate feed and bleed actions – a sequence of actions that involved producing a lineup for high-pressure safety injection, opening a relief path through the pressurizer, and sustaining flow to achieve core cooling. In the case of the steam generator over fill (SGOF) transient, crews had to respond to an uncontrolled feed flow to the affected steam generator(s) and act to control flow by controlling valve positions and feed water pump operations manually to prevent water from reaching the main steam lines and producing a potentially more damaging event on a relatively short time scale. All of the critical mitigation actions involved planning, carefully sequencing a set of actions, and coordinating them within the crew in order for the actions to be timed correctly and to be complete. In all cases, crews could rely on their previous experience, training, and procedural guidance to assist in decision making, action plan formulation, and carrying out mitigating actions. However, all scenarios were designed to be challenging and required crews to carefully time and carry out their actions with precision in order to achieve their goals. The scenarios include PRA- and HRA- relevant human actions, both from the standpoint of the actions needed to mitigate the transient(s) and secure systems as well as the degree of challenge and realism involved.

Following each scenario, operating crew members were asked to evaluate a set of PSFs in terms of the effects these PSFs had on their performance in mitigating the transient(s). A questionnaire-style instrument was used to collect this information from crew members individually. Each questionnaire asked the operator to rate the influence that a set of PSFs had on their performance of the critical mitigation action in the scenario that they had just completed in their training simulator. Thus, each administration of the questionnaire was scenario-specific, as well as focused on a critical mitigation action from the scenario. Prior to data collection, researchers and simulator staff familiarized crews with the method for rating PSFs and the general format of the questionnaire by evaluating a sample task collectively. Following each scenario, crew members filled out a questionnaire. In the case of the first conventional plant setting, 28 crew members provided data on three scenarios. In the second conventional plant setting, 19 crew members provided data. In the advanced plant setting, 10 crew members participated in the study. In all, 57 crew members filled out questionnaires.

Seven PSFs were evaluated and are included here with the definitions as they were presented to the operating crews.

Procedures refers to how easy the procedure is to follow, and how clearly it directs the operator to take action in the task being considered.

Instrumentation (shown as Human-Machine Interface or HMI in subsequent analyses) refers to how easy the displays and controls located on consoles and back panels are to locate, read, and or operate.

Training refers to how well-schooled the operations staffs are in terms of classroom theory and simulator training, and what bearing this has on the task being reviewed.

Information available refers to the extent to which information presented to the operator is accurate and easy to understand.

System Feedback refers to how the operator knows s/he has taken the appropriate control action or how s/he is made aware of the general plant state, i.e., through annunciator systems or plant automatics.

Workload refers to how busy the operator is while performing tasks and how difficult these tasks are to perform.

Stress refers to the pressure experienced. The stress could come from a number of sources, such as being unsure there was enough time to complete the task, having a lack of experience in that particular situation scenario, or bearing the burden of responsibility that unsuccessful completion of the task would have negative consequences for the plant.

2.2.4 Data Collected

A scale for rating the PSFs was provided next to each PSF on the questionnaire. The scale was arranged from 1 to 5, with 1 meaning that the PSF hinders performance, 3 meaning that the PSF has no effect on performance, and a 5 meaning that the PSF helps performance on the task being evaluated.

In addition to the PSF data, performance data related to transient mitigation were obtained in each scenario. A key mitigation activity was identified during the design of each of the scenarios. Training and licensing personnel from the utilities identified operator activities that, if

performed, would materially contribute to the mitigation of the transient and, if not performed, would contribute to plant condition degradation. These are similar to or are the same as human actions represented in PRAs. Performance of each key mitigation activity was obtained from the operator input logs of the simulator following completion of the scenarios.

2.2.5 Analyses Conducted

A number of analyses were conducted on the data collected in this research. These analyses represent different, but related areas of interest important to HRA modeling of human performance. Each analysis is presented below.

2.2.5.1 Question 1: Are PSFs predictive of important aspects of operator performance?

The data were analyzed using multiple regression and correlation techniques to determine the extent to which the PSFs correlate with and are predictive of crew performance on measures of objective performance. In these analyses, the independent measures were the operators' ratings of PSFs, and the dependent measure was the amount of time each crew used to mitigate the simulated transient. The time measure reflects the window from the initiation of the transient to the time at which crews performed the mitigation activity. The linear multiple regression model used to assess the effects of PSFs on transient mitigation can be summarized as:

$$Y = a + b_1x_1 + b_2x_2 + \cdots + b_nx_n \qquad \text{(Eq. 12)}$$

where Y represents the time taken by a crew to mitigate the transient, a is the constant of the linear model (i.e., the intercept), x_1 through x_n refer to the discrete value of a particular PSF, b_1 through b_n represent the weight or contribution a particular PSF makes in predicting the time taken to mitigate the transient. n, in this case, is represented by the seventh PSF. Three analyses were made of the PSF regression model data. The first analyses are of the strength of model predictions as a function of whether the models assessed between PSFs and performance are made at an aggregated level (i.e., pooling data from several scenarios or plant settings) or at an non-aggregated level (i.e., at the level of the individual scenario or plant setting). For these model evaluations, data were used from all plant settings and from all the scenarios that were employed in those settings.

Table 1 shows the values of multiple regression models constructed using data collected in the different plants for all scenarios. At the highest level of aggregation (i.e., two plants, all scenarios), the multiple regression of PSFs on operator and crew performance is 0.41, showing that a relationship exists between the PSFs and operator performance. Note, however, that multiple regressions at the individual plant level reveal consistently higher results than when aggregated across plants or even across scenarios in general.

In both conventional plants, the predictive strength of the PSF models (measured through the magnitude of the multiple regression coefficient) becomes progressively stronger as they approach the individual scenario level. This shows that PSFs become more predictive of performance when analyzed in a specific context. It also suggests that the strength of generalizations about PSFs beyond the original contexts in which they were collected may not be robust. Returning to the question posed at the outset of this section, we may conclude that PSFs are predictive of aspects of operator performance. A couple of notable points attend this conclusion: (1) Inferences appear to be stronger when made about the specific contexts that are

present during data collection and (2) Generalizations beyond these contexts may not be warranted.

These conclusions are intended to be illustrative of the methodology for analyzing the PSF data suggested here. Any conclusions are not yet robust, especially owing to the limited sample sizes used in constructing these models. Green [1991] discusses issues associated with the ratio of the number of cases to independent variables included in linear regression models. A general rule of thumb of $N \geq 50 + 8k$ (where k is the number of independent variables used in the linear regression model) is recommended for testing multiple correlation models. In these studies, limited access to operational staff precluded collecting data sufficient to support the full use of inferential statistics, or generalizations of these results beyond the purposes of demonstrating trends in the strength of relationships among PSFs and performance described here. Yet, some important trends are noted in the results of these model applications, and the general methodology may hold some promise for developing human reliability models that are contextually anchored.

Table 1 Table of Multiple Regression Results Using Data Collected From Different Plants and Scenarios.

Regression Model	Multiple R	Degrees of Freedom (k, n-k-1)
Two Plants, All Scenarios*	0.41	7, 65
Conventional Plant 1		
All Scenarios	0.66	7, 33
LOFW	0.94	7, 6
SGOF	0.51	7, 5
SGTR	0.86	7, 6
Conventional Plant 2		
All Scenarios	0.40	7, 48
LOFW	0.71	7, 16
SGOF	0.84	7, 16
LOCA	0.60	7, 16
Advanced Plant		
All Scenarios	0.34	7, 24
LOFW	0.40	7, 4
SGOF	**	
SGTR	0.75	7, 3

*k=number of PSFs in the model (7 in all cases); n=sample size

**Statistics for this model could not be computed owing to missing data from one crew

2.2.5.2 *Question 2: Can these data be used to distinguish important PSFs from less important factors in contexts of interest?*

The second set of analyses compare the model Beta weights of PSFs between plant settings and scenarios to assess the similarities and differences in model coefficients and contributions of these variables in predicting performance. Table 2 shows the Beta weights of the regression model using the PSFs as predictors of operator and crew performance. Referring back to Equation 12, the model weights (i.e., the $b_1 \ldots b_n$ in the equation) are standardized with a mean of 0 and range. Since the model weights are standardized, values will fall in the range $-1.00 \leq \beta_{weight} \leq 1.00$. The absolute value of the weight of the model coefficient, $|\beta_{weight}|$, indicates the relative strength of the weight of the coefficient. Values closer to 1.00 represent a stronger contribution of the PSF than values closer to 0. The Beta weights in the scenario-specific models range from nearly 0 (a PSF having a negligible effect on performance) to 0.92 (a PSF exerting a strong effect on performance).

Table 2 Beta Weights of PSFs from Regression Models.

	SGTR		LOFW		SGOF	
PSF	Advanced Plant	Plant 1	Advanced Plant	Plant 1	Advanced Plant*	Plant 1
Procedures	0.1420	0.4658	0.1708	0.7689		-0.1052
HMI	-0.4274	-0.2785	-0.2498	0.9211		-0.2798
Training	0.4792	-0.1295	-0.1126	-0.5869		-0.2705
Information Available	-0.2488	-0.2826	0.4553	0.2540		0.2300
System Feedback	-0.0949	-0.4303	0.0038	-0.6758		0.7597
Workload	0.0887	0.2472	-0.3057	-0.1038		0.1522
Stress	0.2207	-0.3268	0.1225	0.4005		-0.1186

*Statistics for this model could not be computed owing to missing data from one crew

The sign of the β_{weight}, whether positive or negative, indicates the kind of effect the PSF has on performance. Since these models were used to predict the time at which a critical mitigation task was performed during a scenario, values of coefficients that are positive indicate that the PSFs exerted an effect that would result in the crew taking longer to perform the mitigation action. Conversely, a negative β_{weight} indicates that the PSF tends to reduce the amount of time taken to perform the important mitigation action. This is because the variable predicted by the multiple regression (with the β_{weight}) is time to perform the critical mitigation action. Positive values of a β_{weight} indicate a PSF that adds to the time to complete the action – a negative β_{weight} sign indicates that the PSF leads to reduced time to perform the action. The magnitude of the multiple correlation coefficient, together with information about β_{weight} can in this way be used to draw insights about the strength and nature of the relationship between PSFs and crew performance. In the case of the β_{weight} analyses, these results show a method for discriminating the magnitude and direction (i.e, whether positive or negative) of a set of PSFs effects on operator and crew performance.

Figure 5 provides a graphical summary of some of the data in Table 2. Data in Figure 5 are taken from two reference case scenarios (the steam generator tube rupture and total loss of feedwater scenarios), from one conventional, and from one advanced plant setting. Figure 5 reveals several insights about the nature of the relationships between the PSFs and performance. Procedures added to crew response time in both scenarios and plants. This is observable as positive β_{weight} loadings in Table 2. This could be reflective of the time burden that accompanies systematic use of procedures by crews through process and function monitoring, immediate action verification, and subsequent step-by-step activities, even for very well-known events with clear event signatures – a potentially time-consuming process. It could also be indicative of crew interaction activities necessitated by procedure, such as communication, coordination, prioritization, goal setting, and others that, while not directly contributing to mitigation, are needed to develop goals, situational awareness, and workload management in order to balance the many competing demands that are required during mitigation and stabilization activities. Group processes are an integral part of control room activities, especially those involving task execution and coordination among control room personnel, though they may add to the time burden crews experience.

In contrast, information from control room systems and other plant sources was available and easily understood by operators and, as a result, this PSF had a positive influence on crew performance in the steam generator tube rupture scenarios. This is likely due to the saliency and obviousness of cues and symptoms that are available to operating crews in a SGTR event: the event signatures, as run in these studies, were apparent and unambiguous, as were the indications of which steam generator was faulted. Control actions were fairly straightforward in the early stages of response, though they became more complex during the progression of the event. In contrast, the initiating event leading up to the sustained loss of feedwater was perhaps more cognitively complex, owing to a simulated common-cause failure of the lubricating oil pumps, which supply main feedwater pumps with lubricating oil, due to a small leakage (20 kg/s) in the feedwater system. Out-of-service backup equipment and intermittent electrical failures due to the water leakage created an ambiguous picture of the availability and status of feedwater and auxiliary feedwater pumps to crews during transient mitigation. Furthermore, crews faced a situation near the end of the transient (as it was run in the simulator studies) in which they had to re-commission a dry steam generator with little procedural guidance. This may be one reason for the positive β_{weight} loadings for the Information Available PSF in the loss-of-feed-water scenario in Figure 1 – the positive β_{weight} loading reflecting a negative impact of the PSF (i.e., one that added to their time to complete the mitigation action).

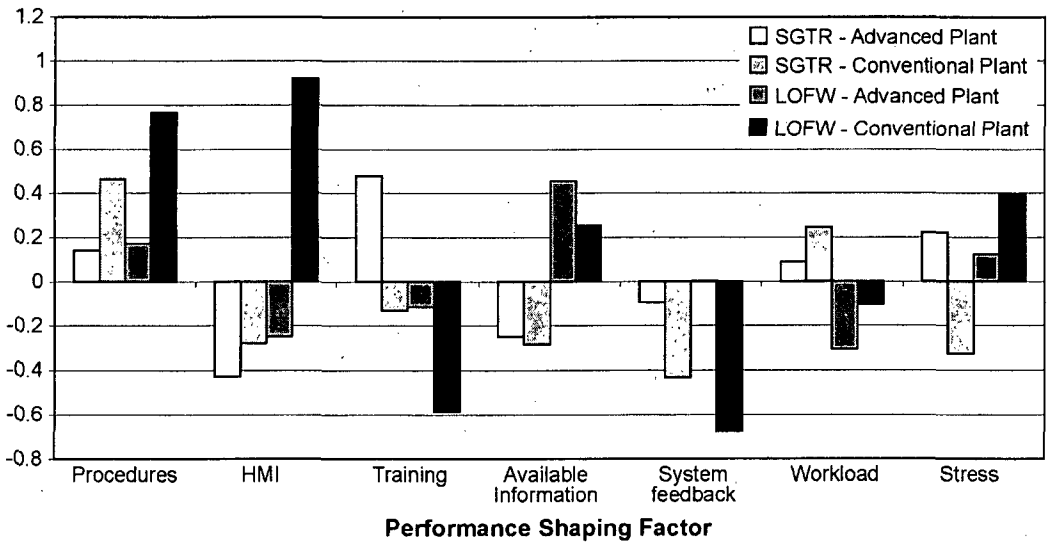

Figure 5 Beta Weights of PSFs in Conventional and Advanced Plant Settings.

The HMI in the loss-of-feed-water scenario in the conventional plant exhibited the greatest negative impact on operator performance of any PSF shown here. This is shown in Figure 5 as a positive loading of the β_{weight} for this PSF. In the view of the crews and as borne out through these analyses, the HMI aspects of the control room during this scenario detracted from their timely completion of control room mitigation activities. As discussed, the complications that stemmed from the ambiguity of the initiating event, equipment response, and vagueness regarding placing a steam generator back into service may have contributed to this,

These findings may account for the emergence of stress as a negative PSF in the loss-of-feed-water scenarios (as shown by the positive β_{weight} loading for the Stress PSF in the loss-of-feed-water scenario). When considered together, procedures, the human-machine interface, and information available interfered with crew mitigation in some ways. Collectively, such impacts may have led to conditions of stress that began to interfere with work and task completion.

Workload had a minor effect on crew performance across scenarios and plant settings, as shown by the relatively small β_{weight} loadings of this PSF. Clear differences in the direction of the β_{weight} of this PSF is observed: in the SGTR scenario, workload added to the mitigation time of crews, whereas in the LOFW scenario, it tended to reduce mitigation time. Workload, thus, had a slightly negative effect in the SGTR scenarios and a slightly positive effect in the LOFW event. Performance by the reactor operator (RO) and balance of plant (BOP) operator in the SGTR scenario are more tightly coupled and require greater coordination and interaction than in the LOFW scenarios which may have contributed, to the slightly negative workload effect. Perhaps more important than a perception of differences on workload between the two events is the observation that workload appears to have exhibited a relatively small influence on crew performance overall. This is not to say that workload was not relatively high, or that crews did not work under conditions of sustained workload during transient mitigation; rather, this indicates that the effects of workload would not appear, from these analyses, to have unduly influenced performance in a negative manner. Crews were successful in mitigating these transients, and so it is apparent that workload did not exceed crew capabilities or capacities.

Figure 6 shows the variation of Beta weights of the same PSFs across three scenarios from a different plant than those of Figure 5. These weights were produced from the analyses of plant

data from a single U.S. plant. The PSFs demonstrated predictive strength with regard to operator performance (i.e., transient mitigation time) and, similar to the results illustrated in Figure 5, demonstrate marked variability and individual predictive strength across contexts (i.e., scenarios)

Inspection of Figure 6 reveals several insights about the relationships between the PSFs and performance. In this plant setting, procedures also tended to add to crew response time in the same two scenarios as in Figure 5. Similarly, the HMI was perceived as mostly hindering performance in the total loss of feed water scenario in this conventional plant setting – similar to the same perception in the same scenario in the other referent conventional plant setting. Conversely, feedback from the system was perceived in both plant settings as mostly aiding control room activities.

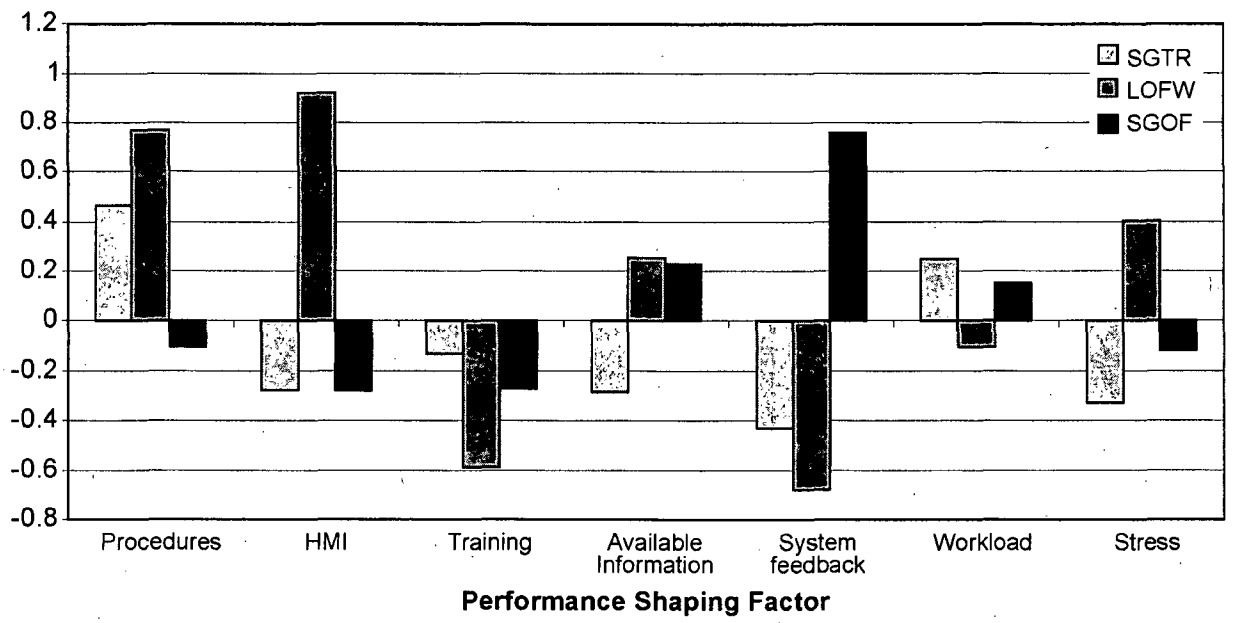

Figure 6 Beta Weights of Performance Shaping Factors.

Based upon these analyses, it seems possible to distinguish important performance-shaping factors from less important factors in contexts of interest to PRA and HRA. As the preceding discussion illustrates, the PSFs showed meaningful variation in simulated accident contexts, as evidenced by differences in both magnitude and direction (i.e., either positive or negative) of the β_{weight} of the PSFs that were included in these assessments. However, we have not attempted to define or suggest a criterion for what constitutes "important" as regards a PSF. Ultimately, a measure of sensitivity in the dependent variable and the variation produced in it by the individual PSF could be advanced to establish a context-specific measure of PSF importance. The β_{weight} of all PSFs provide a suitable starting point for comparing and contrasting differences in contributions from the PSFs.

2.2.5.3 Question 3: Can these data be used to relate the PSF influence to a general model of operator performance?

The analysis of Beta weights permits a general comparison of the effects of model parameters – PSFs – on performance. Through standardization, they are no longer in the same units as the

dependent measure when originally collected. To make predictions of performance under different hypothesized conditions (i.e., sensitivity analyses) requires non-standardized regression models. The b-weights of these models (as distinct from the Beta weights of the standardized regression model) are in the same units of measure as the original dependent measure and can be used to estimate predicted values of performance, using different input for model parameters. In the case of Human Reliability Analysis, it may be reasonable to question how response time (the measure employed in this study) could differ where conditions of the PSFs also differ. For example, just how much of a difference in performance may the quality of procedures create? Many HRAs employ a nominal value for most PSFs that represents the expected condition of PSFs in a specific plant. In most cases, the nominal value represents an assumption that procedures are well-prepared, that personnel are well-trained to employ them, and that they are technically well-suited for conditions to which they may be applied. As the results from the previous section show, operating crews may experience procedures differently than as assumed by the HRA. The resulting implication is that procedures, for example, do not facilitate performance in the manner or to the degree that we believe in all instances.

Previous research on the risk sensitivity to human error has employed an approach involving sensitivity analysis [Wong *et al.*, 1990]. In these studies, the PSFs were set to nominal, optimal, and worst-case conditions allowed by the HRA method employed in the PRA. The effect of variations in PSFs on HEPs and the resulting risk metric showed a clear functional response and sensitivity to PSFs as HRA model parameters. A difference between the sensitivity analyses performed here and those in the reported studies relates to the use of empirical rather than analytical model data.

Multiple regressions were performed on data from one of the conventional plants (i.e., the U.S. plant) using PSF data as independent variables and time to complete critical mitigation tasks as the dependent measure. The b-weights of the multiple regression models for three scenarios using the PSFs are presented in Table 3.

Table 3 b-weights of PSFs from a U.S. plant.

	LOCA	LOFW	SGOF
Procedures	0.74	1.0	1.27
HMI	0.64	0.56	-2.85
Training	-6.13	-1.1	-2.32
Available Information	1.47	-0.51	4.29
System Feedback	0.68	0.47	-2.42
Workload	1.52	-0.56	-1.96
Stress	-2.66	1.65	1.71

Using the b-weights from Table 3, three sets of sensitivity calculations were performed using the PSFs and the time used to mitigate each transient. The purpose of the sensitivity calculations is to demonstrate incremental changes in the predicted time taken to mitigate the operational transients that might occur were the quality of the PSFs to change. Here, the quality of the PSF is represented by the numeric value of the scale used to assess their perceived influence. Since the correlations between the multiple regression model and crew performance (and, hence predictive ability) between the PSFs and the time taken to mitigate the transients are greatest at the scenario-specific level, the sensitivity calculations were conducted on a scenario-specific basis. Two sensitivity calculations were performed for each scenario. The first includes the case where PSFs are set to their most optimal (a value of 5 on the original scale); the second represents the case in which PSFs are assumed at their worst (i.e., a value of 1 on the scale).

The output from the regression model is a prediction of the expected time taken to mitigate the transient. The results of these calculations are presented in Table 4.

Table 4 Sensitivity Calculations of Transient Mitigation time Using Scenario-Specific Model Parameters.

Sensitivity	Scenario	Predicted Time ($\hat{Y}$)	
PSFs =5	LOCA	1.8 minutes	
PSFs=1	LOCA	16.8 minutes	
PSFs =5	LOFW	< 1 minutes	
PSFs=1	LOFW	16.5 minutes	
PSFs =5	SGOF	1.6 minutes	
PSFs=1	SGOF	10.8 minutes	

For the loss of feed water scenario, the regression equation predicts that, given optimal PSFs, the time taken to mitigate the transient may be less than one minute. For the same transient, assuming the worst-case conditions addressed by the model, the predicted time taken to mitigate the transient may be 16.5 minutes. This reflects an increase in the expected time taken to mitigate the transient by at least a factor of 16.

Several important observations are motivated by these results. The b-weight regression model illustrates the effect size of the PSFs in the individual scenarios studied. One result of this model development is a scenario-specific set of PSF weights that can be used to estimate sensitivity of an HRA parameter of interest (i.e., transient mitigation time). Moreover, the regression approach provides a model by which the individual PSFs can be aggregated that accounts for their individual influences and avoids the problem of 'double counting' (i.e., treating each PSF's influence as independent and equal to those of other PSFs in the estimation of human performance reliability). In the case of the particular plant from which data were collected, the sensitivity analysis predicts an operator response range or 'time window' that is within the time allowed for operator performance to prevent further system degradation.

Similarly, for the small break loss of coolant accident, the predicted time to mitigate the transient under optimal conditions is less than 2 minutes. Assuming worst-case PSF model conditions, the time to mitigate the same transient is predicted at nearly 17 minutes. This reflects an increase in the expected mitigation time by a factor of 9. The sensitivity calculation for the steam generator overfill scenario indicates that the predicted time to mitigate the transient assuming optimal PSFs is approximately 1.6 minutes. The corresponding time to mitigate the transient under poor performance conditions is approximately 10.8 minutes, an increase in the expected mitigation time by a factor of 6. This still provides operators with sufficient time to prevent water from entering the steam lines given that the performances of crews who participated in these studies are representative of other crews.

Referring to the original question of this section, it is possible to relate PSFs through a general model of operator performance (i.e., a quantitative model) to predict operator performance. This model is limited to the purpose for which it was originally designed – to illustrate numerical sensitivity of operator performance to PSF parameters. Many types of models could be desired. For example, some HRA methods include causal models of operator behavior that account for such factors as short-term memory, attention, biases in judgment and decision making, etc. The model demonstrated here is not "causal" in nature and is reflective of performance in the limited environment from which data were obtained. Nevertheless, predictive validity of the model for

specific applications (i.e., predicting response time) may be good and support specific applications.

2.2.5.4 Question 4: Can these data be used to identify and characterize systematic interactions of PSFs in PRA-relevant scenarios?

A further question addressed in this research concerns whether systematic relationships exist among the PSFs that may also reflect underlying processes. Identification of underlying processes and characterizing them would represent an advance in our ability to treat PSFs as interdependent and dynamic – not static or deterministic model parameters. Factor analysis is a statistical analysis procedure that has been used by researchers to analyze patterns of relationships among individual variables to produce a smaller set of 'factors' that summarize the unique relationships among the variables and are capable of serving as composite measures of the variables. In the case of the analysis performed using PSF data, the goal was to ascertain whether relationships among factors exist and are reducible to a stable set of factors through which their effects can be uniquely expressed on operator and crew performance. Thus, rather than assessing the seven PSFs separately and treating their influences as independent of one another (an approach that is already contentious), factor analysis may be used to identify a factor structure employing fewer PSFs that are tractable, predictive, and easier or more efficient to assess during analysis than the original factor set.

A further goal of these analyses concerned the extent to which factor structures may replicate across plant settings. Replication is an important aspect of scientific research, and its value in the present study concerns the stability of factor structures across scenarios and plants. In performing these analyses, a distinction is drawn between these analyses and those previously reported in this paper. The goal of analyses using PSFs as model parameters is to evaluate their suitability in predicting operator performance, the sensitivity of performance to a predictor set of PSFs, and for assessing the relative contributions of these PSFs to operator and crew performance. The goal of these factor analyses is to identify systematic relationships that exist among PSFs, to evaluate the extent to which the variability in PSFs can be explained through an emergent factor structure, and to assess the stability of these relationships across settings.

Factor analysis is typically conducted to either test theory or to identify relationships. In the case of theory testing, an expected factor solution would typically be hypothesized prior to factor extraction, and analyses would be conducted to confirm or inform us about the suitability of the hypotheses. The analyses performed here, like those previously employing regression models to predict performance, are exploratory in nature and are intended to illustrate a candidate methodology to identify relationships among PSFs. Similar concerns exist relating to the size of the sample and the reliability of correlation results. This extends to procedures, such as factor analysis, that are based on correlations among variables. Comrey and Lee [1992] recommend having 300 cases or more for factor analysis. This is especially important when the results of analyses are intended for use in measuring psychological attributes and for making decisions about clinical placement or medical treatment (e.g., selecting candidates for employment in sensitive positions, treatment for a medical or psychological condition, etc.). The current analyses were performed and are reported for purposes of demonstrating approaches to improving the prediction of operator performance to support human reliability analysis. While they may exhibit less potential to affect decisions of a personal or medical nature, the guidance of Comrey and Lee should be considered and the results of these analyses interpreted with caution.

The PSF ratings obtained from crews were analyzed using factor analysis.[1] A goal of factor analysis is to obtain variables that load on (i.e., correlate with) a single factor and that make sense (i.e., assist in explaining causal relationships among variables). Factor analysis aims to reproduce the original correlation matrix among variables with a smaller set of orthogonal factors, based on analyses of covariance between variables (i.e., their communality). A factor is interpreted from the variables that have high loadings on it. In this way, factors are emergent from the correlation structure (i.e., in exploratory types of analysis such as this). Table 5 shows the results of factor analyses of the PSF ratings. The first column of the table lists the PSFs that were analyzed. The second column of the table (i.e., under Plant 1, Factor 1 heading) displays the correlations (loadings) between each PSF in turn and the first factor extracted by factor analysis. The third column (i.e., Plant 1, Factor 2) shows loadings of individual PSFs in turn and the second factor that was extracted, etc. Significant factor loadings are bolded in Table 5.

Table 5 Factors and Factor Loadings of PSFs.

	Plant 1		Plant 2			(Advanced) Plant 3		
PSF	Factor 1	Factor 2	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3
Procedures	**0.5685**	0.2175	**0.7885**	-0.342	-0.0496	**0.8799**	0.1639	0.0972
HMI	**0.7096**	0.1187	-0.266	-0.0538	**0.7785**	**0.6079**	-0.3686	0.1341
Training	**0.6662**	0.3189	**0.8545**	0.1299	0.0936	**0.5922**	-0.3378	-0.4733
Information Available	**0.6793**	-0.1164	0.4109	-0.0192	**0.6583**	-0.2828	**0.8521**	-0.1972
System Feedback	**0.7642**	0.1193	0.3772	0.216	**0.4431**	0.0586	**0.9109**	-0.0177
Workload	-0.1484	**-0.7816**	-0.2737	**0.8733**	-0.0508	0.1005	0.0064	**0.8089**
Stress	-0.1703	**-0.8385**	0.3339	**0.7865**	-0.193	0.0165	-0.4188	**0.7336**
Variance Explained	0.34	0.22	*0.61			0.30	0.21	0.21

* This value represents the total variance accounted for by all three factors extracted.

Figure 7 shows the factor loadings of the PSFs on the two factors extracted in plant 1 (a conventional U.S. plant). The figure shows two distinct factors. The first factor relates to the systems and work processes in the workplace that have been designed to assist operating crews in accomplishing their work. These include procedures, the human-machine interface, training, the information available to operating crews, and system feedback. A second factor that is comprised of workload and stress was also extracted. This factor may represent the perceived *demand* that the scenarios placed on crew members. The directional relationship

[1] This involves analysis of the inter-PSF correlations to extract factors. Following extraction of the factors, additional procedures were performed to maximize the variance accounted for by each factor, to reduce shared variance among factors, and to maximize the individual correlations between the variables in the analysis (the PSFs) and the resulting factors. Factor rotation was performed to maximize high correlations or factor loadings between individual items and each factor, and to minimize low correlations. This factor rotation technique, *varimax* rotation, also maximizes variance accounted for by individual factors, and produces factors, after rotation, that are orthogonal to one another [see Tabichnick and Fidell, 2001]. This aids in interpretation since the variance that each factor accounts for in the variable set is unique, and shared variance is avoided. It also means that the unique solution of factors that is obtained in each factor analysis is additive in its explanation of the total variance accounted for in the variable set. A factor loading of each variable on each factor (a correlation actually) is produced that assists in explaining the factors that emerge from the variable structure.

between these factors is also apparent from the factor loadings. In both factor structures, operators' higher ratings of PSFs that support performance (e.g., Procedures, HMI, etc.) are accompanied by lower ratings of PSFs that load on the *demand* factor, and vice versa.

The factors extracted and factor loadings from Plant 2 are shown graphically in Figure 8. Factors in Plant 2 (a conventional non-U.S. plant) show greater specificity in the factor loadings than in Plant 1. Differentiation of a third factor in Plant 2 shows a slightly different, though conceptually compatible, factor structure than extracted from the PSF data from Plant 1. The first factor extracted is composed of procedures and training. This factor probably represents *preparedness* – that is, how well-suited procedures and training were to the demands of the scenarios in enabling operators to effectively mitigate the event. A second factor was extracted that probably represents the *human-system interfaces* and elements – that is how well the instrumentation, control, layout, information available, and system feedback to the operating crew supported their mitigation and control activities. A third factor was extracted that represents the *demand* of the transients on crew performance. This factor is identical to that extracted in Plant 1.

Figure 9 shows the factors and factor loadings of PSFs in Plant 3. Similar to plant 2, three factors were extracted in this plant. Two differences are observed between the factor structures in plant 3 and the others. The first factor extracted relates to procedures and training. This factor includes not only how well-prepared the crews were to manage the event through the design of procedures and its training program, but also the quality of instrumentation and control-room layout that supported their efforts in achieving and executing control activities. The second factor is comprised of the information available and the quality of system feedback. Together these represent the systems that provide support to cognitive *information processing* activities, such as detection, information gathering, monitoring, diagnosis, etc. A third factor, similar to the *demand* factor noted in the other plant settings, was also extracted from this data.

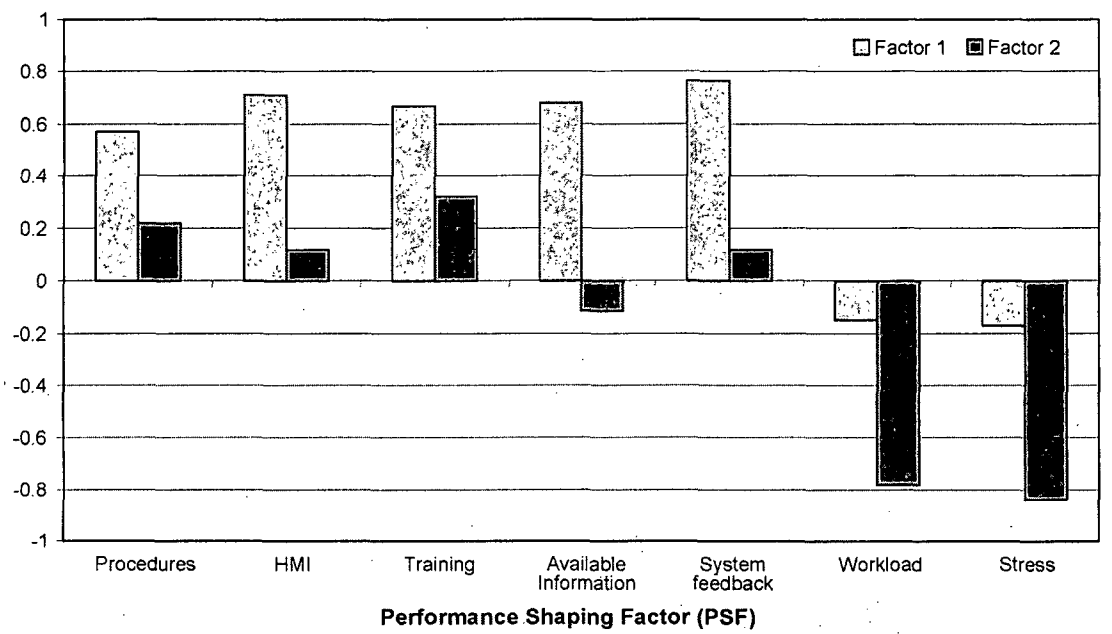

Figure 7 Factors and Factor Loadings of PSFs from Plant 1.

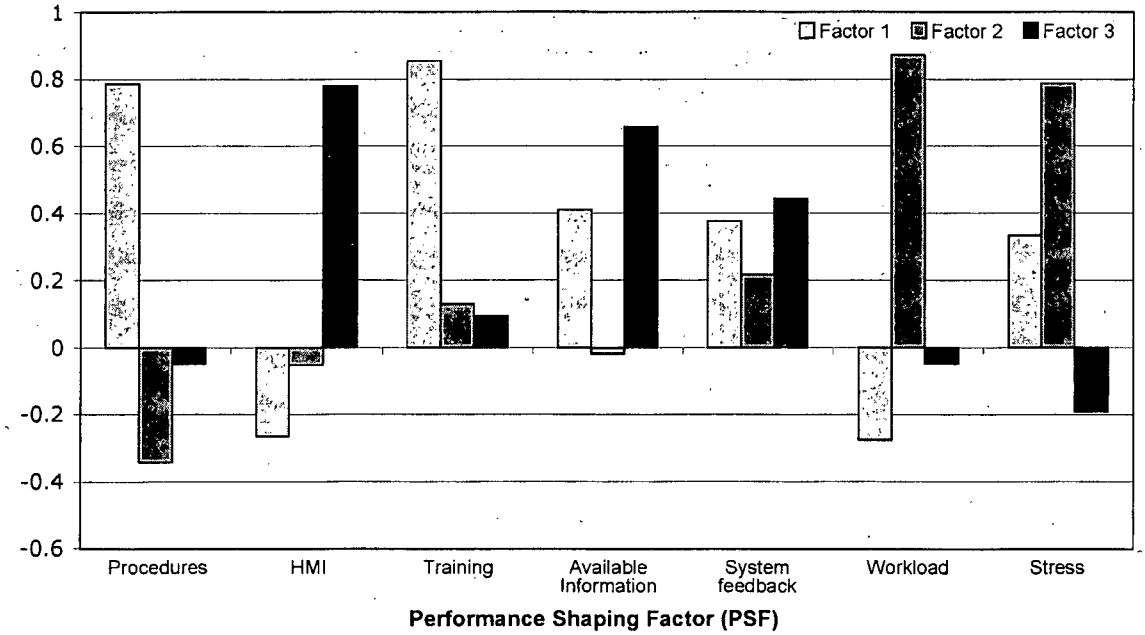

Figure 8 Factors and Factor Loadings of PSFs from Plant 2.

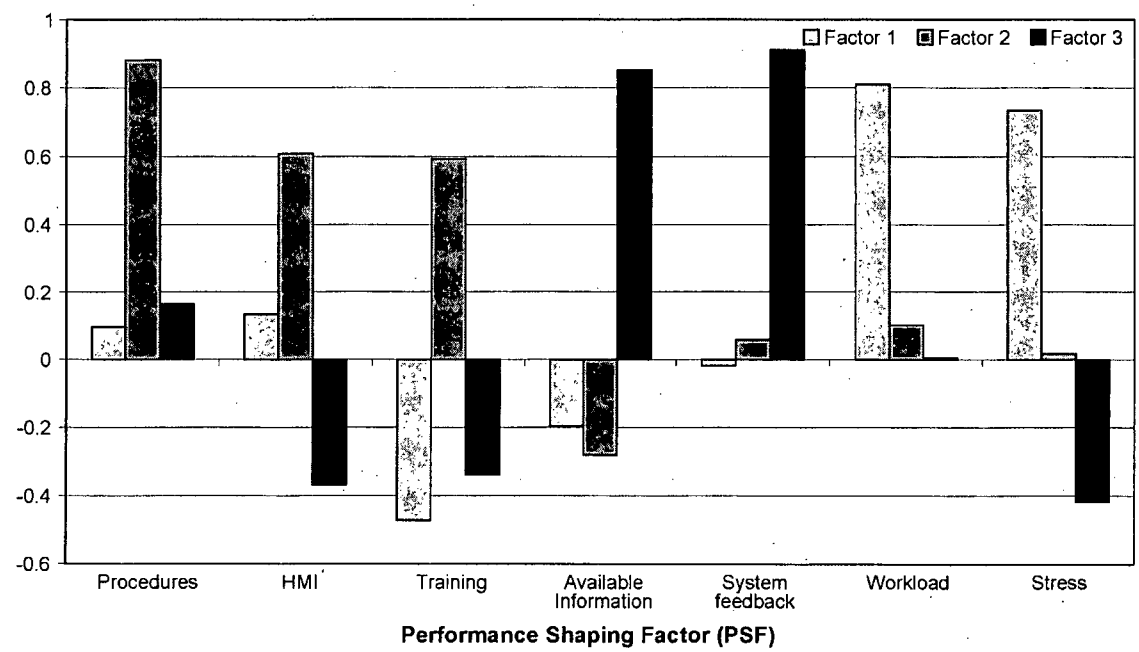

Figure 9 Factors and Factor Loadings from Plant 3.

Some similarities in factors and factor loadings are observed in these results, as well as some differences. In all plants, workload and stress were extracted as a single factor termed *demand* for the mental and physical loads they place on operating crew members. This demand factor always had the same unique solution; these two variables loaded on this factor and not on other factors. The implication of this factor is clear; crews experience workload and stress as a collective influence on their capacities and abilities that typically functions as a complement set to the other PSFs which are designed to support operator performance. As discussed, operating crews in these studies were exposed to long duration (i.e., 1-2 hrs.) design-basis events that included realistic role-playing of ex-control room and off-site personnel activities. They

performed during conditions of high workload transition following the initiating event, and sustained workload thereafter as they restored and stabilized plant conditions. Workload and stress were experienced by operating crews in a similar manner that is evidenced by PSF correlations and relatively stable factor extractions across different plants.

Differences were observed in the other factors extracted across plants. In the case of the first plant, the remaining five PSFs (after accounting for workload and stress) comprised a single factor, with similarly high factor loadings. In the second plant, operators discriminated between the effects of PSFs in supporting their preparedness to manage the challenges of the scenario, and the instrumentation and ergonomics aspects of process control. In the third plant setting, procedures, training, and the HMI were viewed as more related to one another in how they supported crew activities. The more information-related aspects of the human-system interface were perceived distinctly and separately from these other two factors. Some of the differences observed in factors that were extracted may relate to subtle plant differences – especially in the case of differences observed between plants 1 and 3, which involved operating crews from the same plant performing in their own simulated control room and in an advanced control room simulator, which had greater automation and passive system functions. The differences may also relate to subtle variances in the way that the scenarios played out in different settings, given the small differences that were present in the timing and thermal-hydraulic response of the simulated plants. Given the relatively small sample sizes employed and the limited number of scenarios that were involved in the studies, too much significance should not be attached to small variations in factor composition, especially given the similarities, overall, in factor structures across plant settings.

A final purpose of the factor analyses was to ascertain whether the factors extracted account for an appreciable amount of covariance among the PSFs – that is, their correlations. The last row of Table 5 displays the amount of variance accounted for in the PSFs by the individual factors. In the case of the first plant, 56% of variability in the data is accounted for by the two factors, the rest being residual or unexplained variance. In the second plant, the factor structures accounted for 61% of variability, and in the third plant, 72% of the variability is accounted for by the factor structures. Collectively, we can say that the majority of covariance among PSFs is accounted for by the emergent factors that we have discussed here. The factor analysis results thus illustrate meaningful interactions among PSFs, some stability and predictability in the kinds of factors that emerge, and a potentially more tractable and systematic way of combining individual PSFs to account for their influence on operator performance.

2.2.6 Summary

This study demonstrates a methodology for collection and application of human performance data from simulator settings and has discussed some of its potential uses to support human performance modeling and HRA method development or improvement. The data from such studies are potentially valuable but require careful treatment to advance their use in HRA. Attempts to fit data to broad classes of reliability model-type curves without regard to the performance determinants (i.e., PSFs and context) was a factor that eventually led to the disuse of some time-reliability HRA methods. The data from this study were employed to better understand the factors that influence operator behavior in PRA-relevant contexts and to relate them to meaningful, objective measures of operator performance. Results from this study indicate that PSFs and context are inter-dependent; the extent to which a particular PSF or latent factor influences operator behavior is dependent on and conditional upon the particular context. These results also argue for the use of formalisms that account for the multivariate nature of context when using data from such complex performance domains for estimating the reliability of specific human actions.

The study also identified a means of addressing specific questions that are important to understanding and predicting human performance and its determinants in contexts of interest to PRA. Recasting the questions posed earlier in light of the findings, we summarize some of the important insights.

1. Performance-Shaping Factors appear to be predictive of some important aspects of operator performance.

The high multiple correlation coefficients obtained by using the PSF data as a predictor of crew response time to take important mitigation actions showed a consistently high degree of explanatory power. In most behavioral science studies, to be able to account for 25% of human performance variability in dependent measures is considered strong evidentiary support for a predictive relationship between the independent and dependent measures. In these studies, the majority of crew variability was accounted for by the predicted relationships between PSFs and operator performance in most of the scenarios. This means that the majority of crew variability in mitigating the transients was accounted for by the combination of PSFs measured in these studies. Given the many other sources of individual and crew variability that one can imagine, this is significant and seems to confirm what human reliability analysts have hypothesized for some time, that PSFs are predictive of human performance and, thus, human reliability in PRA contexts.

2. The importance of individual PSFs varies across contexts.

Considering the Beta weights of individual PSFs, a good deal of variability in the importance of individual PSFs was observed across scenarios and across plants. This was evidenced in the value and sign of the Beta weight of PSFs. Some PSFs seemed to add to the time crews took to mitigate transients, while the same PSFs appear to have facilitated the timely completion of mitigation actions in others. In addition, there was marked variability in both the sign and value of individuals across and within individual scenarios. Post-scenario analyses and crew debriefings seem to bear out this point and provide specific reasons for the finding. However, this would seem to argue for the need to conduct scenario-specific assessments of performance shaping factors and the need to ascertain realistic information from subject matter experts about the effects of PSFs rather than using static look-up tables and expert judgment without the benefit of operational insight. This issue has been much debated within the HRA community, and data from this study would argue for context-sensitive assessments of PSFs, as well as for a thorough description of the many variations of PRA scenarios as they produce differing performance contexts for operating crews. That is to say that steam generator tube rupture and other PRA-relevant event sequences can have different initiating events, vary in terms of their severity, complexity, and other factors known to produce stress and affect the way crews use human and plant resources. Such differences may provide evidence as PSFs and contextual factors, and their characterization in PRA should be thorough and accurately reflect the nature of their influence on crews.

3. A general model of operator performance can be derived that accounts for the influence of PSFs and operator performance.

The analysis of individual PSF impact on operator performance was able to account for important aspects of crew variability in performing critical mitigation actions. Moreover, the linear model employed showed sensitivity to differences in PSF strength and magnitude (i.e., facilitating or impeding performance) across transients. As a consequence, we can say that the general model was able to distinguish between the relative importance of PSFs and the kinds of effects they have on performance, in general. The results prompt an investigation of what

specific features influenced operator performance and why. In some scenarios, for example, procedures were viewed by crews as facilitating their mitigation of the event. But we don't know, from this data, which aspects of procedures and their use were important to crews. Similarly, the data do not provide insights into what conditions lead procedures, in the view of crews, to impede performance in performing a specific mitigation activity. Such data would not be difficult to obtain in studies such as these, but they were not a part of the scope of this research. Nonetheless, the human reliability and PRA communities have debated the utility of employing the "HEP cum PSF" approach to human reliability estimation – either because empirical evidence is lacking about the influence of PSFs in PRA-relevant contexts, or because of observed variability in HRA outcomes using such approaches. The present research is intended to provide an illustration of methods that may be helpful in reducing uncertainty about employing PSFs and how to gather data on their influence on operator behavior in PRA relevant contexts.

4. The data obtained in these studies do show that some systematic interactions occur among PSFs in the simulated PRA contexts.

An important aspect of model development and model validation is to understand the relationships between model parameters – especially systematic interactions. These analyses showed a number of systematic interactions among the PSFs across operational contexts. Firstly, stress and workload demonstrated strong inter-PSF correlation and common factor loadings across all of the simulated contexts. This degree of predictability and replication of the underlying construct demonstrates stability of the measurement and some generalizability of the methodology to assess PSF interactions. Beyond that, some of the PSFs tended to load together on factors across operational contexts. *Procedures* and *training*, for example, loaded on the same factor across contexts, as did *information available* (through the human system interface) and *system feedback*. Such predictability is desirable as it shows that the operators' views of them are consistent, and provides some additional insight into patterns of emergent interactions among PSFs. This information may be valuable for ascertaining a minimum set of PSFs that may be operant in certain PRA-relevant contexts, as well as point out ways of reducing unnecessary factors for inclusion in HRA models.

Most HRA methods in use today direct analysts to account for the effects of context, plant conditions, and performance-shaping factors that may influence the likelihood of human performance errors and unsafe acts. These methods each provide differing guidance on how to assess the influence of these factors and how to incorporate their influence in the final estimate(s) of human reliability. The uncertainties accompanying estimates of human reliability produced exert an effect on the overall uncertainty of results in PRAs. Better models of human performance and data are needed that are capable of predicting qualitatively accurate results (i.e., the kinds of human errors that should be included in PRAs, etc.) and producing quantitative estimates of human reliability that are sufficiently accurate while reducing or better characterizing their uncertainty.

This study demonstrated a methodology for characterizing the relationship among a candidate set of PSFs and operator performance. Results of the research (considering the limited amount of data) show that the predictive strength of PSFs is best at the individual plant and scenario level. This demonstrates plant- as well as scenario-specific relationships between PSFs and operator performance. HRA practitioners routinely address the question of generalizing HEP estimates from one context to another (i.e., whether or not to use the results of analyses of human performance from one event sequence in another event sequence that is in some ways different). At the very least, these results would support a thorough reassessment of PSFs under qualitatively different performance conditions, especially when applying estimates from methods that do not distinguish between nominal failure rates and context.

The results also motivate a reconsideration of the way to best assess the influence of PSFs. As discussed earlier, the current prevailing hypothesis underlying some HRA methods is that fixed multipliers can account for the effects of a PSF in a range of situations. While this may be acceptable for purposes of screening some human actions, it is not reflective of these data nor of more recent advances in human reliability assessment (see Barriere *et al.*, 2000 for a more thorough treatment of this issue). Although preliminary, these data argue against generalizing PSF effects across scenarios, plant settings, or plants.

Several limitations in this study warrant recognition. Firstly, as previously discussed, the ratio of cases to independent variables in regression analyses and factor analyses did not meet the minimum that has been recommended in psychometric research. No intent is intended to imply a general model of performance for use as a PRA tool in these results. As importantly, all of the analyses have been performed using similar dependent performance measures – that of transient mitigation time. The time measure was selected as a single, objective outcome variable that is reflective of crew decision making, progress through the mitigation plan, etc. Time, as a measure, is also important from a thermal hydraulic standpoint because in all of these scenarios, plant conditions tend to worsen as time progresses, without operator action. While this is a useful measure and has good face validity for purposes of demonstrating a relationship between performance and PSFs, we must recognize that other aspects of performance may be at least equally important.

The methods and results presented here are intended to demonstrate a means for better assessing the effects of PSFs for use in PRA, and for developing improvements in model parameters to reduce the uncertainty in HRA. They demonstrate the systematic variability in performance that can be accounted for through explicitly modeling the effects of contextual factors on performance, and indicate the types of systematic variation in model parameters that occur during simulated accident conditions. Approaches such as this, when included as a part of data collection for event sequence modeling in PRA, may be used to improve the identification of relevant contextual factors, assessing their effects and improving the accuracy of model parameters employed in HRA methods. Together with quantitative methods that can make use of different forms of evidence, such sources of empirical information can be used to improve the accuracy of estimated human failure events and the uncertainty associated with such events.

To refine the methodology and results presented here, systematic efforts are needed to extend the amount of available data in hopes of determining whether these results and trends hold up under different conditions. For example, the NRC participates in the Halden Reactor Project research program and collaborates on research in human reliability already. Some attention may be given within that collaboration to continue this line of research to (1) determine to what extent PSFs are predictive of operator performance in different contexts; (2) extend and refine the methodologies for data collection, analysis, and human reliability model development; (3) explore modeling techniques such as structural equation modeling, latent variable methods (i.e., factor analysis) and others to gain insights into subtle and systematic interactions between contextual elements; and (4) increase the potential sources of data that can be included within HERA, thereby increasing its usefulness as a tool to support and improve human reliability analysis activities. Furthermore, additional attention needs to be given to how to use the quantitative results of such analysis to modify and improve existing HRA methods and quantification efforts. For example, can some sources of analyst judgment be reduced by benchmarking PSFs in different contexts to determine which may be most relevant and how much of an impact on performance reliability they have? How can these kinds of studies simplify approaches to human reliability assessment? Ultimately, the utility of this approach rests with the goals of improving the accuracy of predictions based upon more systematic use of information related to PSFs and context, reducing the uncertainty currently associated with

human reliability analysis, and providing a simplified process for assessing a relevant set of performance shaping factors that are predictive of human reliability in the performance domain(s) of interest.

2.3 Bayesian Updating of PSF Effects and HRA Estimates

Prepared by Sankaran Mahadevan

2.3.1 Introduction

This presentation develops the application of the Bayesian methodology to update HEP estimates and PSF statistics based on empirical data. The SPAR-H model [Gertman, *et al.*, 2005] is used for illustration, and sample information from the HERA database is used. The Bayesian analysis makes use of prior estimates on HEPs, human performance data, and PSF occurrence data to compute updated HEPs, and updated probability mass functions (PMFs) for the performance shaping factors. The methodology includes cases when data is available on PSF occurrences, as well as human error rates, the output state is either binomial or multinomial, correlation exists between trials, and correlations exist between PSFs. A numerical example is presented in which the eight PSFs in the SPAR-H model were updated using sample data from the HERA database, based on typical nuclear power plant incident reports.

The Standardized Plant Analysis Risk – Human Reliability Analysis (SPAR-H) method developed by Idaho National Laboratory (INL) makes use of eight performance shaping factors (PSFs). These are treated as discrete variables, each having several possible multiplier values corresponding to different conditions. Thus, under normal conditions, the multiplier value for a given PSF will be equal to one, and under aggravating conditions, the multiplier will be greater than one. The HEP is computed as the product of a baseline HEP, p_0, and the multiplier values for each PSF.

Thus, the probability of human error is expressed as

$$p = p_0 \prod_{i=1}^{8} F_i \quad , \qquad \text{(Eq.13)}$$

where p is the probability of a human error occurring given the performance conditions described by the eight PSFs, and the multiplier value of the ith PSF is represented by the variable F_i. Each of the multipliers F_i is a discrete variable with possible values $\{f_{i1}, f_{i2},...f_{ij}\}$ and corresponding probability mass function (PMF) $p_{Fi}(f_i)$. Limited knowledge of the actual shape of the PSF distributions has been combined with expert judgment to estimate the prior PMFs for each of the PSFs. The factors are listed in Table 6, along with the possible multiplier values for each, and the corresponding assumed prior PMFs.

Table 6 PSF multipliers and assumed prior probabilities.

PSF	Description	Multiplier	Prior PMF
F_1 Available Time	Time available to complete task	10	0.159
		1	0.683
		0.1	0.136
		0.01	0.023
F_2 Stress	Stress level of operator	1	0.841
		2	0.136
		5	0.023
F_3 Complexity	Complexity of task	1	0.500
		2	0.341
		5	0.159
F_4 Experience/Training	Operator's experience level	0.5	0.333
		1	0.333
		3	0.333
F_5 Procedures	Existence and clarity of documented procedures	1	0.450
		5	0.300
		20	0.200
		50	0.050
F_6 Ergonomics	Interaction between operator and equipment	0.5	0.159
		1	0.683
		10	0.136
		50	0.023
F_7 Fitness for Duty	Mental and physical state	1	0.841
		5	0.159
F_8 Work Processes	Organizational factors	0.5	0.159
		1	0.819
		5	0.023

In Table 6, "available time" describes the extent to which the operator has ample time to perform the task. "Stress" is broadly defined to account for negative motivating forces influencing the worker, such as mental stress, excessive workload, or physical stress. "Complexity" refers to the relative difficulty of the task at hand. "Experience/Training" accounts for years of experience, whether or not the operator has been trained for the specific task, and the amount of time since training. "Procedures" describes whether or not appropriate documentation exists outlining the proposed task, and whether or not such documentation is clear and correct. "Ergonomics" refers to the quality and ease of use of the instrumentation as well as the interaction between the operator and the necessary equipment. "Fitness for Duty" can be both mental and physical, and it includes factors such as fatigue, sickness, drug use, and other personal problems. "Work Processes" refers to organizational factors such as the safety culture, communication, and management policies. For a more detailed description of the performance shaping factors, refer to [Swain and Guttmann, 1983].

Since the SPAR-H human error probability model is multiplicative, it is possible for certain combinations of PSFs to result in probabilities that are greater than one. Thus, the model makes use of a correction factor which is applied as follows:

$$p = \frac{p_o \prod F_i}{p_o(\prod F_i - 1) + 1}, \qquad \text{(Eq. 14)}$$

whenever three or more PSFs have multiplier values greater than one. Although not outlined in the SPAR-H model, it has been observed by the authors that the conditions for applying the correction factor did not cover all cases where probabilities greater than one could occur. For example, only two PSFs, $F_5 = 50$ and $F_6 = 50$ could result in a probability greater than one if all other PSFs take a value of 1 (with $p_0 = 0.001$). Thus, for the analysis that follows, the correction factor will be applied whenever probabilities greater than one would result, or when three or more multipliers are greater than one.

2.3.2 Proposed Methodology

This subsection outlines the methodology for using Bayes' theorem to update the PSF distributions (and hence the probability distribution for the HEPs predicted by the SPAR-H model) when human reliability data is available. Four cases are considered, and a numerical example for each case is provided for illustrating purposes, after presenting the methods.

1. Empirical data gives the number of errors observed along with the number of opportunities.

2. Data is also available on the frequency of occurrence of the performance shaping factors.

3. Correlation is believed to exist between trial outcomes.

4. Correlation is believed to exist among the occurrences of certain performance shaping factors.

Case 1: Data Available on Binomial Outcome

Consider that data is available which tells us that k errors were observed out of n trials. We can use Bayes' theorem to obtain the posterior PMF of each performance factor as

$$P(f_1, f_2, \ldots, f_8 \mid k, n) = \frac{P(k \mid n, p) P(f_1, f_2, \ldots, f_8)}{\sum_i \sum_j \cdots \sum_p P(k \mid n, p_r) P(f_i, f_j, \ldots, f_p)}, \qquad \text{(Eq. 15)}$$

where $f_1, f_2, \ldots, f_8$ are the particular multiplier values being updated; p is calculated from (Eq. 13) using the values $f_1, f_2, \ldots, f_8$; and p_r is calculated from (Eq. 13) using the values $f_i, f_j, \ldots, f_p$.. The summation in the denominator is taken over all possible multiplier values for each of the eight PSFs. The first term in the numerator on the right hand side is the likelihood function, and the second is the prior PMF. Since the denominator term is a constant and the posterior PMFs can be normalized so that their sum is one (Eq. 15) can be reduced to

$$P(f_1, f_2, \ldots, f_8 \mid k, n) \propto P(k \mid n, p) P(f_1, f_2, \ldots, f_8). \qquad \text{(Eq. 16)}$$

The marginal distributions for each factor can be computed by summing over all possible values for the other factors.

If we assume that the trials are independent with constant error probability, then the number of human errors is a binomial random variable. The likelihood term is thus given by

$$P(k \mid n, p) = \binom{n}{k} p^k (1-p)^{n-k}. \qquad \text{(Eq. 17)}$$

We can drop the $\binom{n}{k}$ term when using (Eq. 16) since it becomes part of the normalizing constant in future calculations. In addition, if we assume that the PSFs are independent of each other, then we can use the product rule to write

$$P(f_1, f_2, \ldots, f_8) = \prod_{i=1}^{8} P(f_i) \quad \text{(Eq. 18)}$$

With these simplifications, the posterior PMFs of the PSFs can be obtained as

$$P(f_1, f_2, \ldots, f_8 \mid k, n) \propto p^k (1-p)^{n-k} \prod_{i=1}^{8} P(f_i). \quad \text{(Eq. 19)}$$

<u>*Case 2: Data on both Input and Output*</u>

Consider the case where, in addition to the human error data, we have data on the occurrences of the various performance conditions, or PSFs, such as in nuclear power plant incident reports. In that case, the HRA model can be updated in a two-step procedure, first using the data on the PSF occurrences, and then using the human error data, k and n.

Let a particular factor (say stress, or available time) have l possible states with respective PMF values p_1, p_2, ..., p_l such that $\sum_{i=1}^{l} p_i = 1$. Consider that under n opportunities, each state i occurred k_i times, with $\sum_{i=1}^{l} k_i = n$. The probability of obtaining any i^{th} state k_i times is $p_i^{k_i}$ and the number of ways in which ($k_1 + k_2 + \ldots + k_l$) opportunities can be divided into $k_1, k_2, \ldots, k_l$ groups is $\frac{(k_1 + k_2 + \ldots + k_l)!}{k_1! k_2! \ldots k_l!}$. The logic behind this is that whenever we choose k_1 groups out of n opportunities, there will be (n-k_1) opportunities left to choose k_2 from, (n - k_1 - k_2) opportunities left to choose k_3 from, and so on. Thus the likelihood of observing the data, or the joint distribution of $k_1, k_2, \ldots, k_l$ can be derived as

$$f(k_1, k_2, \ldots, k_l \mid p_1, p_2, \ldots, p_l) = \left[\frac{\left(\sum_{i=1}^{l} k_i\right)!}{\prod_{i=1}^{l} (k_i)!} \right] \prod_{j=1}^{l} p_j^{k_j} \quad \text{(Eq. 20)}$$

The above expression is also the well-known multinomial distribution [see, for example, Johnson, *et al.*, 1997]. If the subjective prior joint PDF of probability masses $f(p_1, p_2, \ldots, p_l)$ is known, one can use the likelihood function to obtain the posterior joint PDF, $f(p_1, p_2, \ldots, p_l \mid k_1, k_2, \ldots, k_l)$.

Further, the choice of a suitable conjugate prior will yield a posterior density of the same form as the prior. This eliminates the computation of complicated integrations needed during Bayesian updating [Jeffreys, 1961]. When the likelihood is multinomial, a Dirichlet prior will result in a Dirichlet posterior, and thus has been the undisputed choice throughout the Bayesian statistics

literature [see Jeffreys, 1961; and Leonard and Hsu, 1999] for a more theoretical analysis. Thus, the probability masses $p_1, p_2, \ldots, p_l$ are assumed to follow the Dirichlet distribution with parameters $\alpha_{i\,1}, \alpha_{i\,2}, \ldots, \alpha_{i\,l}$, where the joint PDF is given by

$$f(p_1, p_2, \ldots, p_l) = \left[\frac{\Gamma\left(\sum_{i=1}^{l} \alpha_i\right)}{\prod_{i=1}^{l} \Gamma(\alpha_i)} \right] \prod_{j=1}^{l} p_j^{\alpha_j - 1} . \qquad \text{(Eq. 21)}$$

The mean and variance of the distribution are given by

$$E(p_i) = \frac{\alpha_i}{\sum_{j=1}^{l} \alpha_j} \qquad \text{(Eq. 22)}$$

and

$$Var(p_i) = \frac{\alpha_i \left(\left[\sum_{j=1}^{l} \alpha_j \right] - \alpha_i \right)}{\left(\sum_{j=1}^{l} \alpha_j \right)^2 \left(\left[\sum_{j=1}^{l} \alpha_j \right] + 1 \right)} . \qquad \text{(Eq. 23)}$$

From (Eq. 22) α_i is given by the prior estimate of the probability of state *i*.

By combining the data and prior using the Bayes theorem, we find that the posterior density of each of the PMF values also follows a Dirichlet distribution with parameters $\alpha_i + k_i$ (see, for example, [Johnson, *et al.*, 1997 and Leonard and Hsu 1999]. Again, from Eq. (10), we find that the expected value for the posterior PMF of the ith performance factor will be given by

$$P(f_i \mid k_1, k_2, \ldots, k_l) = \frac{\alpha_i + k_i}{n+1} \qquad \text{(Eq. 24)}$$

We then update the PMFs a second time using (Eq. 19) with the human error data. The updated PMF values obtained from (Eq. 24) become the prior for this second step.

Case 3: Correlation between Trials

For this case, assume that the probability of error on a given trial can be influenced by whether or not errors occurred at previous opportunities. This is to say that, within some set of trials, correlation exists among the trial outcomes. For example, multiple errors could be related to the same incident or occur on the same day. This would be an example of a positive correlation between trials. Negative correlation is also possible, as in cases where operators could learn from previous incidents, or where errors could result in increased awareness or caution. In these cases the correlation would be negative because the occurrence of an error would make it less likely for an error to occur again in future opportunities.

For this case we still consider binomial outcome (failure or success); however, we will allow that the usual assumptions associated with the binomial distribution may be violated. These assumptions are as follows:

1. The trials are independent

2. The probability of success is constant for all trials.

Thus, a generalized binomial distribution is needed. Several such distributions have been developed. Drezner and Farnum [1993] explored a generalized binomial which uses a recursive relationship to discard the assumption of independence between trials. Altham [1978] proposed two binomial generalizations which were based on "multiplicative" and "additive" definitions of interaction between the discrete variables. A chi-square test was used to compare the three distributions using data showing a strong negative correlation. The multiplicative generalized binomial distribution was found to give the best fit for the available human reliability data, and it is applied for this case.

For the multiplicative generalized binomial, the PMF of the distribution of the number of successes, X, is given by

$$P(X=x)=\frac{\binom{n}{x}p^{x}(1-p)^{n-x}\theta^{x(n-x)}}{f(p,\theta,n)}, \qquad \text{(Eq. 25)}$$

where n is the number of trials, p and θ are the distribution parameters, and

$$f(p,\theta,n)=\sum_{j=0}^{n}\binom{n}{j}p^{j}(1-p)^{n-j}\theta^{j(n-j)}. \qquad \text{(Eq. 26)}$$

As in the case of the binomial distribution, the parameter p is related to the probability of success for a given trial. The parameter θ must be greater than zero and describes the amount of correlation among trials. Note that (Eq. 25) reduces to the binomial distribution when the correlation factor, θ, is equal to 1. Otherwise, $\theta < 1$ corresponds to positive correlation between trials and $\theta > 1$ corresponds to negative correlation. In effect, for positive correlation between trials, the variance of X will be greater than that predicted by the binomial distribution, and vice versa.

Because this is a two-parameter distribution, estimating the parameters will require that data be available up to the second moment. Whereas the probability of failure for the binomial distribution can be estimated with only one set of data (number of trials n and number of failures k), estimating both a probability and a correlation parameter will require multiple sets of data, in which the number of trials, n, must be the same for each set of data. The method of maximum likelihood can be used to estimate the parameters p and θ. The likelihood equations are derived below:

$$\frac{\partial \ln L}{\partial p}=\frac{\sum x_i}{p}-\frac{\sum(n-x_i)}{1-p}-\frac{n}{f}\frac{\partial f}{\partial p}, \qquad \text{(Eq. 27)}$$

$$\frac{\partial \ln L}{\partial \theta} = \frac{\sum x_i (n - x_i)}{\theta} - \frac{n}{f}\frac{\partial f}{\partial \theta}, \qquad \text{(Eq. 28)}$$

where

$$\frac{\partial f}{\partial p} = \sum_{j=0}^{n} \binom{n}{j} \theta^{j(n-j)} \left(j p^{j-1} - n p^{n-1} \right), \qquad \text{(Eq. 29)}$$

$$\frac{\partial f}{\partial \theta} = \sum_{j=0}^{n} \binom{n}{j} p^j (1-p)^{n-j} j(n-j) \theta^{j(n-j)-1}. \qquad \text{(Eq. 30)}$$

Altham [1978] gives slightly different expressions for (Eq. 27 – 30)). Note that to determine p and θ, (Eq. 27 and 28) must be solved iteratively.

Consider that z experiments were conducted, each having n trials, so that the number of successes for each experiment are given by k_1, k_2, ..., k_z. In the context of human reliability analysis, each experiment may correspond to a certain operational facility or time period, and k_1, k_2, ..., k_z represent the number of human errors that occurred. (Eq. 16)_can be rewritten as

$$P(f_1, f_2, \ldots, f_8 \mid k_1, k_2, \ldots, k_z, n, \theta) \propto P(k_1, k_2, \ldots, k_z \mid n, p, \theta) P(f_1, f_2, \ldots, f_8). \qquad \text{(Eq. 31)}$$

If we can assume that the experiments themselves are independent of each other, then the likelihood term in (Eq. 31) can be rewritten as

$$P(k_1, k_2, \ldots, k_z, n \mid f_1, f_2, \ldots, f_8, \theta) = P(k_1 \mid n, p, \theta) P(k_2 \mid n, p, \theta) \cdots P(k_z \mid n, p, \theta). \qquad \text{(Eq. 32)}$$

Thus, the PSF distributions can be updated as

$$P(f_1, f_2, \ldots, f_8 \mid k_1, k_2, \ldots, k_z, n, \theta) \propto P(k_1 \mid n, p, \theta) P(k_2 \mid n, p, \theta) \cdots P(k_z \mid n, p, \theta) P(f_1, f_2, \ldots, f_8), \qquad \text{(Eq. 33)}$$

where p is calculated from (Eq. 13).

Case 4: Correlation between PSFs

As shown in Table 6, the PSFs are discrete random variables that multiply a baseline probability, p_0, to compute the overall HEP. Thus, under favorable conditions, the multipliers will take values less than 1, and under unfavorable conditions they will take values greater than 1. It may be the case that one or more pairs of PSF multipliers tend to take favorable or unfavorable values under the same overall operating conditions. If so, these factors can be said to be correlated. Consider two PSFs; one accounting for the stress level of the operator and one accounting for the time available to complete the task. If when the time available is favorable then the stress level also tends to be favorable, and vice versa, then these two factors may be said to have some degree of positive correlation.

If this is the case, then the PSFs are not independent, and (Eq. 18) does not hold. It is necessary to modify (Eq. 18) by substituting joint PMFs for the correlated factors; however, the joint PMFs are rarely known. It may be possible, though, to simulate the necessary joint PMFs if data is available on the occurrence of the factor values, or with expert opinion. It will first be

necessary to estimate the correlation coefficient between the factors under question. This could be done using available data or expert opinion.

The correlated discrete variables, (f_1, f_2) say, can then be simulated by first generating correlated standard normal variables, (V_1, V_2), and transforming them using

$$f_1 = F_1^{-1}[\Phi(V_1)] \qquad \text{(Eq. 34a)}$$

$$f_2 = F_2^{-1}[\Phi(V_2)], \qquad \text{(Eq. 34b)}$$

where F_1 and F_2 are the discrete CDFs for the PSFs f_1 and f_2. Alternatively, we may also generate correlated uniform variables instead. In general, the correlation coefficient between f_1 and f_2 and that between V_1 and V_2 will not be the same. That is, we know $\rho_{f_1 f_2}$ from the data or expert opinion, but we do not know $\rho_{V_1 V_2}$. Using the definition of covariance, the following relation can be derived [see, for example, Haldar and Mahadevan, 2000]:

$$\rho_{f_1 f_2}\sigma_{f_1}\sigma_{f_2} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left\{F_1^{-1}[\Phi(u_1)] - \mu_{f_1}\right\}\left\{F_2^{-1}[\Phi(u_2)] - \mu_{f_2}\right\}\phi_2\left(u_1, u_2, \rho_{V_1 V_2}\right) du_1\, du_2 \qquad \text{(Eq. 35)}$$

where ϕ_2 is the standard binormal PDF.

The solution to (Eq. 35) can be found using a non-derivative numerical method such as the bisection method, where the double integral is evaluated numerically, such as with Gaussian quadrature. It was found by the authors that depending on the discrete PMFs of F_1 and F_2, the transformation of (Eqs. 34) may not always be capable of achieving the desired value of $\rho_{f_1 f_2}$. This result was also observed by [Van Ophem, 1999], who proposed a method for estimating the joint probability distribution for correlated discrete variables when the marginal distributions are known by relating the discrete variables to the bivariate normal distribution.

Once a large set of correlated variables (f_1, f_2) has been generated, their joint PMF can be constructed. The updating process can then proceed by simply using the appropriate pair-wise joint probabilities in the calculation of $P(f_1, f_2 \ldots, f_8)$ in (Eq. 16).

2.3.3 Numerical Example

The objective of this numerical example is to illustrate the methodologies for the four cases discussed above. Typical nuclear power plant incident data reports are used in this numerical example to update the SPAR-H model. The SPAR-H model is specified for two modes: "action" and "diagnosis." Here we deal only with the "action" portion of the model, for illustration. The PSF multipliers for action and diagnosis are identical in the SPAR-H model. The prior PMF estimates for each of the performance factors are given in Table 6, and the nominal HEP, p_0, is assumed equal to 0.001. The necessary details for the application of each case in the proposed methodology are presented first. Then, plots of the updated PSF distributions are given.

Case 1: Data Available on Binomial Outcome

As described earlier, the necessary data for this case are the number of errors, k, and the number of trials (or opportunities), n. From the available data, we have $k = 52$ and $n = 80$. These values will also be used in Cases 2, 3, and 4.

Case 2: Data on both Input and Output

For this case we use the same human error data, $k = 52$ and $n = 80$. However, we will now use empirical data on the model inputs as well as on the model outcome. The data on the model inputs give reported frequencies for the PSFs, and are given in Table 7. Note that the total number of reports for each performance factor does not add up to 80. This is because each performance condition does not get reported for every trial, but the Bayesian approach allows for the use of whatever data is available.

The updating process is split into two steps. First, the PSF distributions are updated using the input data given in Table 7 by applying (Eq. 24), where α_i is given by the prior estimated probability of observing the ith state for that particular factor. For the second step, these updated distributions become the prior distributions and the model is updated as in (Eq. 19) using the human error (outcome) data.

Table 7 Reported PSF frequency data.

PSF	Multiplier	Frequency	Total
	10	0	
F_1 Time	1	11	73
	0.1	58	
	0.01	4	
	1	0	
F_2 Stress	2	26	73
	5	47	
	1	0	
F_3 Complexity	2	14	72
	5	58	
	0.5	19	
F_4 Experience/Training	1	44	69
	3	6	
	1	0	
F_5 Procedures	5	4	69
	20	9	
	50	56	
	0.5	0	
F_6 Ergonomics	1	8	68
	10	60	
	50	0	
F_7 Fitness for Duty	1	0	5
	5	5	
	0.5	31	
F_8 Work Processes	1	46	78
	5	1	

Case 3: Correlation between Trials

Recall that correlation between trials means that the probability of human error for a given trial is not only dependent on some probability, p, but also on whether or not errors occurred at previous trials. To account for this, we use a generalized binomial distribution to calculate the likelihood of observing the data. Altham's multiplicative generalized binomial distribution is used here, and is given by (Eqs. 25 and 26).

The first step is to use the available data to estimate the distribution parameters, specifically the correlation factor θ, since p in (Eq. 33) will be replaced by the value calculated from (Eq. 13). In order to estimate the correlation parameter, it is necessary to have data available for multiple records, ideally with each having the same number of trials. For this example, there are eighty data points indicating whether or not a human error occurred. Thus, the data must be artificially divided to give multiple sets. The data is grouped so that trials occurring on the same day are part of the same set. This introduces another problem, since each set does not contain the same number of trials. This is overcome by calculating the average number of trials per record, and then normalizing each record to contain this number of trials. This will introduce error in the computation, but it is necessary to allow for the estimation of the distribution parameters.

The modified data now contain 20 sets of 4 trials each, and the average probability of error is 0.606. The sample variance of the number of human errors is 0.30, which is less than the expected binomial variance of 0.96, suggesting a negative correlation between trials within each set. Using the maximum likelihood formulas given by (Eqs. 27 and 28), the parameters for the multiplicative generalization of the binomial distribution were estimated to be: $\hat{p} = 0.72$ and $\hat{\theta} = 2.68$. Recall that for the multiplicative generalized binomial, $\theta = 1$ corresponds to no correlation, and $\theta > 1$ corresponds to negative correlation. We can now update the model by applying Eq. (21).

Case 4: Correlation between PSFs

The nature of the PSFs used in the SPAR-H model suggests that several of the PSFs may be correlated. For this example, the method described for case 4 can be used to simulate the joint PMF between two PSFs. The available PSF occurrence data is used to estimate the correlation coefficients between pairs of PSFs. This is done by first transforming the occurrence data into multiplier values, then calculating the sample correlation coefficient between the factors.

From the available data, one of the strongest correlations between PSFs is found to be between F_5, procedures, and F_6, ergonomics. The sample correlation coefficient from the data is found to be 0.68. Using the CDFs for the two factors, it is found that this value fell within the range for $\rho_{f_5 f_6}$ that could be simulated, which is (-0.21, 0.75). This range is easily found by generating perfectly correlated normal deviates and performing the transformation of (Eqs. 34). Using the sample correlation coefficient, $\rho_{f_5 f_6} = 0.68$, the numerical solution to (Eq. 35) is found to be $\rho_{V_1 V_2} = 0.957$. Upon simulating samples of F_5 and F_6, the authors found by trial and error that adjusting $\rho_{V_1 V_2}$ to 0.93 gave simulation results that more closely matched the desired correlation between F_5 and F_6. This discrepancy could possibly be due to the inaccuracies associated with solving (Eq. 35) numerically.

Thus, $\rho_{V_1 V_2} = 0.93$ is chosen and forty-thousand pairs for F_5 and F_6 are then simulated and used to construct the joint PMF. It is assumed for the sake of illustration that the remaining

factors occurred independently, and the prior PMF in (Eq. 16) is calculated as $P(f_1, f_2, \ldots, f_8) = P(f_5, f_6) \prod_{i \neq 5,6} P(f_i)$, where $P(f_5, f_6)$ is calculated using the simulated frequencies for F_5 and F_6. A comprehensive analysis could consider the significant correlations among all the PSFs.

Numerical Results

The results achieved using each of the updating processes described above are presented in Figures 10 – 17. The figures show both the prior and updated PMFs for each of the eight PSFs for the four cases. Note that for the majority of cases and factors, the updating process shifted the PMFs to show higher probabilities for larger multiplier values. This was expected since the data gave such a large proportion of failures. Also note that as expected, the results for Case 2 agree with the PSF frequency data given in Table 2. For each factor, the PMF updated by Case 2 shows that the multiplier with the largest probability corresponds to the multiplier that was reported with the highest frequency.

A significant difficulty that is present when working with HRA is the lack of reliable empirical data sources. Although efforts are being made to improve the quantification of and availability of human reliability data, it is necessary to be aware of the issues associated with using these data. One major problem is that when human error figures are reported in a format such as "*k* failures out of *n* opportunities," it is very difficult to know the correct value for the total number of opportunities for failure, *n*. For example, the data used in this paper are based on 80 incident reports. This does not necessarily mean 52 human errors out of 80 opportunities, since there are no incident reports when the plant is functioning normally. Also, this figure is grossly inconsistent with the estimated nominal failure probability of 0.001, and it is most likely the case that these samples are failure-biased. If this is indeed the case, then using these data to update HEPs, as in the above example, will introduce error. To actually implement this process, better estimates of the number of opportunities would be required.

The purpose of this presentation was to present a methodology for the application of Bayesian updating to refine quantitative human reliability models with empirical data. The SPAR-H model for human error probability was employed for the purpose of illustrating the theory for several cases of Bayesian updating. The Bayesian analysis made use of prior estimates on HEPs, human performance data, and performance factor occurrence data to compute updated HEPs and updated PMFs for the performance shaping factors. Methods for analyzing correlation between event trials as well as correlation between PSFs were outlined as well. Also note that any or all of the cases described above could be combined, given that the appropriate data are available.

A case of interest for future study is the case where the outcome of each trial is allowed to have more than two states, as opposed to only failure or success. The additional outcomes might correspond to some intermediate state such as a malfunction or degraded performance. The analysis would remain the same, except that the likelihood term would be calculated using the multinomial distribution as opposed to the binomial. Since the current SPAR-H model does not provide multipliers for any outcome other than failure, it would need to be modified to allow for the calculation of more than two outcome probabilities. This would involve expanding (Eq. 13); the PSF multipliers, and PMFs would need to be modified as well.

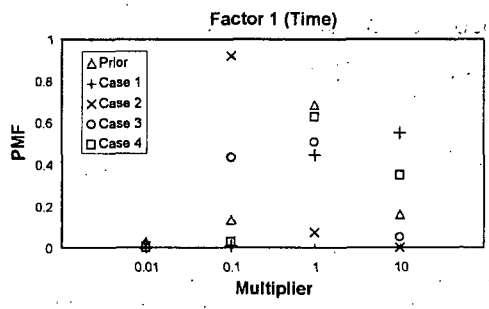

Figure 10 Prior and updated PMF for factor 1

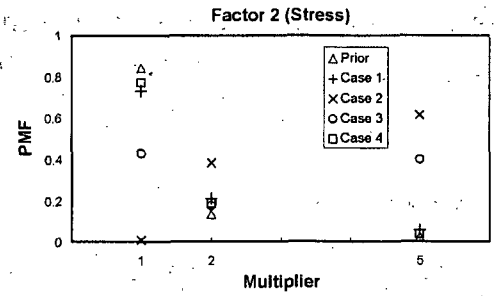

Figure 11 Prior and updated PMF for factor 2

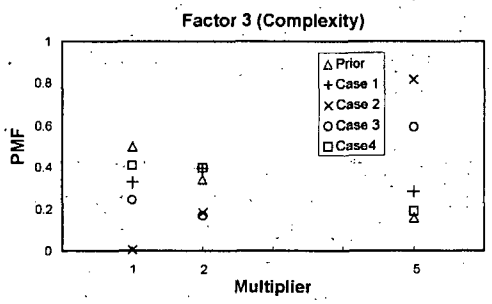

Figure 12 Prior and updated PMF for factor 3

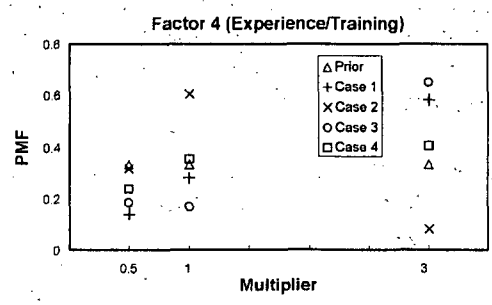

Figure 13 Prior and updated PMF for factor 4

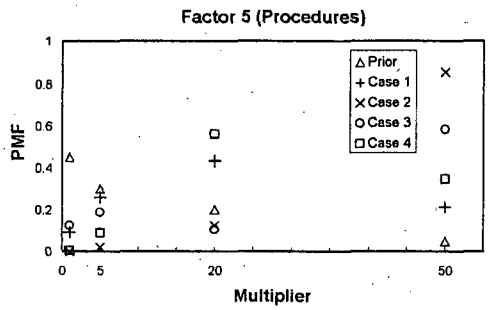

Figure 14 Prior and updated PMF for factor 5

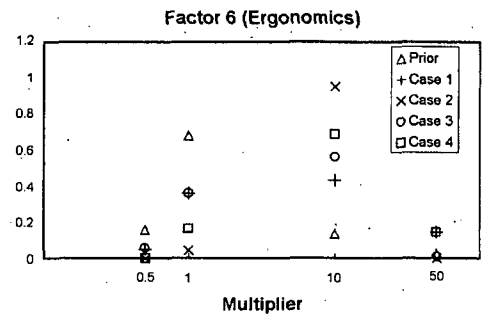

Figure 15 Prior and updated PMF for factor 6

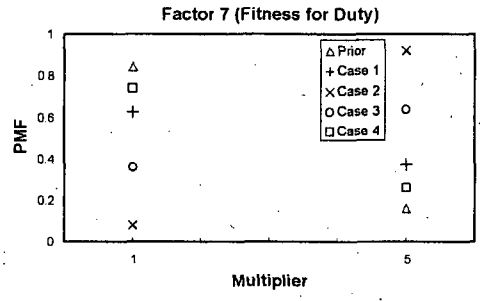

Figure 16 Prior and updated PMF for factor 7

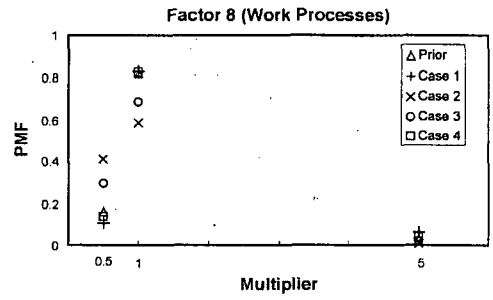

Figure 17 Prior and updated PMF for factor 8

2.4 The Use of HERA and Bayesian Analysis to Support Quantification in Context-Based HRA Methods

Prepared by Dennis Bley

2.4.1 Introduction

Could Bayesian approaches be helpful in second-generation HRA methods that emphasize the importance of context in evaluating the likelihood of human response? We begin by outlining why consideration of context, the conjunction of plant conditions and human performance conditions, is essential in analyzing and quantifying human performance. Next is an example of how context is used in ATHEANA and how its current quantification process is organized. This sets the stage to discuss two ways in which Bayesian approaches could enhance and, perhaps, simplify the quantification of such context-based models. Finally, we discuss the quantification of an HRA model.

Why do we focus on context? Because *serious accidents* in nuclear plants and many other industries almost invariably involve significantly difficult context, sometimes called cognitively complex situations or error-forcing context (EFC) [Bley, *et al.*, 1987]. The problem is that under strong EFC, operators can mentally lock onto their first assessments and fail to update their evaluation as the situation evolves. Often, they even fail to recognize incoming evidence as related to the situation. Multiple cues are no longer independent signals. Strong EFC involves adverse plant and human conditions, and the most severe cases are typically driven by deviation from expected plant response (i.e., mismatch between operator mental models and the actual situation) or significant mismatch between PSFs and the actual situation.

In this regard, the following characteristics have been identified in many serious accidents:

1. Deviation—operations outside expectations

2. Resulting physical regime not understood

3. Operators "refuse to believe" evidence coming to them.

It is not that this list is a proved requirement for situations where unsafe actions (UAs) are likely to occur and to persist. Rather, it is an observation that in many accidents in many industries, similar sets of occurrences are observed. Plausible arguments about why this is so can be offered, but none are yet proved. When deviations from expectations are minor or within the realm of training and mental models held by operators, their procedures, experience, and the operators' own situation assessments are all likely to align and all but assure success in time to prevent severe damage. Likewise, if plant conditions are awry but human conditions are favorable, operators are likely to salvage the situation.

This type of situation (i.e., the way operators can miss multiple strong cues and fail to correct dangerous situations when faced with strong EFC), makes us suspect that a simple model using a base failure rate, λ, and a multiplier based on simple descriptions of context (Figure 18, left side) cannot be an appropriate model. At a minimum, it would seem that such a multiplier function is not well-behaved.

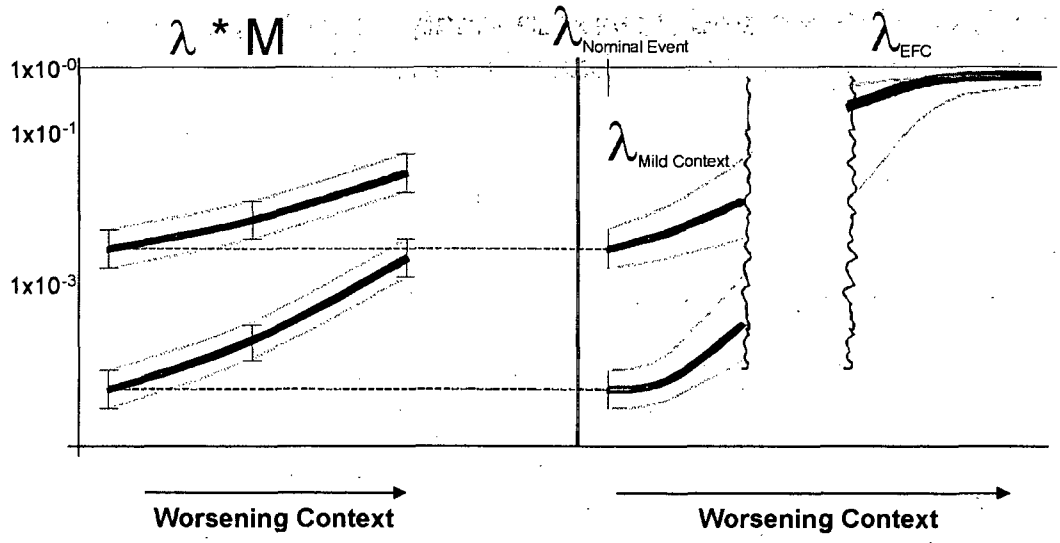

Figure 18 Possible Models for Human "Error" Rate Variability.

Perhaps it is reasonable that, in some conditions where λ is well-known, slight changes in conditions can be modeled or represented using a simple multiplier. However, at some point, as the EFC becomes more and more severe, there is a jump to a different regime, where failure is very likely, say approaching p=0.5, as shown in the right side of Figure 18. This is similar to the control regimes of Hollnagel [1998], in which the failure causes are much different than during the nominal condition.

2.4.2 Bayesian Approaches to Improve Quantification in ATHEANA: Description and Examples

The approach for defining and modeling context in ATHEANA is based on the multidisciplinary ATHEANA Framework sketched in Figure 19 [Barriere, *et al.*, 2000]. Here we see that the PRA models human failure events (HFEs) and includes a simplified model of reality, i.e., plant states. However, the EFC is driven by the real world, where "plant conditions" include equipment that operators interact with directly and that may not be modeled in the PRA. Together, plant conditions and performance shaping factors (PSFs) trigger human error mechanisms that lead to unsafe acts. Unsafe acts or combinations of unsafe acts are the events modeled in the PRA as HFEs. Thus, the likelihood of an HFE is driven by context (plant conditions and PSFs), rather than by the simplified plant states of the PRA.

Currently, ATHEANA quantifies P(HFE) by breaking the problem into its constituent parts. For a simple HFE relating to a single UA, this can be expressed as:

$$(HFE) = P(EFC) \times P(UA|EFC)$$

The probability of the context, P(EFC), is a systems analysis problem and is calculated by the same systems analysis methods as used in the PRA. The probability of the unsafe act, given the context P(UA|EFC), is currently evaluated using a consensus expert elicitation approach.

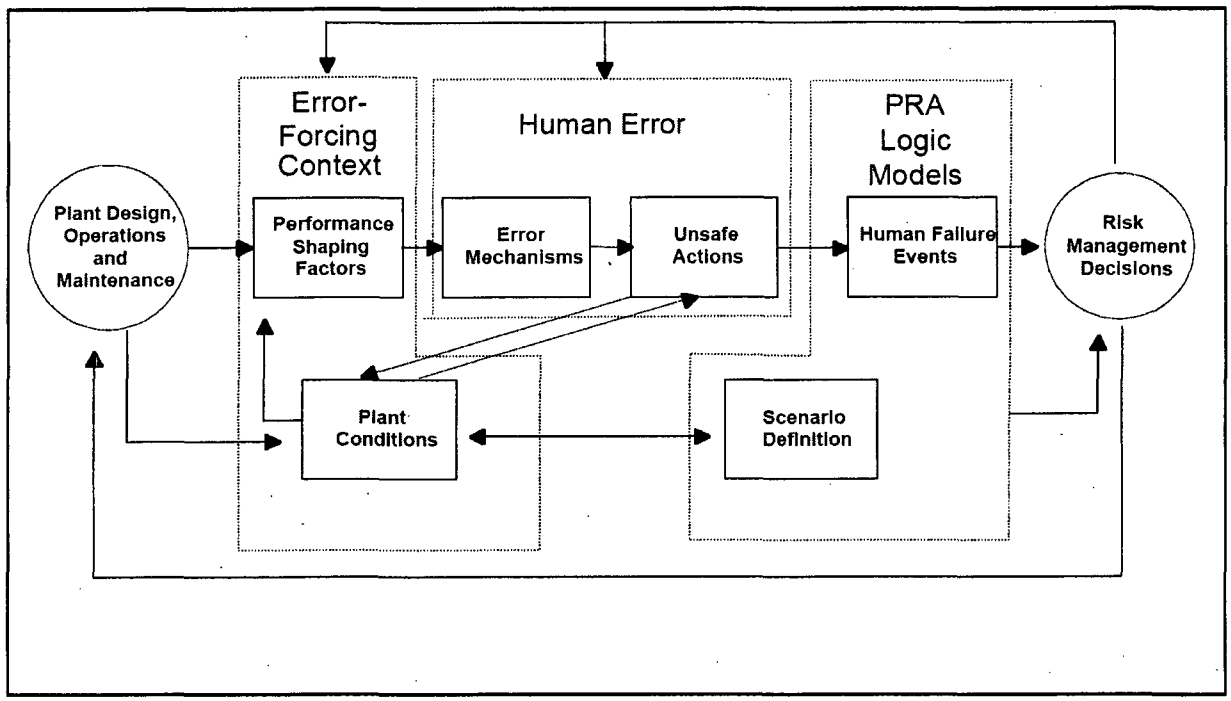

Figure 19 The Multidisciplinary ATHEANA Framework.

This has been found to be a reasonable means to translate the impact of the EFC into an uncertainty distribution on the likelihood of the UA. The approach requires care to control bias, ensure full consideration of all available information, and fully question and check the consistency of estimates [Barriere, *et al.*, 2000]. The approach strives to facilitate the derivation of realistic quantitative estimates. It encourages the analyst to consider all sources of information, including partially relevant data. However, there are always potential limitations in any approach that relies on individuals to directly transform qualitative, or even semi-quantitative information, into probability distributions. There is also a stringent requirement to have experienced and capable team members representing the multiple disciplines outlined implied in the framework of Figure 19 – behavioral scientists, engineers, operators, PRA/HRA experts. Therefore, alternatives to simplify the process and to create as objective an approach as possible are worthy of attention. Several of these can be Bayesian approaches, either philosophically Bayesian or actual Bayesian inverse probability calculations.

Simplified approaches might be possible, once a catalogue of detailed contexts has been identified, examined, and organized, as part of the HERA database[2]. To clarify this point, consider two distinct types of methods:

[2] HERA has the capability and structure to include qualitative descriptions of context and human performance including details of the plant conditions, PSFs, mismatches, and deviations. It includes the kinds of information that the ATHEANA team found helpful in performing analyses. This encompasses the following kinds of qualitative information:

1. *Methods for uncovering the mental mechanisms* that are associated with "human error" for the purpose of determining P(UA|EFC)

2. *Calculational methods that could incorporate collected data and results from many ATHEANA-style analyses* into a simple-to-apply high-level tool.

An example of the first is the current ATHEANA method. It requires an experienced and broadly multi-disciplinary team to perform a competent analysis. The team generates detailed descriptions of context and unsafe acts and develops consensus probability distributions for P(UA|EFC). Results of many such analyses could be used to build a catalogue of EFCs and P(UA|EFC).

The second type of method would try to use the results of many expert-generated ATHEANA-type analyses to provide a basis for quantifying new events with somewhat similar context. Two possible Bayesian approaches to improve, simplify, or extend the current quantification method are described below–one based on "interpolating" among similar contexts and one introducing a Bayesian updating scheme based on updating more general data on human error with context-specific error-of-commission data. Note that, even if these never become simple to apply, they could systematize context-based HRA, helping ensure consistency and providing more convincing anchor values to support quantification. There is some hope that less broadly-based expertise would be required to use these methods.

2.4.3 Developing Generalized Contexts for Interpolation

2.4.3.1 Description

Recall that, in an application of ATHEANA as described above to a particular UA-EFC pair, the analysis team develops a consensus probability distribution for the likelihood of the UA under that particular EFC, P(UA|EFC). Let us call the results generated in each application of the ATHEANA quantification process a "context anchored probability" (CAP). Note that each CAP includes a detailed description of the error-forcing context and a probability distribution quantifying P(UA|that specific context).

The next step would be to build a library of contexts and associated P(UA|EFC) distributions. The source of CAPs in the library could include results from:

- ATHEANA proactive analyses
- ATHEANA quantification of real operating event descriptions

---

Event Summary	ATHEANA Summary
Event data	Deviations from the "expected" scenario
Event description	Event timeline
Event surprises	UAs, equipment failures, human actions, recovery actions
Key mismatches (between training and unusual plant conditions, supervision and plant conditions, procedures and the specific scenario, etc.)	Dependencies linking actions on timeline
	Accident diagnosis log, a table with three columns
Key parameter status; key facility/process status (initial and during the accident)	Time
	Accident progression and observed symptoms
Action summary	Human response to symptoms
Corrective actions	

- Experimental results for certain actions under a limited range of contexts
- Results extracted from existing experimental data

Exactly how each of these sources is developed needs to be carefully defined and illustrated. While we can imagine the process, it has yet to be demonstrated. However, once a substantial library is assembled, we propose to use it as the basis for quantifying new events. The basic idea is that the analysis team will define the context of new events by using the ATHEANA process. Then a search of the catalogue will identify previously quantified events with somewhat similar contexts; the new event will be quantified based on using some measure of closeness to the existing contexts. As the catalogue grows, we expect that the analysis will have a more sound underpinning.

2.4.3.2 Example

A large catalogue of specific CAPs is likely to be difficult to use. How could it be organized to permit identification of the most relevant CAPs for current purposes? One approach that appears promising is to sort CAPs into families of similar contexts. Each family could be considered to be a generalized CAP – a Generalized Context Anchored Probability (GCAP) – that represents all the members of the family.

As an example of the process for developing GCAPs, suppose we take five CAP distributions from the library as shown in Figure 20. Each distribution, θ_1 through θ_5, corresponds to a single CAP context. Note that the distributions θ_1nd θ_2 are very similar, as are curves θ_3 and θ_4. The descriptions of context for the members of each pair are also similar, but not given here.

Each of these two pairs of CAPs can be represented by single GCAP curves as shown in Figure 21, where our five CAP curves have been replaced by three GCAP curves. GCAPS, then, can be thought of as descriptions of classes of events that have been characterized according to the factors driving their occurrence and that have a probability distribution associated with them that is also strongly affected by those factors. The next question is, how can we use these GCAPs?

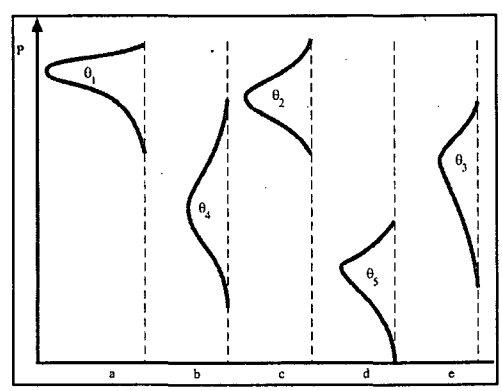

Figure 20 Five CAP Distributions

GCAPs are an unproved concept that appears to offer potential for:

- Standardizing or normalizing the judgments that are applied in assessing the probabilities of different unsafe actions
- Using operational experience from other events as reference cases in assessing the probabilities of the different unsafe actions
- Explaining the assessments to peers and outside reviewers.

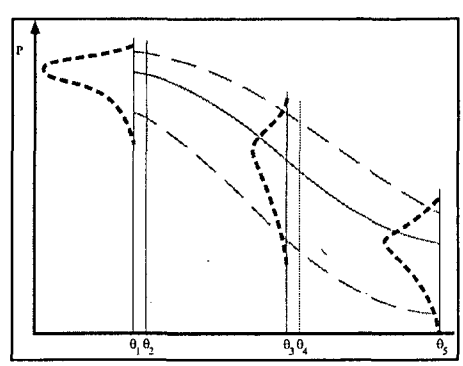

Figure 21 The Five CAPs of Figure 3, Grouped as GCAPs

Use of GCAPs for expert quantification and analysis would begin following the usual ATHEANA search process, including the identification of the full context of interest. The analyst would use the new approach to simplify quantification, drawing on the library of existing results. We might be able to use them to place a newly analyzed event within groups of event signatures.

Then, using some as yet undefined measures of "distance" between a UA-EFC set being analyzed and related GCAPs, we could generate a probability distribution for the new event through an interpolation process. Early on, this interpolation could be accomplished using the structured expert elicitation process described in recent ATHEANA papers and reports [e.g., Forester, *et al.*, 2004].

For some simple but powerful types of context, those with strong EFC, GCAP probability distributions can be directly assessed. Figure 22 shows how judgment concerning the relative strength, when compared to very strong context, would shift such a probability distribution curve.

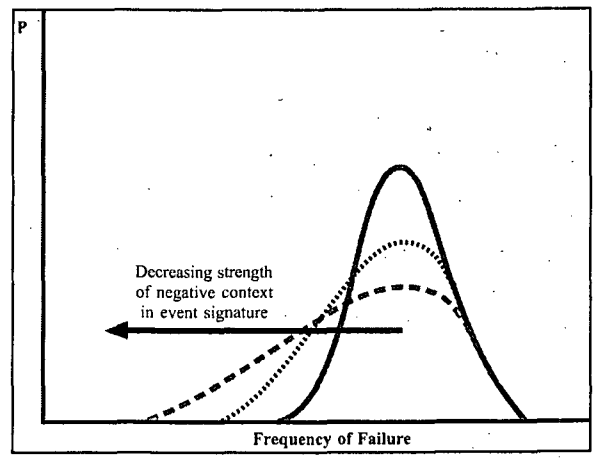

Figure 22 Shifting Strong-EFC Curve vs EFC

2.4.3.3 Information Needs

The primary information needed to test and use this approach is a viable catalog of GCAPS. For GCAPS from proactive ATHEANA analyses, we need reports of ATHEANA applications. For GCAPS based on ATHEANA quantification of real operating event descriptions, the collection of retrospective analyses of events with significant context in the HERA database needs to be expanded, followed by an ATHEANA expert evaluation of P(UA|EFC) for each UA in the database. For GCAPS based on experimental results for actions under a limited range of contexts, results from Halden or other simulator experiments are needed.

2.4.3.4 Future Activities

No activities to develop and test this approach are currently planned. The following activities can be performed as the HERA database is expanded:

- Develop estimates of P(UA|EFC) for events in the HERA database
- Build an associated catalog of GCAPS
- Attempt to quantify newly analyzed events from PRA applications using the GCAPS and expert elicitation to interpolate new P(UA|EFC) from the catalog
- Develop a Bayesian update model to permit evaluation of P(UA|EFC) from the GCAP catalog, rather than the expert process

2.4.4 Quantification Method Based on Data for Errors of Commission

2.4.4.1 Description

A Bayesian approach to quantifying P(UA|context) for EOCs has been proposed that could directly support quantification in ATHEANA [Barriere, M., et al.]. The underlying idea is based on an observation that a set of documented errors of commission fell into a limited number of

groups, based on characteristics of the context that existed at the time of the events. Prior human failure rates obtained from a large "human error" database can then be updated, based on elements of their context, using a likelihood function from the more limited EOC database.

2.4.4.2 Example

The published work [Reer, 2004] identified 180 events that involved EOCs; however, the number of opportunities was unknown. Therefore, these data are only helpful in estimating the fraction of EOCs with similar context. Used qualitative findings on context to define sub samples specific to context types; i.e., from random to various degrees of EFC; examples of this data are show in Table 7.

With this kind of information, it becomes possible to identify context types. Then, for each context type, an analyst can update "generic" failure rates to the specific general type of context through the use of Bayes theorem. The approach updates a general failure rate with context-specific likelihood functions. Results are context-specific failure rates.

2.4.4.3 Information Needs

For such an approach to be useful, a database that includes a large number of EOCs that occurred in a wide variety of contexts is needed. Useful context types must be defined and supported.

Table 8 Examples of Data Including EOCs.

#	Plant	Date	LER title	EOC identified?
1	North Anna 2	820101	Reactor shutdown due to high RCS leakage	No
9	Ginna	820125	Reactor shutdown due to SGTR	Yes, EOC1: isolation of steam relief control
99	Trojan	830122	SGs reach lo-lo level after AFW pumps fail to restart [LMFW]	Yes, EOC2: shutdown of AFW pumps
125	Hatch 2	830714	Rods inserted out of sequence [Loss of condenser vacuum]	Yes, EOC3: bypass of Rod Sequence Control System
157	San Onofre 2	831007	CEAC failure causes scrams	No
158 (2nd count)				
180	Browns Ferry 1	831231	RCS has high chloride concentration	No

2.4.5 A Caveat on Context: Plant-to-Plant Crew Variability

Only a limited number of accident reports provide information about how crew members interact and what administrative procedures they operate under. Nevertheless, there is evidence that these practices form an important part of the context and, under specific alignments of conditions, can have great impact on success or failure of the crew to prevent or control damage. Any approach using Bayesian reasoning needs to account for the observed effects of crew characterization, if the results are to give an accurate portrayal of risk.

Experience has clearly demonstrated the importance of crew operating characteristics to plant-specific HRA. That experience included visits and training exercises at four US PWRs in

gathering data for the NRC-industry PTS study [Kirk, M.E, 2005] and an ATHEANA development effort at another US PWR. When we observed simulator drills, our purposes included:

- Observing timing
- Observing use of procedures and "informal rules"
- Determining if scenarios were cognitively challenging.

During these visits and observations we found some surprises that emphasize the importance of crew characterization. We found dramatic plant-to-plant differences in how crews behaved with respect to the following attributes, differences that have strong influence on the likelihood of success, under specific contexts.

Do crews act in concert or independently?

We saw cases with nearly opposite approaches. At some plants the crews acted as tightly coupled teams led by the Shift Supervisor, who follows the procedures step-by-step and each member informed the others step-by-step, throughout the entire exercise. There is something of a mythology that this approach is universal. We found several alternative, formalized approaches among the plants we visited, where crewmembers carried out independent actions followed by reports to the Shift Supervisor. Several of these approaches, sanctioned by plant administrative procedures were observed:

1. Crew members followed independent "Initial Action Cards," then report.
2. Individual crew members carry out long verification lists and report when done.
3. "Rules" (quick action cards on fast-response-needed situations) are carried out independently, followed by verbal reports to the Shift Supervisor.
4. The Shift Supervisor keeps his head out of the procedures; ROs have spiral-bound EOPs.

Generally, these alternatives evolved from the recognition that, in some designs, specific accidents can progress very quickly, and executing prearranged sets of activities can provide operators with improved performance and flexibility. Since these plans are prearranged, the Shift Supervisor can truly perform a supervisory role, following the overall progress of the event and focusing on higher level diagnostics.

Different plants use very different "key indicator screens"

In many plants, a set of "first out" indicators lock-in lit tiles to show which specific signals first initiated key response systems (i.e., reactor SCRAM in the event of a reactor trip). Such key indicator screens are helpful in diagnosing the sequence of events that lead to plant upset. The particular key indicator screens vary plant-to-plant. In our plant visits, four distinct types were used:

1. First out panel (some plants do not have these)
2. "Strip chart" displays of multiple parameters – some were standard displays with pre-selected parameters, others were custom made as they were needed
3. Computer alarm screens
4. Reactor pressure-temperature (P-T) plot.

The first simply provides confirmation of the specific signal that initiated the reactor trip. This is helpful in supporting the operators' situation assessment and moving them to or through their emergency procedures. Crews having the strip chart displays use them to confirm that progress of the event matches their expectations and the flow of the emergency procedures. They are sometimes helpful in identifying situations that diverge from expectations, hastening necessary corrections in response. Properly organized computer alarm screens can provide a very helpful history of the event, yielding the information of the first out panel and some of the advantages of

the strip chart displays. Poorly organized computer alarm screens can lead to information overload and confusion. The P-T plot approach was very interesting and only works if the Shift Supervisor is not buried in the step-by-step activities of the emergency procedures. When the Shift Supervisor is overseeing operations and is conversant in using the P-T plot, it can be a powerful tool for tracking the trajectory of an accident. It can help break the mindset sometimes created by strong context.

Formality of communications

Communications is an essential part of crew response to plant upset. Miscommunications, especially unfulfilled assumptions, about the intentions and actions of other crew members has been at the heart of many events (including simulator drills) that get out of control The industry has developed formal communications strategies that can minimize the chance of miscommunications and unfulfilled expectations (assumptions). Best practice has settled on "three-way communications," where, when a command is given, it is repeated back by the person receiving the command, and then the originator confirms, repeating the full command. Names (or position titles) of the individual who is to receive the communication are explicitly stated. Although this is most common, it is not yet universal and sometimes is not carried out in the spirit of cooperation and positive control.

When the spirit is awry, lip service is given to the process, but the crew is not carefully listening to the responses. Communications can go as far wrong, as when less formal modes of communication are used.

Briefing strategies of many types were observed

During the sequence of events, briefings can help detect when the crew's responses (and likewise the steps of the procedures) are not effectively resolving the upset conditions. Briefing practice runs the gamut from simply informing the full crew of the situation (when it seems convenient to do so) to a formal process performed at regular intervals (when deemed useful by any crew member uncomfortable with the current situation) to confirm the situation assessment and uncover any divergence from expectations. The practices we have observed include:

1. Occasional briefs by Shift Engineer
2. EOP-driven briefs
3. Structured briefs (e.g., "BAG" or Before-At-Going), when the Shift Supervisor can break action or when called by another crewmember. The Shift Supervisor asks all members of the control room team to agree on what has just happened (Before), where the plant is currently (At), and what is expected to happen next (Going); this questioning process and drive for consensus can uncover misperceptions, and the agreement on where it is going ensures early detection of a plant behaving in unexpected ways. This process may be the most effective means of breaking the effects of error-forcing-context.

Approach to verbatim compliance

Finally, the idea of verbatim compliance with procedures has evolved into multiple interpretations, most of which give crews an improved chance of success, when the actual situation is not well-matched to their training and emergency procedures. The range of flexibility we found covers the following range:

1. Absolute verbatim compliance, unless a public safety hazard is imminent.
2. Use of a structured agreement process to deviate from the current procedure. In one plant, the Shift Supervisor offers or entertains a proposal to deviate, based on observations that the current approach is not working. If a deviation is to be carried out, each member of the crew must state their opinion and all must agree.
3. The Shift Supervisor (or Shift Supervisor and Shift Engineer) orders a deviation.
4. The plant has a formal deviation procedure. In one case, the plant has a separate procedure that is essentially a template for constructing tailored EOP on the fly.
5. The plant has defined certain conditions that allow "parallel actions" (simultaneous use of multiple procedures) or define a set of "continuous steps" that can be carried out in the middle of the current procedure.
6. Certain defined procedure jumps are permitted under specific, well-defined conditions.
7. The Shift Supervisor can pull any step forward (conduct it earlier than called for), if he has "evaluated" it.

Essentially, all these deviations from verbatim compliance have been designed to cover unusual situations that are not well-covered in the printed procedures or specific situations that analysis has shown to require more rapid action.

It seems reasonable to believe that the range of characteristics observed at these five plants is representative of PWR plants. Importantly, the differences observed change the way crews interact and move through their emergency procedures, and this means that the crews from different plants have very different vulnerabilities when operating under strong, unusual context.

The net effect of these variations in plant practice can have significant impact on operator actions in complex event sequences. The ensemble of these practices at a specific plant is what we call "crew characterization."

3. SUMMARY OF PEER REVIEW

Prepared by Alan Kolaczkowski and Bruce Hallbert

3.1 Peer Review Team and Purpose

This section documents the peer review that was conducted on the draft version of this document. Seven peer reviewers were selected on the basis of their involvement in PRA and HRA in particular, as well as having first-hand knowledge in the use of Bayesian and other quantitative techniques and how such techniques might be useful to HRA. A questionnaire was provided to the reviewers as a means to establish the purpose and expectations for the review and to specifically guide the reviewers in focusing their reviews. While most reviewers provided explicit responses to each question on the questionnaire, not all did so nor was that required. But all did, in one form or another, generally address the considerations outlined in the questions that were provided. The questionnaire that was used to obtain peer review comments contained questions designed to evaluate the following:

1. Whether the draft report established Bayesian and other techniques as theoretically valid frameworks for using empirical evidence and supports their use in Human Reliability Analysis

2. Whether the examples of methods provide adequate illustration of their stated purpose(s), are clear, may potentially advance HRA in specific ways, and meet or have the potential to meet stated workshop objectives

3. Whether the methods appear to be feasible and may help with some of the current needs in HRA discussed in the workshop.

3.2 Summary of Peer Review Comments

The reviewers provided numerous written responses, ranging from general comments on the overall objective of the report and approach taken to accomplish it to very detailed technical comments on each of the proposed approaches and recommendations on the technical quality or the usability of these methods. The reviewers also provided comments to otherwise improve the document.

A primary theme of the comments was that the document successfully addressed the goals of the workshop, though the examples and methods require greater maturity and testing to achieve their objectives. Generally, based on the comments, it can be concluded that while the document provided reasons and illustrations to be positive about the feasibility of using Bayesian and other quantitative techniques to improve HRA, the document is considered to be too general in its content to be of use for specific HRA applications at this time. Reviewers agreed that another major step is needed before the more formalized use of Bayesian and other techniques for improving HRA can be practically implemented. This next step needs to address general and specific issues confronting the field of HRA, how specific techniques can be applied to address these issues, and expectations regarding the results of applying such techniques.

3.3 Addressing the Peer Review Comments

Following the peer review, some comments were addressed through modification of the draft report. Such changes included revising the presentation of the information, the style, and the reference materials and references, and consolidating or relocating information in the section and subsection to which reviewers thought it best applied. In addition, some reviewers expressed the concern that the draft report emphasized Bayesian concepts over the other quantitative techniques that were also included in the draft report. To address this comment, pertinent sections were rewritten to better reflect the concepts proposed that involve Bayesian techniques and accommodate the discussions and information pertaining to the other quantitative techniques described in the workshop summary.

The reviewers provided constructive suggestions for improving the usability of the quantitative techniques that were introduced or discussed in this workshop. To some, especially those in the HRA community, these techniques are new. To others in the PRA community, such techniques have been employed successfully and have had a positive impact in formal risk assessment. However, these techniques have been applied much less in the HRA field. In future activities that may involve development of or application of quantitative techniques, the reviewers suggested several concrete activities that we also recognize including:

- Focus initial efforts to a limited number of problems that may be well-described already in the HRA field. This could include, for example, how to combine several sources of relevant information to better estimate the reliability of a human action; using human performance data obtained from plant simulators to estimate HRA model parameters of interest, etc.

- Develop the quantitative examples and applications as though applying them to a particular problem of interest, documenting the steps involved.

- Describe important assumptions that must be made to use the quantitative technique that analysts may not be familiar with already. For example, Bayesian methods in particular rely upon a likelihood function to describe the process that governs the predicted posterior probability. The particular form of the likelihood function is important and should be chosen on the basis of how to best describe the uncertainty in the underlying psychological mechanism being modeled. Little attention has been paid this issue to date and due consideration will need to be paid this issue if Bayesian methods are employed with human performance data.

- The types of information and sources of information for the types of analyses discussed in the workshop must be considered in order to make use of the quantitative techniques. Sources of information for HRA are notoriously scarce and recommendations for improvements in estimating HRA quantities of interest should bear in mind the availability and suitability of data to support their implementation and use. In other words, the nature of the evidence that can be obtained from different sources needs to be understood to determine whether it can be used in a quantitative framework.

4. SUMMARY AND CONCLUSIONS

Prepared by Bruce Hallbert and Alan Kolaczkowski

4.1 Introduction

Reiterating the statement provided in Section 1, the overall purpose of this document is to summarize discussions and proposals offered by invited workshop presenters to address the feasibility as well as the associated issues relevant to using quantitative methods to employ evidence for improving our human performance models and gaining more confidence in the qualitative and quantitative results produced by HRA methods (perhaps even to the point of validating the methods to some degree). In particular, these specific questions were raised in Section 1 that should be addressed in order to be responsive to the overall purpose:

1. Do quantitative techniques offer a theoretically valid framework for using empirical evidence to inform our current HRA methods?

2. What are some examples of ways we could inform current HRA methods (i.e., provide illustrations)?

3. What more needs to be done to demonstrate the feasibility of using these methods and empirical evidence to inform current HRA methods?

The following subsections provide our observations as to the progress made in answering the above questions. These observations are based on the information provided in this report and particularly the contents of Section 2 that reflect the discussions held at the workshop in Washington, D.C. on August 10-11, 2005 to discuss approaches to the use of data and forms of evidence to support the prediction of human performance in PRA applications.

4.2 The Validity of Using Quantitative Techniques and Empirical Evidence for HRA Use

To address the use of empirical evidence for HRA quantitative techniques, it is first important to understand the parameters of interest that analysts are typically required to estimate for PRA purposes. In its simplest form, this involves estimating the probability of success or failure in performing some action along with understanding the most influential factors affecting the probability, whether that action is physical or cognitive in nature.

The information provided in Section 2.1 addresses this subject from a theoretical perspective. In that section, we explain that the quantity of interest in PRAs is the probability of a failed outcome regarding a human action of interest, which is labeled as "p" in the discussion. Further, because decision makers need to be aware of the analyst's level of confidence and hence the robustness associated with the estimation of "p", we are also interested in "$\pi(p)$", the probability distribution for "p" that represents the sources of variability and uncertainty in "p".

The estimate of "p" for human performance is not like predicting an equipment failure rate because typically for equipment, it is often assumed that the context or environment for the operation of the equipment is generally uniform. That is, the environment in which equipment performs can be largely disregarded for most conditions once specified, unless the nature of the event sequence creates changes that are important to the functioning of the equipment. With regard to human failures, we need to be concerned with how the personnel perceive the

conditions under which performance is demanded. These perceptions can include their interpretations of events, their use of plans and procedures, and the applicability of their training, among other influences that affect the actions that crews will take in ways that materially affect "p". Most HRA methods typically model human performance by evaluating, in some way, the relevant influences for an action of interest (typically termed performance shaping factors [PSFs]) for a plant condition/sequence, so as to ultimately estimate "p" based on these influences.

Ideally, direct evidence would provide us with a basis for estimating the rate of certain kinds of errors in specific contexts, but such data are not yet available. Bayesian methods are appealing since they afford the ability of combining our current evaluations of human performance with different types of evidence, via likelihood functions, to produce updated assessments (updated in that they are a reflection of the collected evidence) of human performance manifested as posterior probability distributions for "$\pi(p)$," as well as an updated best estimate for "p". Further, as illustrated by the various topics addressed in this workshop, Bayesian methods can employ different amounts of evidence including sparse evidence.

Together, the discussions presented in Section 2 introduce a variety of ways to characterize the relationship between evidence use and probability estimation. They also demonstrate the use of Bayesian techniques with sparse data, and that meaningful probability estimates can be made with less evidence than one might thought to be necessary using classical statistics as a frame of reference. The same is shown to be true even when data that we would like to have is missing or incomplete (such as the coin tossing example discussed in Section 2.1). These examples provide expectations that the use of Bayesian methods holds a promise of being able to utilize the types of empirical evidence regarding human performance that is potentially available despite being incomplete, sparse, anecdotal, and/or otherwise undesirable.

Committing to Bayesian methods to assist in estimating probabilities for some human actions does not, in itself, provide an answer to the difficult questions of what information to use, how to structure such calculations, how to weight the many forms of evidence that are available, and how to best account for context in a way that most appropriately represents the spectrum of situations to which the results may be applied.

It is critical to the use of Bayesian methods that all of the relevant evidence for estimating the quantity of interest be known and be made available, whether it is incomplete, partially relevant or not. There has been some debate among practitioners regarding how to identify HRA parameters of interest for a particular PRA application, how to measure them, and how to weight them in analyses. The theoretical and practical illustration in Section 2.2 uses actual data from simulator studies involving licensed NPP operators and addresses some of methodological issues in attempting to address these questions. The methodology is flexible and easy-to-use and may help address some of the more fundamental questions in HRA such as: (1) What PSFs are most important to model in a PRA accident sequence? (2) How much do the PSFs themselves contribute to crew performance and how much is explained by other factors? (3) Relative to one another, how much weight should be attached to individual PSFs in predicting crew performance? (4) As they are perceived by crews, are these PSFs more systematically related to one another on a higher level that can be accounted for by emergent, situationally dependent factors? Having the ability to better answer these questions will provide the HRA community with better information about which parameters to incorporate in HRA methods and how to account for their influence in PRA contexts of interest.

These are issues suggested by many of the proposals offered in Section 2. Nevertheless, via the theoretical discussion in Section 2.1, and some of the practical examples provided in the

other portions of Section 2, we have illustrated how Bayesian mathematics and other formal methods such as regression analysis can explicitly deal with different evidence to estimate the quantities of interest and/or provide a better qualitative basis for understanding, modeling, or predicting human performance.

These observations are coupled with the knowledge that Bayesian methods and the use of empirical evidence have been, and continue to be, employed in many applications. This suggests a high confidence that the use of a combination of methods and empirical evidence can be used to inform HRA and improve the current methods used to better understand and predict the most influential factors and the resulting human failure probabilities in contexts of interest to PRA. Hence, we see no significant obstacles that would invalidate or otherwise prevent the use of these methods and empirical evidence for HRA. Further, the use of quantitative methods with empirical evidence seems to hold the promise of improving the following aspects of HEP estimates and the identification of their significant contributors:

- Credibility, because the estimates would be based, in part, on actual experience
- Accuracy, by reducing their associated uncertainties (or at least improve understanding of their uncertainties)
- Validity, by the inclusion of relevant evidence
- Scrutability given the formality of the mathematical constructs.

4.3 Examples of Informing Our Human Performance Assessments

Section 2 provides a number of proposals concerning how the use of evidence and various analytical techniques including Bayesian methods and other treatments (e.g., regression analysis) could be employed. For instance, the research covered in Section 2.2, illustrates a method and approach to collect data that can enable the development of insights into the conditions that give rise to behavior and dependencies in human actions. This research proposes the study of behavioral prediction through formal study of situational factors or PSFs and the contexts that produce systematic variation in them. From this research, we can surmise that behavior is dependent on the situation and systematically influenced by the kinds of PSFs that we attempt to account for in HRA. The manner in which they influence behavior was shown, considering the limited data collected, to be dependent upon the specific accident context, and may be predicated upon the influence of other PSFs. Such interactions appear important in accounting for variability in performance and may also be a key to reducing some of the uncertainty in estimates of reliability if it can be employed in formal models of human reliability. This may argue against more simplistic likelihood models used in Bayesian approaches in favor of more explicit and parametrically elaborate likelihood models. For example, the multiple regression models that were employed to explain variation in performance may also be suitable for use as a likelihood model in a Bayesian formalism.

Section 2.3 describes an approach for using evidence from data such as that available from HERA to improve upon our knowledge of some of the parameters typically employed in HRA methods, using one HRA method for illustration. This approach employs data about PSFs that are associated with human actions that are stored in HERA to develop posterior probability mass functions of the PSFs. With this information, analysts and model developers can assess the extent to which PSFs are associated with human actions in specific contexts. This is a slightly different use of the Bayesian framework for developing posterior probability distributions. Whereas other examples in this report use evidence to estimate the probability of a human

action, Section 2.3 demonstrates how to use evidence to provide information about characteristics of model parameters of an HRA method. Both illustrate the use and value of evidence employed in the Bayesian framework but from different perspectives.

These considerations are mirrored in discussions about the specific types and form of the likelihood functions employed in the Bayesian model. For example, accounting for the hypothesis that an individual's awareness, goals, and activities are constantly changing with regard to specific tasks and factors based in part on what previous successes and failures have occurred, suggests that behaviors may often be correlated. Workers often behave in a certain way in certain environments, or they may exhibit mindset in other cognitive phenomena that result in behavior that is highly dependent on individual, situational, and organizational factors. Such dependencies result in a powerful source of correlation among individual observations that challenges the way we must treat certain forms of human performance data. Section 2.3 demonstrates a general method for addressing the issue of correlated data in a generalized form.

Section 2.4 illustrates ways for conceptualizing human performance in accident contexts and for employing the kinds of data described in other sections. It seems likely that in order to be of the greatest use for PRA, the results of human performance observations, whether from operating experience, simulator studies, or other sources, will require development of *bona fide* probability density functions that relate the human action to a probability scale. The suggestions in Section 2.4 on quantitative treatment of human performance data match well to some sources of data that may be available. They also tie to probability estimation through the suggested development of context anchored probabilities (CAPs) and their generalized form (GCAPs) for describing classes of distributions from similar or related contexts. Interpolation from such contexts to the application at hand may be facilitated through direct estimation, when such distributions become available, or through Bayesian updating using data from similar contexts as forms of partial evidence.

Whether or not the specific proposals in Section 2 are eventually implemented, they do provide pragmatic illustrations of how we might use available human performance evidence along with mathematical formalisms, including Bayesian methods, to improve HRA. These illustrations provide yet further confidence that we can use quantitative methods and empirical evidence.

4.4 Demonstrating the Feasibility of Using Empirical Evidence and Quantitative Techniques for HRA

While we believe the information in the Sections 4.2 and 4.3 provides considerable optimism, if the use of empirical evidence and quantitative techniques is to be employed in HRA, it is necessary to demonstrate the successful implementation of these concepts. With such demonstrations, and hopefully with very usable products from these demonstrations, the HRA community will be more inclined, as a whole, to investigate other ways these concepts can be used to advance our human performance modeling and failure probability predictions.

Hence, while the illustrations in Section 2 suggest ways human performance evidence and quantitative methods might be used, there needs to be a specifically focused attempt to define and implement at least a few meaningful pilot projects whose results would be directly useful in improving one or more of current HRA methods. To do so will require the collection, interpretation, and analysis of relevant human performance information from such sources as actual events, simulator studies, inspections and investigations, and special experiments. Pilot projects can then be established to investigate ways to use this data to inform our HRA models and the human error probabilities they produce. Hence, progress needs to be made on several

fronts to demonstrate the value of applying quantitative techniques to use real data in a more prevalent manner in HRA.

First, and as an ongoing effort, data collection efforts such as the HERA project and nuclear plant simulator experiments at the Halden Research Facility need to continue. While some analysis and interpretation of these data for HRA use is expected and encouraged, the raw data should also be maintained since we cannot know, at this time, how the data might be used and further analyzed or interpreted using quantitative techniques to address issues of interest to HRA modeling and probability estimation. Further, advantage should be taken of data from the international community to complement and add to our available information about human performance in nuclear power plant settings. Given the ability to make use of various forms of data using Bayesian methods, these and other sources of human performance data potentially can all be used.

Second, and closely following the initial phases of the above projects, once a reasonable representation of the possible data have been collected and organized, small proof-of-principle pilot projects need to be defined and implemented to demonstrate how we can use the data in ways similar to the illustrations in this report. It appears that qualitatively informing our HRA models and particularly the structures that they use (typically using PSFs), may hold the best initial promise for uses of the data. Even so, questions, such as how to weigh the different evidence that will be made available by various data collection programs, need to be answered by these pilot projects. Later, as even more data is collected, attempts for more directly estimating human error probabilities can be made. An effort to pull together interested parties with the task of defining and implementing such pilot projects is highly encouraged.

Third, demonstration activities should be organized around accepted issues within the field of practice. The peer review of the workshop discussions indicates that most reviewers believed that the examples chosen and the illustrations that were developed were too academic and were lacking in practical merit. While some of this may be understood in view of the nascence of these methods within the HRA field, the need to better illustrate the information, steps involved in performing analyses, and assumptions that are made must be clear to the practitioner community.

4.5 Possible Steps for Implementing Suggestions to Improve the Use of Empirical Evidence and Quantitative Techniques in HRA

The previous section has indicated several initial steps toward making the use of empirical evidence and quantitative techniques eventually more commonplace in HRA. These are the continuation of existing data collection programs and the need to define and implement some useful proof-of-principle pilot projects using that data.

Assuming successful performance of these steps, we then need to communicate the results of these efforts to the HRA community. In doing so, the value to HRA must be clearly delineated and suggestions for further uses of available data sources should be indicated. One vision of how to communicate this information is to develop a document similar in nature to the *Handbook of Parameter Estimation for PRA* (Atwood *et al.*, 2003). Just as that handbook provides an excellent reference for using data and appropriately selecting formalisms to estimate the reliability of systems and to characterize the uncertainty in analyses, HRA needs a complementing guide or handbook if HRA is to achieve a level of maturity similar to that currently encountered in other aspects of PRA. Such a HRA handbook would conceptually provide, for instance:

- Background information on the use of empirical evidence and quantitative methods for HRA
- Sources and the nature of empirical evidence about human performance
- Ways to use this information in HRA, including the descriptions and results of the pilot projects
- General procedures for using data and quantitative techniques in HRA.

On the basis of the peer review of this workshop summary, several concrete suggestions have been made regarding future efforts to achieve a common working approach to employing empirical data and quantitative formalisms in HRA. These include:

- Identifying the information and sources of information for these kinds of analyses.
- Focusing initial efforts to a limited number of problems that may be well described already in the HRA field
- Developing the quantitative examples and applications as though applying them to a particular problem of interest, and documenting the steps involved
- Describing important assumptions that must be made to use the quantitative technique that analysts may not be familiar with already

In yet the longer term, other steps are proposed. Assuming the pilot projects are successful in demonstrating the beneficial use of quantitative techniques together with empirical data to improve HRA, programs should be developed to specifically validate or otherwise modify, in formal ways, existing HRA methods. Validation, at some level, is a general concern for all HRA methods and it is an issue that needs to be performed if HRA is to be seen as providing credible and most importantly, experience-informed results. It may eventually be possible to supplant existing HRA models with more direct ways to improve our estimation of human error probabilities. Further, with the publication of the aforementioned HRA handbook along with any applicable training, the HRA community could actively use these concepts in future HRA assessments for risk-informed applications.

5. REFERENCES

1. Altham P. "Two Generalizations of the Binomial Distribution". *Applied Statistics* 1978;27(2):162-167.

2. Atwood, C.L., *et al.* "Handbook of Parameter Estimation for Probabilistic Risk Assessment". *NUREG/CR-6823.* U.S. Nuclear Regulatory Commission, Washington, DC, 20555. September 2003.

3. Barriere, M., *et al.* "Technical Basis and Implementation Guidelines for A Technique for Human Event Analysis (ATHEANA)" *NUREG-1624, Rev. 1.* U.S. Nuclear Regulatory Commission, Washington, DC, 2005. May 2000.

4. Bley, D.C., Wreathall, J., and Cooper, S.E. *Common Elements in Operational Events across Technologies*, OECD/NEA Specialists Meeting on Human Performance in Operational Events. Chattanooga, Tennessee, October 13-17, 1997.

5. Comrey, A.L., and Lee, H.B. *A First Course in Factor Analysis* (2nd edition). Hillsdale, NJ: Lawrence Erlbaum Associates, 1992.

6. Cooper, S., *et al.* "Application of ATHEANA: A Technique for Human Error Analysis". *In: Proc IEEE 6th Conf on Human Factors and Power Plants.* Orlando, FL, 1997. p. 9.13-9.17.

7. Drezner Z. and Farnum N. "A Generalized Binomial Distribution". *Communications in Statistics, Theory and Methods* 1993;22(11):3051-3063.

8. Embrey, D.E., *et al.* "SLIM-MAUD: An Approach to Assessing Human Error Probabilities Using Structured Expert Judgment". *NUREG/CR-3518*, U.S. Nuclear Regulatory Commission, Washington, DC, 20555. 1984.

9. Forester, J., *et al.* "Expert Elicitation Approach for Performing ATHEANA Quantification". *Reliability Engineering & System Safety*, v. 83, pp 207-220, 2004.

10. Gertman D., *et al.* "Human Performance Contribution to Risk in Nuclear Power Operating Events". In: *Proceedings of IEEE 7th Conference on Human Factors and Power Plants.* Scottsdale, AZ, 2002. p. 340-349.

11. Gertman, D., *et al.* "The SPAR-H Human Reliability Analysis Method". *NUREG/CR-6883.* US Nuclear Regulatory Commission, Washington, D.C. 2005.

12. Green, S.B. *How Many Subjects Does it Take to do a Regression Analysis?* Multivariate Behavioral Research, v.26, pp. 449-510, 1991.

13. Groen, F.J., and Mosleh, A. "Foundations of Probabilistic Inference with Uncertain Data". *International Journal of Approximate Reasoning*, 39, 2005, p.49-83

14. Haldar, A., and Mahadevan, S. *Probability, Reliability and Statistical Methods in Engineering Design.* New York: Wiley, 2000.

15. Hallbert, B., *et al.* "The Use of Empirical Data Sources in HRA". *Reliability Engineering & System Safety*, 2004;83(2):139-143.

16. Hallbert, B.P., *et al.*. "Using Information from Operating Experience to Inform Human Reliability Analysis". In Spitzer, C., Schmocker, U., & Dang, V.N. (Eds.), *Proceedings of Probabilistic Safety Assessment and Management*, Springer: June 14-18, 2004, pp. 977-984.

17. Hallbert, B.P., *et al.* "A Study of Staffing Levels for Advanced Reactors". *NUREG/IA-0137*, U.S. Nuclear Regulatory Commission, Washington, D.C., November 2000.

18. Hallbert, B.P., *et al.* "Human Event Repository and Analysis (HERA) System Overview". *NUREG/CR-6903, Volume 1*, U.S. Nuclear Regulatory Commission, Washington, D.C., June 2006.

19. Hanson, D.J., *et al.* "Evaluation of Operation Safety at Babcock and Wilcox Plants Volume 1 – Results Overview". *NUREG/CR-4966, Volume 1*, U.S. Nuclear Regulatory Commission, Washington, D.C., October 1987.

20. Hollnagel, E. *Cognitive Reliability and Error Analysis Method CREAM*. New York: Elsevier, 1998.

21. Jeffreys, H. *Theory of Probability* (3rd ed.) London: Oxford University Press, 1961.

22. Johnson, N., Kotz, S., and Balakrishnan, N. *Discrete Multivariate Distributions.* New York: John Wiley, 1997.

23. Kirwan, B., *et al.* "Nuclear Action Reliability Assessment (NARA): A Data-based HRA Tool". In Spitzer, C., Schmocker, U., & Dang, V.N. (Eds.), *Proceedings of Probabilistic Safety Assessment and Management*, Springer: June 14-18, 2004, pp. 1206-1212.

24. Leonard, T., Hsu, and J., Gill, R. *Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers*. New York: Cambridge University press, 1999.

25. Kirk, M.E., *et al.* "Technical Basis for Revision of the Pressurized Thermal Shock (PTS) Screening Limit in the PTS Rule (10CFR50.61) Summary Report," *NUREG-1806*, U.S. Nuclear Regulatory Commission, Draft 2005.

26. Mosleh, A. and Apostolakis, G. "Models for the Use of Expert Opinions", in *Low Probability, High Consequence Risk Analysis: Issues, Methods and Case Studies,* R.A. Waller, and V.T. Covello, (Eds.) Plenum Press, New York, 1984.

27. Mosleh, A. and Apostolakis, G. "The Development of a Generic Database for Failure Rates", *Proceedings of the International Topical Meeting on Probabilistic Safety Methods and Applications*, San Francisco, California, February 24-March 1, 1985.

28. Mosleh, A. "Expert-to-Expert Variability and Dependence in Estimating Rare Event Frequencies", *Reliability Engineering and System Safety*, 38 (1992) 47-57.

29. Mosleh, A. "Hidden Sources of Uncertainty: Judgment in Collection and Analysis of Data", *Nuclear Engineering and Design*, 93 (1986) 187-198

30. *R-DAT: Reliability Data Analysis Tool*, Commercial Software by Prediction Technologies Inc. http://www.prediction-technologies.com/products/r-dat.brochure.pdf, Dec. 2007

31. Reer, B., "Sample Size Bounding and Context Ranking as Approaches to the HRA Data Problem", *Reliability Engineering & System Safety*, vol. 83, pp. 265-274, 2004.

32. Samanta, P., *et al.* "Risk Sensitivity to Human Error". *NUREG/CR-5319*, U.S. Nuclear Regulatory Commission, Washington, D.C., 1989.

33. SECY-04-0118, July 13, 2004. "Plan for the Implementation of the Commission's Phased Approach to Probabilistic Risk Assessment Quality".

34. Sträter O. "Considerations on the Elements of Quantifying Human Reliability". *Reliability Engineering Systems Safety* 2004;83(2):255-264.

35. Sträter O. *Evaluation of Human Reliability on the Basis of Operational Experience.* GRS-170. Koln, Germany, 2000.

36. Swain, A.D. and Guttmann, H.E. "Handbook of Human Reliability Analysis with Emphasis on Nuclear Power Plant Applications". *NUREG/CR-1278*. Washington, DC: USNRC, 1983.

37. Tabichnick, B.G. and Fidell, L.S. *Using Multivariate Statistics* (4th Edition). Boston, MA: Allyn and Bacon, 2001.

38. U.S. Nuclear Regulatory Commission, "Use of Probabilistic Risk Assessment Methods in Nuclear Activities: Final Policy Statement," *Federal Register*, Vol. 60, pg. 42622 (60 FR 42622), Washington, DC, August 16, 1995.

39. Van Ophem, H. "A General Method to Estimate Correlated Discrete Random Variables". *Econometric Theory* 1999;15:228-237.

40. Williams, J.C. "A Databased Method for Assessing and Reducing Human Error to Improve Operational Performance". In: *Proceedings IEEE 4th Conf on Human Factors and Power Plants.* New York: Institute of Electronic and Electrical Engineers, 1988. p. 436–50.

41. *WINBUG*, Open Source Software, 2003

42. Wong, S., Higgins, J., and O'Hara, J. "Risk Sensitivity to Human Error in the LaSalle PRA". *NUREG/CR-5527*, U.S. Nuclear Regulatory Commission, Washington, D.C., 1990.

43. Yoshikawa, H. "An Experimental Study on Estimating Human Error Probability (HEP) Parameters for PSA/HRA by Using Human Model Simulation". *Ergonomics* 1999;42(11):1588-1595.

NRC FORM 335
(9-2004)
NRCMD 3.7

U.S. NUCLEAR REGULATORY COMMISSION

BIBLIOGRAPHIC DATA SHEET

(See instructions on the reverse)

1. REPORT NUMBER
(Assigned by NRC, Add Vol., Supp., Rev., and Addendum Numbers, if any.)

NUREG/CR-6949

2. TITLE AND SUBTITLE

The Employment of Empirical Data and Bayesian Methods in Human Reliability Analysis: A Feasibility Study

3. DATE REPORT PUBLISHED

MONTH	YEAR
December	2007

4. FIN OR GRANT NUMBER

JCN: Y6496

5. AUTHOR(S)

Edited by:
Bruce P. Hallbert, Idaho National Laboratory
Alan Kolaczkowski, Science Applications International Corporation

6. TYPE OF REPORT

Technical

7. PERIOD COVERED *(Inclusive Dates)*

8. PERFORMING ORGANIZATION - NAME AND ADDRESS *(If NRC, provide Division, Office or Region, U.S. Nuclear Regulatory Commission, and mailing address; if contractor, provide name and mailing address.)*

Idaho National Laboratory
PO Box 1625
Idaho Falls, ID 83415-3605

Science Applications International Corporation
10260 Campus Point Dr.
San Diego, CA 92121

9. SPONSORING ORGANIZATION - NAME AND ADDRESS *(If NRC, type "Same as above"; if contractor, provide NRC Division, Office or Region, U.S. Nuclear Regulatory Commission, and mailing address.)*

Same as above

10. SUPPLEMENTARY NOTES

Erasmia Lois, NRC Project Manager

11. ABSTRACT *(200 words or less)*

The U.S. Nuclear Regulatory Commission (NRC) is addressing issues related to the quality of Probabilistic Risk Assessment (PRA), including issues related to human reliability analysis (HRA) performed as part of PRA. Among the issues of concern is an inadequate use of human performance data in the estimation of human error probabilities (HEPs), as well as in testing or otherwise validating underlying models used in HRA to predict human performance under cognitively demanding conditions. In order to address issues related to the use of human performance data in HRA, the NRC is developing the Human Event Repository and Analysis (HERA) database (NUREG/CR-6903). In addition, in August 2005, the NRC hosted an expert workshop on the use of Bayesian and other quantitative formalisms in conjunction with empirical data, such as that available in HERA, to improve both the estimation of human error probabilities and the underlying assumptions and quantitative algorithms employed by different HRA methods.

This report contains a collection of papers that were produced as a result of the workshop. It also summarizes the peer review comments of a draft version of this report, includes conclusions about the feasibility of using empirical data and quantitative methods for HRA, and provides suggestions on how to proceed to address the issues under consideration.

12. KEY WORDS/DESCRIPTORS *(List words or phrases that will assist researchers in locating the report.)*

Human Reliability Analysis (HRA)
Probabilistic Risk Assessment (PRA)
Bayesian

13. AVAILABILITY STATEMENT

unlimited

14. SECURITY CLASSIFICATION

(This Page)
unclassified

(This Report)
unclassified

15. NUMBER OF PAGES

16. PRICE

PRINTED ON RECYCLED PAPER

Printed
on recycled
paper

Federal Recycling Program

UNITED STATES
NUCLEAR REGULATORY COMMISSION
WASHINGTON, DC 20555-0001

OFFICIAL BUSINESS