U.S.NRC
United States Nuclear Regulatory Commission
Protecting People and the Environment

NUREG-2127

The International HRA Empirical Study

Lessons Learned from Comparing HRA Methods Predictions to HAMMLAB Simulator Data

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
Publications Branch
Washington, DC 20555-0001
E-mail: DISTRIBUTION.RESOURCE@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, 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-2127

The International HRA Empirical Study

Lessons Learned from Comparing HRA Methods Predictions to HAMMLAB Simulator Data

Manuscript Completed: December 2012
Date Published: August 2014

Prepared by:
John Forester[1,5], Vinh N. Dang[2], Andreas Bye[3],
Erasmia Lois[4], Salvatore Massaiu[3], Helena Broberg[3],
Per Øivind Braarud[3], Ronald Boring[5], Ilkka Männistö[6],
Huafei Liao[1], Jeff Julius[7], Gareth Parry[4,9], Pamela Nelson[8]

[1]Sandia National Laboratories, USA
[2]Paul Scherrer Institute, Switzerland
[3]Institute for Energy Technology, OECD Halden Reactor Project, Norway
[4]U.S. Nuclear Regulatory Commission, USA
[5]Idaho National Laboratory, USA
[6]VTT, Finland
[7]Scientech, USA
[8]Universidad Nacional Autónoma de México, Mexico
[9]ERIN Engineering

Erasmia Lois, NRC Project Manager

NRC Job Code N6129

Office of Nuclear Regulatory Research

ABSTRACT

This report documents the overall conclusions and lessons learned from the International Empirical HRA Study, documented in NUREG/IA-0216, Vols. 1-3, as well as in Halden Reactor Project reports (HWR-844, HWR-915, and HWR-951). The International HRA Empirical Study has developed an empirically based understanding of the performances, strengths, and weaknesses of a set of HRA methods through comparisons between human reliability analysis (HRA) predictions of crew performance in simulated scenarios and actual crew performance outcomes. The simulator experiments were conducted at the Organization for Economic Co-Operation and Development (OECD) Halden Reactor Project's Human-Machine Laboratory (HAMMLAB), Halden, Norway. This is a large-scale study; organizations from ten countries, representing industry, regulators, and the research community, participated.

This report summarizes the findings and insights for the individual HRA methods empirically tested in this study, as well as the overall observations and conclusions regarding the HRA discipline as a whole. In addition, it summarizes the methodology developed to allow comparisons between HRA results and crew performance and its merits for future studies, and reflects on individual analyses of crew simulator performance, providing evidence for improving both HRA practices and plant safety. It has also been published as a Halden report (HPR-373).

TABLE OF CONTENTS

ABSTRACT iii
LIST OF FIGURES ix
LIST OF TABLES ix
EXECUTIVE SUMMARY xi
ACKNOWLEDGMENTS xxv
ACRONYMS AND ABBREVIATIONS xxix
1. INTRODUCTION 1
1.1 Intended readers 1
1.2 Background and motivation (aims) 1
1.3 Scope 2
1.4 Overview of the study design, tasks 3
1.5 Study organization, participants, and roles 4
1.6 Phases of the Empirical Study 4
2. METHODOLOGY 7
2.1 Simulation design 7
2.1.1 Data Collection and Participants 7
2.1.2 Defining Human Failure Events 8
2.1.3 Scenarios and HFE definitions 9
2.2 Empirical Data Analysis Methodology 11
2.2.1 Raw Data 11
2.2.2 Crew - Level Analysis 12
2.2.3 Crew Performance and HFE Failures 12
2.2.4 Operational Description 12
2.2.5 PSF Assessment: Operational and HRA ratings 13
2.2.6 HRA PSF ratings 17
2.2.7 Difficulty and Ranking of the HFEs 18
2.3 HRA predictive analyses performed for the study 19
2.3.1 HRA analysis inputs 19
2.3.2 Reporting of HRA analyses and predicted outcomes 19
2.4 Methodology for assessment of HRA methods 19
2.4.1 Assessment criteria 20
2.4.2 Structure of summary assessment of each method 20
2.4.3 Comparison of method’s qualitative predictions 23
2.4.4 Comparison of method’s qualitative predictions in terms of drivers 23
2.4.5 Comparison of method’s qualitative predictions in terms of operational expressions 23
2.4.6 Comparison of quantitative predictions (including ranking) 24
2.4.7 Assessment of traceability 25

2.4.8 Assessment of adequacy of method guidance 25
2.4.9 Insights for error reduction 25
3. EMPIRICAL HAMMLAB RESULTS 27
3.1 Variability of performance 27
3.2 HFE difficulty ranking 28
3.2.1 SGTR difficulty ranking of HFEs 28
3.2.2 LOFW difficulty ranking of HFEs 32
3.3 Operational descriptions and PSF assessments 34
3.3.1 LOFW HFE-1B (B&F before dryout in complex scenario) 34
3.3.2 SGTR HFE-3B (depressurization in complex scenario) 39
4. OVERALL QUANTITATIVE RESULTS 43
4.1 Role of the quantitative data 43
4.2 LOFW, Overall quantitative results from HRA method predictions 44
4.3 SGTR, Overall quantitative results from HRA method predictions 45
4.4 General discussion of the quantitative results of the HRA predictions 46
4.5 The empirical HEPs (Bayesian results) (LOFW part) 47
4.6 LOFW, Predicted HEPs vs. empirical HEPs (Bayesian results) 48
4.7 SGTR, predicted HEPs vs. empirical HEPs (Bayesian results) 49
5. HRA METHOD ASSESSMENTS 51
5.1 Overall Assessment of ASEP (UNAM) 51
5.1.1 Strengths 51
5.1.2 Weaknesses 52
5.2 Overall Assessment of ASEP/THERP (NRC) 54
5.2.1 Strengths 54
5.2.2 Weaknesses 55
5.3 Overall assessment of ATHEANA (NRC) 56
5.3.1 Strengths 57
5.3.2 Weaknesses 58
5.4 Overall Assessment of CBDT+THERP (EPRI) 59
5.4.1 Strengths 59
5.4.2 Weaknesses 60
5.5 Overall Assessment of CESA-Q 61
5.5.1 Strengths 61
5.5.2 Weaknesses 62
5.6 Overall Assessment of CREAM (NRI) 63
5.6.1 Strengths 63
5.6.2 Weaknesses 64
5.7 Overall Assessment of Decision Trees (DTs) + ASEP (NRI) 66
5.7.1 Strengths 67

5.7.2 Weaknesses 67
5.8 Overall Assessment of ENHANCED BAYESIAN THERP (VTT) 68
5.8.1 Strengths 69
5.8.2 Weaknesses 69
5.9 Overall Assessment of HEART (Ringhals) 70
5.9.1 Strengths 71
5.9.2 Weaknesses 71
5.10 Overall Assessment of K-HRA (KAERI) 72
5.10.1 Strengths 72
5.10.2 Weaknesses 73
5.11 Overall assessment of MERMOS (EDF) 74
5.11.1 Strengths 74
5.11.2 Weaknesses 75
5.12 Overall Assessment of PANAME (IRSN) 76
5.12.1 Strengths 76
5.12.2 Weaknesses 77
5.13 Overall Assessment of SPAR-H (INL) 78
5.13.1 Strengths 78
5.13.2 Weaknesses 78
5.14 Overall assessment of SPAR-H (NRC) 80
5.14.1 Strengths 80
5.14.2 Weaknesses 81
6. CONCLUSIONS ABOUT HRA AND THE HRA METHODS 85
6.1 Explaining variability in results 85
6.2 Summary of Lessons Learned about the HRA Methods 86
6.3 Specific recommendations for improving guidance, practice, and methods 92
6.3.1 Address both qualitative and quantitative aspects of HRA 92
6.3.2 Enhancing qualitative analysis 93
6.3.3 Improving guidance, practices, and methods 95
6.3.4 Towards hybrid HRA methods 97
7. CONCLUSIONS ON THE USE OF EMPIRICAL HRA DATA AND BENCHMARKING .. 99
7.1 Feasibility of benchmarking against simulator data 99
7.2 Usefulness and acceptance of results 99
7.3 Assessment of benchmarking methodology features 101
7.4 Future HRA studies utilizing simulator data 102
8. ACHIEVEMENTS AND OVERALL CONCLUSION 105
9. REFERENCES 109
APPENDIX A DETAILED SCENARIO DESCRIPTION AND HFE DEFINITIONS 111

LIST OF FIGURES

Figure 1-1 Overview of the HRA Empirical Study 3
Figure 4-1 LOFW, range of predicted mean HEPs for the HRA methods, in decreasing order of difficulty 45
Figure 4-2 SGTR, range of predicted mean HEPs of the HRA methods 46
Figure 4-3 Bayesian posterior distributions resulting from update of Jeffrey's prior 47
Figure 4-4 Bayesian confidence bounds of the LOFW empirical HEPs vs. all predicted HEPs 49
Figure 4-5 Bayesian confidence bounds of the SGTR empirical HEPs vs. all predicted HEPs 50

LIST OF TABLES

Table 1-1 Phases of the Empirical Study. 5
Table 2-1 Performance-shaping factors – definitions. 14
Table 2-2 Structure of assessment summary for each HRA method. 22
Table 3-1 Summary table of ranked HFEs and their difficulties in SGTR. 29
Table 3-2 Summary table of ranked HFEs and their difficulties in LOFW. 33
Table 3-3 Operational descriptions of the crews in LOFW HFE-1B. 35
Table 3-4 Overall PSF evaluation for LOFW HFE-1B. 36
Table 3-5 Operational modes observed in the SGTR HFE-3B. 40
Table 3-6 Overall PSF assessment for SGTR HFE-3B. 41

EXECUTIVE SUMMARY

Background

Diverse human reliability analysis (HRA) methods are currently available to treat human failure in probabilistic risk assessments (PRAs). Given the differences between the methods and their associated models, there is substantial interest in assessing HRA methods, and ultimately in validating the approaches and models underlying them. Such a validation is warranted to assess the credibility of HRA results when decision makers have to use those results to make risk-informed decisions.

To that end, the Office of Nuclear Regulatory Research (RES) of the U.S. Nuclear Regulatory Commission (NRC) supported the initiation and execution of an International HRA Empirical Study (hereafter called "Empirical Study"). The study is an international collaborative effort, and involves the use of the Organization for Economic Co-Operation and Development (OECD) Halden Reactor Project's HAMMLAB (HAlden huMan-Machine LABoratory) research simulator, a full-scope nuclear power plant (NPP) simulator located in Halden, Norway. It aims to develop an empirically based understanding of the performance, strengths, and weaknesses of different HRA methods. The empirical basis is developed through experiments performed at the HAMMLAB simulator, with real crews responding to simulated initiating events (IEs) based on those included in PRAs.

The scope of the study is limited to HRAs for internal events during full-power operation of current light water reactors, and focuses on the control room personnel actions. Thus, the results in this report are mainly valid for control room actions within a Level 1 PRA. Nonetheless, it is likely that the findings about the methods and HRA in general will be relevant to any HRA application. As the scenarios reflect the typical use of emergency operating procedures (EOPs) after an IE, it is believed that the results may be generalized to other IEs.

Objective of this Report

The objective of this report is to summarize the findings and insights regarding individual HRA methods empirically tested in this study, as well as the overall insights and conclusions of the study. It provides overall perspectives on the HRA discipline, as well as for individual methods. Furthermore, it provides evidence about human performance in simulated adverse conditions and supports improvements in NPP operations, including plant changes and improvements in procedures and training. NUREG/IA-0216, Vol. 1-3 (corresponding to Halden reports HWR-844, HWR-915, and HWR-951) [1], [2], [3] provide details on the technical basis for the findings and conclusions in this report, and will provide readers with additional insights regarding the information presented in this report.

Overview of the Study Design and Methodology

The International HRA Empirical Study focused on control room personnel actions taken in response to initiating events modeled in a PRA. The simulator experiments were conducted in the HAMMLAB pressurized water reactor (PWR) simulator, a full-scope simulator of a three-loop Westinghouse plant, with fourteen crews that operate two units at the home plant. Thirteen HRA teams participated in the study, representing a wide range of organizations and using a variety of methods. The number and variety of HRA teams and methods helped us to develop a good understanding of the methods and their applications.

The study comprised four tasks:

- **Task 1.** The definition of the scenarios and of the human failure events (HFEs) to be analyzed and the compilation of the inputs for the HRA analyses.

- **Task 2.** The production of the empirical or reference data for the comparison, starting with the collection of raw data in simulator experiments conducted at the OECD Halden Reactor Project's HAMMLAB (HAlden huMan-Machine LABoratory) research simulator facility and followed by the analysis of this data.

- **Task 3.** The analysis of the HFEs with HRA methods, which produced the predicted outcomes.

- **Task 4.** The comparison of the predicted outcomes with the empirical data and the development of insights for improving HRA methods and practices.

These tasks were carried out by the following groups:

- **Halden experimental staff** (Tasks 1, 2): The simulator sessions were conducted in the OECD Halden Reactor Project's HAMMLAB research simulator facility. The staff was responsible for the scenario development and the collection and analysis of the experimental data.

- **Operator crews** (Task 2): A set of licensed operator crews responded to a series of scenarios in the HAMMLAB simulator. Each crew responded to four scenarios, which each consisted of a base and a "complex" variant of two scenario types.

- **HRA teams** (Task 3): Each team applied an HRA method to obtain predictions for the HFEs in the scenarios defined for the study. Organizations representing industry, regulators, and the research community have participated.

- **Assessment group** (Overall organization and Tasks 1, 4): This group was responsible for organizing and implementing the study. It collaborated with Halden staff to design the experiments, prepared the analysis inputs for the HRA teams, and answered their requests for additional information. After the HRA teams delivered their analyses, this group reviewed and summarized the predicted outcomes before performing the actual comparison.

To avoid bias in the comparison, a "blind" study protocol was used. The operating crews had no prior knowledge of the scenarios; the assessment group did not receive any information about the actual crew performances until after their review and evaluation of the HRA submittals, and the Halden staff analyzed and documented the crew performance data without any knowledge of the HRA predictions.

The simulator experiment comprised two steam generator tube rupture (SGTR) scenarios and two loss of feedwater (LOFW) scenarios. The SGTR scenarios included nine HFEs, while the LOFW included six HFEs. The study was carried out in three phases. In Phase 1, the Pilot, documented in NUREG/IA-0216 Volume 1 (HWR-844) [1], the study methodology was developed and tested. Improvements were made to the method and incorporated into Phase 2, documented in NUREG/IA-0216, Volume 2 (HWR-915) [2]. In Phase 2, seven SGTR HFEs, along with those examined in Phase 1, were used to benchmark HRA methods. In Phase 3, documented in NUREG/IA-0216, Volume 3 (HWR-951) [3], HRA results were compared to the empirical results for the HFEs of the LOFW scenarios. These reports also include the general observations made in each phase.

Empirical Data Development

To collect and interpret the empirical data for comparison, a data collection team of experts in experimental psychology, human factors, NPP operations and operator training, and PRA/HRA developed a multifaceted methodology, which involved the following:

1. Collecting raw data in operator logs, audio/video recordings of crews' activities, and crew interviews.

2. Crew-level analysis to determine whether and to what degree crews accomplished the tasks related to the different HFEs. In this task, crew performance is investigated at a detailed operational level.

3. Determination of crew failures associated with various HFEs. Failure determination was based on a comparison between the information gathered through crew/scenario reviews and the quantitative performance information (e.g., performance times, SG level).

4. Development of operational descriptions, which summarize how the crews handled the various tasks involved in the HFEs.

5. Identification of performance-shaping factors (PSFs) in crew performance and PSF ratings, that is, evaluating the presence and the strength of a PSF as an underlying performance driver.

6. Ranking of the HFEs based on the empirical evidence on the level of difficulty involved in the diagnosis and execution of the associated human actions.

These steps were performed iteratively to ensure an accurate analysis and interpretation of the empirical data for the purpose of benchmarking HRA methods.

HRA method assessment approach

HRA methods were assessed primarily from a qualitative analysis perspective; however, quantitative results were also considered in the evaluation. The evaluation was based on the following desirable HRA attributes:

- The predictive power of a method (i.e., the extent to which the method application predicted the empirical evidence from both qualitative and quantitative perspectives).

- The traceability of the qualitative and quantitative analyses.

- The adequacy of the guidance provided by each method for qualitative and quantitative analysis.

- The usefulness of qualitative and quantitative results in human error reduction.

Qualitative predictive power assessments examined the extent to which the methods provided the capability, as well as the extent to which the analysts used their methods to perform adequate qualitative analyses so that their results reflected the empirical evidence. It included:

- PSF Assessments: Evaluated how well the method applications predicted the specific performance issues and drivers observed in the reference data.

- Operational expression assessments: Evaluated how well the method applications predicted the failure mechanisms (i.e., the reason for the difficulties (or ease) with which the crews performed the tasks associated with each HFE, and how these difficulties were expressed in operational and scenario-specific terms).

Quantitative predictive power assessments examined the extent to which the absolute and relative values of the human error probabilities (HEPs) produced were supported by the empirical data. The limitations of the quantitative reference data (e.g., the limited numbers of trials on which to base HEP estimates) were accounted for in this assessment. It is worth noting that these limitations were not uniform across the HFEs; they were more significant for the less challenging HFEs, in which the HEP estimates were based on a relatively small number of trials, than for the challenging HFEs, in which crews were observed to have performance difficulties, and in some cases failed the actions. Quantitative assessments were based on comparisons with the observed data in the following ways, listed in order of priority:

- Potential optimism of the most difficult HFEs (i.e., did they predict relatively low HEPs for the most difficult HFEs?).

- Consistency of the ranking of the HFEs (based on estimated HEPs) with the difficulty rankings based on the empirical evidence, accounting for observations of crew performance and failure rates where applicable.

- Predicted HEPs relative to the confidence and uncertainty bounds of the reference data.

- Quantitative differentiation of the HFEs by HEP (i.e., were the method application results sensitive to the apparent magnitude of the variations in the identified difficulty of the HFEs?).

Evaluations of HRA submittals based on the desirable HRA attributes *traceability*, *adequacy of guidance*, and *capability to produce insights for error reduction* were based mainly on examination of the submitted analyses; where appropriate, comparisons of HRA results to reference data (e.g., did the method's guidance appear to help the analysts capture relevant aspects of the data?) were also performed. The desirable attribute of *repeatability* (i.e., the ability to reproduce the results if a different analyst performs the analysis using the same method) was not examined. Such an assessment would require multiple HRA teams using the same method.

Empirical results

The HAMMLAB research facilities were used to simulate PRA-based scenarios. Some scenarios were designed to challenge the crews. The objective of designing easy (base case) and complex scenario variants was to produce variability in crew performance outcomes, which makes it possible to observe both failure and success, the difficulties encountered, and the quality of performance. Producing empirical evidence of variability in the quality of crew performance provided a basis to go beyond mere failure counting; it allowed the study to examine and address a broader spectrum of performance issues, as well as to rank the HFEs.

For example, the SGTR base case scenario involved a very typical SGTR initiating event in which crews had to identify and isolate the ruptured steam generator (SG), cool the reactor, and depressurize the reactor coolant system (RCS). The complex case involved the same initiating event, with the complication of failure of secondary radiation indications. As shown below, the HFE definitions were exactly the same, but the conditions under which they would have to be performed were different.

- HFE-1A: Failure of the crew to identify and isolate the ruptured SG in the base SGTR
- HFE-1B: Failure of the crew to identify and isolate the ruptured SG in the complex SGTR
- HFE-2A: Failure of the crew to cool down the RCS expeditiously in the base SGTR
- HFE-2B: Failure of the crew to cool down the RCS expeditiously in the complex SGTR
- HFE-3A: Failure of the crew to depressurize the RCS expeditiously in the base SGTR
- HFE-3B: Failure of the crew to depressurize the RCS expeditiously in the complex SGTR

During data analysis, the crews' simulator results were first characterized in terms of successes and failures. For example, all crews accomplished HFE-1A, but 7 out of 14 crews did not accomplish 1B; however, the analysis went much further in developing operational descriptions, assessing PSFs and ranking the HFEs based on their observed difficulty. Examples are provided in the following paragraphs.

The "operational descriptions" used to represent crew performance were developed to compare empirical evidence with method predictions. The analysis of crew performance exemplified how HRA concepts, such as diagnosis, can be observed in actual crew responses to initiating events and then documented. Furthermore, the results of these observations and associated operational descriptions pointed out that HRA practices in which cognitive demands on operators are frequently not well examined can cause the analysts to miss important impacts on performance. For instance, from an operational story from the complex LOFW scenario, in which the SG level indications were malfunctioning (Table 3-3, 4a): "The crews do not identify the abnormal SG levels, although they monitor the SG levels trends." In the Comments Section it is explained that "[E]ven though SG WR level trends are displayed, the crews are absorbed by the procedure work on restoring feed water to the SGs, and do not react to or stop to analyze the SG levels situation before dry out, and hence did not diagnose the real levels in the SGs."

The PSF analysis produced evidence of the presence and strengths of various PSFs in crew responses. For example, the “adequacy of time” PSF was judged not to be a negative driver for HFE-1B (Table 3-4) because “[I]f the criterion to start bleed and feed is detected, there is adequate time to do this before SGs are empty. However, depressurizing the SGs in procedure FR-H.1 step 7 will reduce the time to when the SGs are empty.” However, the “indication of conditions” PSF was judged to be a main negative driver because “[T]he criterion to start bleed and feed on SG levels is masked by two of three SG levels being failed.”

The empirical data analysis showed a large degree of variability in the way different crews responded to the scenarios and the necessary actions in both the SGTR and the LOFW scenarios. This finding established, among other things, that the frequent HRA assumption that the comprehensive emergency response guidelines will lead most crews to perform the tasks within strict boundaries may not always be the case. Variability in crew performance is generally not explicitly handled in most HRA methods, so analysts implicitly model the “well trained” scenario and assume the crews will generally behave the same; the possible impact if a crew takes a different path in executing the procedures is not addressed in most methods. However, the empirical observations confirmed that crew variability contributes to success or failure, and that analysts should be cognizant of the potential for crew variability and should strive to consider it in their analysis (e.g., assess different failure paths and their likelihoods).

Documentation and analysis of crew performance at the level achieved in this study goes beyond addressing HRA method assessment. It provides a basis for addressing other areas dealing with understanding and improving human performance. For example, the PSF concept is used frequently in many human performance-related disciplines; however, there is a lack of consistency and clarity in the definitions and the use of PSFs. The PSFs used in this study had to be defined in operational terms so that it could be determined whether a factor such as “scenario complexity” was observed, and, if so, whether it was a performance driver. As a result, although not necessarily mutually exclusive, steps were taken to clearly define and operationalize the PSFs; furthermore, an empirical basis for their “presence” was established. Based on the PSF definitions, analysts were able to identify factors contributing significantly to the crew behavior; thus, the usefulness of the PSF definitions in explaining crew behavior was to some extent validated, and provided a means of evaluating the HRA results. Similar improvements in the use of other HRA concepts were achieved. Establishing clarity for the fundamental HRA concepts is a major achievement of this study. Other obvious examples are the identification and evaluation of potential modifications to plant and operator training. It is also possible to use the empirical results to ask what-if questions for new reactor designs since the results provide detailed analyses of crew performance in baseline, as well as very challenging conditions.

HRA Method Assessment Results

The HRA submittals were evaluated from both qualitative and quantitative perspectives. A summary of the results of individual method assessments is provided in the body of this report. The main questions addressed by this study and its findings are summarized below. Only a summary of the results across all methods is provided in the Executive Summary.

Were the HEPs obtained with these HRA methods consistent with the evidence (quantitative predictive power)?

The predictive power of methods, based on the quantitative results, was addressed by considering such issues as potential optimism in HEP estimations for the most difficult HFEs, consistency of the HEPs with the difficulty ranking of the HFEs in terms of both order and magnitude, relationship between the HEPs and the confidence or uncertainty bounds of the reference data, and quantitative differentiation of the HFEs.

The major finding from the quantitative results is that there is significant variability from method to method in the HEPs produced for the same HFEs despite the care taken to provide very detailed descriptions of the scenarios and very strict HFE definitions. Variability in HEP estimation was present for both easy and difficult HFEs. Specific findings include:

- *Optimistic HEPs for the most difficult HEFs:* Despite the small sample size, the empirical evidence for the most difficult HFEs, where one or more failures were observed, is relatively strong. In this case, the analyses were expected to produce appropriately high HEPs. However, the HEPs obtained in some analyses appeared to be optimistic compared to the failure rates and to the evaluations of difficulty. The optimism seems to be related to either not addressing diagnosis in the analysis of these events or to an apparent disconnection between qualitative and quantitative analysis (i.e., some analyses identified many of the crew factors observed in the empirical data as having a negative impact on performance, but did not account for the impact of these factors on the corresponding HEPs).

- *Ranking of HEPs*: In many cases, the method applications did not produce HEPs with a ranking reflecting the relative difficulty levels of the HFEs observed in the evidence. This was more evident for the HFEs of the SGTR scenarios than for those of the LOFW scenarios. Ranking deserves attention in HRA practices as a whole because one of the fundamental HRA goals is to ensure that the HEPs produced have internal consistency (i.e., that the relative values of the HEPs produced reflect the relative difficulty of the HFEs).

- *Range and differentiation of HEPs*: The analyses did not always adequately discriminate between the difficulty levels, even in cases where they produced an appropriate ranking. A number of the analyses produced a narrow range of HEPs; in some cases, the HEPs were less than an order of magnitude different for very different levels of HFE difficulty. Although analysts would frequently recognize differences in the difficulty of accomplishing the actions from a qualitative perspective, they did not seem to verify whether the range of produced HEPs reflected these differences as a whole. Differentiation between the HEPs in HRA is as important as the relative ranking of the HEPs; it ensures that the HEPs produced are relatively coherent.

- *Conservative or realistic HEPs*: None of the methods consistently produced high (or low) HEPs for the set of HFEs. In other words, none of the methods were systematically more conservative or optimistic than the other methods. This finding should be considered in regulatory applications in which it is frequently assumed that a particular method by its nature produces conservative bounding values, and can therefore be used as a “screening tool.”

- *Comparisons of HEPs to the reference data’s confidence or uncertainty bounds*: The uncertainty bounds of the reference data were broad for the easier HFEs and relatively narrow for the more difficult HFEs, reflecting the relative strength of the evidence for small HEPs vs. large HEPs. As a result, comparing the predicted HEPs against the bounds to determine whether they were consistent with the evidence was generally more revealing for the difficult HFEs than for the easier HFEs. In this context, there were notable misses for the difficult HFEs (optimistic predictions) and some, but fewer, misses for the easier HFEs. In both cases, such outliers were subject to detailed examination of the analyses of the corresponding HFEs, associated method features, and analysis assumptions.

Another aspect of quantification which was addressed only in the LOFW scenarios is *treatment of dependencies*, which is an important issue in HRA. Failure in an action is typically considered to have a potential negative effect on a subsequent action. However, in the LOFW scenarios, all crews who failed to implement bleed and feed (B&F) before dryout were subsequently able to initiate B&F before core damage. Most HRA teams, consistent with common practices, analyzed the conditional HFE using the Technique for Human Error Rate Prediction (THERP)-based dependence model and obtained HEPs that were pessimistic compared to the empirical data. The empirical data suggest that, when considering dependencies, it may be important to balance potential positive dependence effects from an initial failure with the impact of new information and changes in conditions. *It is concluded that significant improvement in the treatment of dependence is needed for all methods*. In particular, it would appear that analysts need to understand the dynamic nature of the plant status evolution and the information flow and procedural guidance that the evolution entails, rather than the current emphasis on factors like same crew, same procedure, or same location, which focus on more static aspects.

Did the HRAs identify the performance issues identified in the empirical data (qualitative predictive power)?

Qualitative predictive power assessments examined the extent to which methods provide capability, as well as the extent to which analysts used their methods to perform adequate qualitative analyses so that their results reflect the empirical evidence. The qualitative analyses were mainly examined from the following perspectives:

- Ability to capture the significant influences on behavior and the degree to which this leads to an understanding of the underlying dynamics of the scenario and driving factors

- The depth of qualitative analysis acceptable to the method

- The ability of the method to accommodate the analysts' knowledge and understanding of the HFE and scenario context

Insights obtained on the strengths and weaknesses of the methods are summarized below:

- *Handling of crew cognition tasks*: Crews continually perform cognitive and information-gathering activities when responding to an event and cognition is a major contribution to variability in the results of crew performance. In HRA, however, it is frequently assumed that the comprehensive emergency response guidelines will lead most crews to perform the tasks within strict boundaries, such that diagnostic or cognitive tasks will not have an important effect. That is, although HRA methods commonly use the term "diagnosis" to refer to cognitive activity, analysts (and some methods) frequently emphasize only "initial diagnosis," that is, understanding the plant situation and deciding which procedure to enter, as opposed to also considering the fact that crews are making decisions while working through the procedures and during the execution of the "response plan." The results showed that failure to adequately consider the crews' cognitive activities and related potential failure mechanisms can lead to a failure to identify important influencing factors and result in HEP underestimations.

- *Addressing crew characteristics*: The study provided evidence that crew factors such as team dynamics, work processes, communication strategies, sense of urgency, and willingness to take knowledge-based actions can have significant (positive or negative) effects on crew performance. While such factors can certainly be worth investigating, it is often difficult in the context of the PRA to observe enough crews to identify systematic crew characteristics and evaluate their potential influence on the scenarios. Moreover, crew-to-crew variability is not explicitly considered by many methods. Most methods consider a “representative” crew (a crew with characteristics judged to be average or typical) in a “base case” quantification, while a few (A Technique for Human Event Analysis (ATHEANA) and Méthode d’Evaluation de la Réalisacion des Missions Operateur la Sûreté (MERMOS)) explicitly consider scenario variations and can address crew-to-crew variability in estimating the HEP. The question for HRA is to what degree these issues need to be taken into account, and how feasible it is to try and do so. Given the current state-of-the-art in HRA, it appears that crew variability effects can be evaluated using sensitivity analyses on the HRA results, allowing the evaluator to examine whether the effects are important enough to investigate and try to explicitly incorporate into the analysis.

- *Incorporation of failure mechanism and contextual factors*: The study produced substantial evidence that methods that focus on identifying failure mechanisms (ways the crews could fail a particular task) and the contextual factors that enable these mechanisms tended to produce richer content in the qualitative analysis than the PSF-focused methods. Moreover, the resulting operational stories frequently predicted actual crew performance, providing evidence that HRA does have the ability to predict what could or would occur when the crew responds to the scenario. However, methods with richer operational stories did not necessarily lead to HEPs that were more consistent with the empirical data, as other factors are apparently involved (e.g., availability of reliable quantification processes and associated guidance for translating the richer information into HEPs).

- *PSF Treatment*: The study produced evidence that selecting an appropriate PSF and judging its degree of influence on performance is an important factor, and contributed to both over and underestimation of HEPs. In both the LOFW and SGTR scenarios, the inconsistencies were identified in the selection and the weighting of the PSFs thought to be important. Inconsistencies are explained on the basis that the HRA teams did not develop to the same degree a qualitative understanding of the details of the scenario, and that there were differences in the interpretation of the scope of the PSFs and in the ratings assigned to the PSFs. In most of the HRA methods using PSFs, the guidance provided to support these judgments is limited. Another PSF-related issue concerns whether an adequate range of PSFs are addressed by a given method. The study provided evidence that PSF-based methods did not capture some of the relevant influencing factors identified in the data simply because they were not addressed by the method. This finding suggests that to be able to reliably predict performance, HRA methods need to cover an appropriate range of PSFs.

The traceability of the qualitative and quantitative analyses

Traceability is an important aspect of HRA. In this study, two different aspects of traceability were evident in the HRA analyses: (1) traceability of the quantification itself, given the choices made in the analysis, and (2) traceability of how the judgments from any qualitative analysis are reflected in the method's representation (e.g., basis for the chosen PSFs and their weights). The study suggests that for some PSF-based methods, the first aspect of traceability may be good. For example, in SPAR-H, the simplicity of the base probabilities and the adjustments of the multipliers make the quantification very traceable. However, with respect to the second aspect of traceability, the same methods may not be as good; assigning weights to PSFs relies heavily on the documentation provided by the analysts to make it traceable.

For the context-based methods (e.g., MERMOS or ATHEANA), this picture was almost the opposite. These methods have established good approaches for identifying and transferring qualitative analysis and judgments into an understanding of the conditions facing the operators, and they develop strong operational stories as a basis for quantification. However, since these methods rely on expert judgment, they lack an easily traceable way of translating these scenario stories into HEPs in the quantification process, and there is no guarantee of reproducibility, even when the analysts agree on the assumptions and aspects of the scenario descriptions.

Generation of insights for error reduction

This assessment addresses the degree to which the qualitative analysis and the performance influence evaluation addressed by the HRA method provide information that would allow insights into error reduction.

Most methods do not offer specific guidance for error reduction. Analysis by the more traditional PSF-based methods, if performed in conjunction with a good qualitative analysis leading to a sufficient examination of PSFs and situational factors, allows insights into improving safety and reducing errors: that is, examining aspects that, when identified as problematic, could be improved to facilitate error reduction. However, this capability depends heavily on the rigor of the judgments made about the different potential situational factors and the underlying qualitative analysis (including the range of the PSFs addressed by the method, for which additional guidance is needed in many methods).

The newer, narrative-based methods describe how elements of the scenario, task, human-machine interface, and operator aids may contribute to the HFE. The failure scenarios can be directly understood by plant experts, and the specificity and detail level of these narratives make them directly usable in error reduction. For example, the ATHEANA search strategy is useful in identifying ways errors occur, and it lends itself to use in error reduction.

Insights for improving guidance and methods

Many findings on HRA predictive performance can be attributed to weaknesses in HRA guidance found in the documentation of HRA methods. Improvements in guidance could help to improve the qualitative and quantitative predictive performance of the methods. Specific examples of guidance improvements include:

- *Guidance improvement in the selection and treatment of PSFs*: As discussed above, the study produced evidence that inappropriate selection of the PSFs and/or their

degree of influence on performance contributed to both over and underestimation of the HEPs. Also, methods that include a limited range of PSFs often did not capture relevant performance factors. However, if appropriate factors were identified, some analysts were able to "stretch" their method to account for such factors (i.e., they interpreted the method's factors and features more broadly than described in the method guidance).
Therefore, for methods that include a limited set of PSFs, the application could potentially be improved by including guidance to ensure that (1) a comprehensive set of factors is identified in the qualitative analysis, (2) the identified factors can be assigned to the method's PSFs, and (3) the strength of the PSF is appropriately selected.

Methods that use PSF ratings as inputs to quantification could benefit from PSF scales with anchored ratings. The PSF scales and anchors can support a consistent, comprehensive qualitative treatment of the PSFs (the performance issues and challenges that are sought), as well as increase the consistency of the inputs to quantification (e.g., factor ratings). Another issue is that the PSFs provided by a method frequently include PSF definitions that may be interpreted as overlapping. Methods could improve their applicability by addressing overlapping so that effects such as double-counting are avoided.

- The HRA analyses that focus on identifying the ways that the crews could fail the tasks modeled by the HFE and examine crew performance in its operational aspects tended to yield richer, more insightful qualitative analysis results. These analyses were typically (but not solely) associated with the newer methods that quantify HFEs in terms of narrative-based failure scenarios and the contextual factors that enable these failure scenarios. For all methods, qualitative analysis guidance that would support analysts in a thorough assessment of potential failure mechanisms in connection with a variety of possible operational contexts compatible with the PRA scenario would be expected to lead to a more comprehensive and insightful qualitative basis as an input to HFE quantification.

- Several methods that quantify an HFE by decomposing it into sub-tasks (or allowing decomposition) lacked guidance for performing the decomposition and for determining a decomposition level appropriate to assign to subtasks the basic failure probabilities provided by the method. This quantification guidance needs to be method-specific. Furthermore, the development of such guidance needs to consider that HFEs may be defined at different levels of detail in the PRA. For example, some HFEs are defined at a functional level (e.g., initiate B&F), while others are defined at the specific task level (e.g., open a valve). This applies to both methods that use a very generic decomposition, such as diagnosis/cognition and execution, and methods that use a more detailed decomposition.

- While many HFEs in a PRA are fairly simple (small set of cues, a simple manipulation task), PRAs may also include HFEs with multiple assessment/decision subtasks and/or multiple execution subtasks taking place while the plant conditions evolve. The guidance provided by many methods for analyzing, modeling, and quantifying HFEs that involve such dynamic crew-plant interactions is quite limited. Guidance for handling dynamic interactions, as well as examples illustrating how such an analysis should be performed, will enhance the methods' capabilities.

- For several methods, a related issue is the need for guidance to address the cognitive activities of the crew that support the response modeled by an HFE: in other words, guidance for considering not only the primary diagnosis/situation assessment, but also for cognitive activities associated, for instance, with selecting among alternative strategies, prioritizing among tasks, and assessing the plant response to operator actions.

- Reasonableness checks are often performed in external (peer) reviews of probabilistic safety assessments (PSAs)/HRAs, where each individual HEP cannot be reviewed in detail and emphasis is placed on the relative values of the HEPs. The results of a number of the analyses submitted in the Empirical Study suggest that a reasonableness check was not performed because HEPs did not reflect either the level of difficulty or the qualitative differences of HFEs. Currently, we rely on the expertise of the analysts; this result suggests a need to develop guidance on performing reasonableness checks, regardless of the method used to perform an HRA.

Towards hybrid HRA methods

The Empirical Study results, taken as a whole, support the concept of combining the effective elements and features of the different HRA methods by integrating these elements into a hybrid method. The concept of developing hybrid method(s) is supported by the following overall conclusions, which include:

- Over the set of HFEs and the various human performance issues relevant to these HFEs, no method consistently outperformed the other methods.

- The methods did not perform equally well in the SGTR and LOFW HFEs. Overall, some analysis teams performed better in the later LOFW phase, which could be an effect of the analysts learning to analyze the second set better in terms of the measures used within the Empirical Study. However, the predictive performance for the LOFW HFEs was poorer for other teams, suggesting that some methods may be better at treating some kinds of HFEs than others.

- There were fairly clear differences in the ability to model assessment/decision and implementation/execution. The PSF-based methods tended to handle the latter better, while narrative-based methods had advantages for assessment/decision issues.

- Most methods could be used to model simple HFEs, but the simpler methods, with fewer degrees of freedom for structuring the quantification model, were difficult to apply to the complex, multiple-subtask HFEs in the sense that they did not provide the elements needed to represent the subtasks and their interactions.

- No method showed consistently conservative or optimistic tendencies in the HEPs obtained in this study. Thus, at least based on the evidence from this study, it appears that none of the methods may be suitable for producing “scoping” values or performing a conservative bounding analysis (e.g., SPAR-H or ASEP).

The aim of a hybrid method would be to guide a richer qualitative analysis, providing a broad scope of performance factors and failure modes and failure mechanisms, while maintaining the repeatability of the methods with more structured quantification. In summary, the Empirical

Study has not only highlighted some of the overall requirements for HRA methods, especially those related to guidance and supporting consistent analysis practices, but its method assessments also identify features of the various HRA methods that a hybrid method could incorporate.

Conclusions

The International HRA Empirical Study is the first major study to directly compare HRA predictions with actual operating crew performance in PRA-related accident scenarios conducted in a full-scope NPP simulator (HAMMLAB). The study has produced a large amount of information on human performance in both routine and challenging situations by studying fourteen crews addressing fifteen HFEs. Thirteen analysis teams performing fourteen HRA analyses have also produced large evidence of the HRA methods' ability to predict operator performance following accident-initiating events. Consequently, the findings of this study should be taken into consideration for improving human performance in NPPs, as well as in improving the tools used to evaluate performance.

With respect to improving human performance at the plants, the study provides a strong indication that challenging situations, such as those modeled in a PRA, should be regularly examined to improve plant design features, as well as operational features involving procedures, training, communications, team interactions, and leadership. The study's extensive documentation of crew performance in the simulator is also a rich source of information for practitioners dealing with human performance in general.

With respect to HRA, the study provides evidence that HRA has the ability to predict crew performance in responding to initiating events, but important improvements are needed. The study's findings and insights cast light on the adequacy of the set of performance-shaping factors addressed by a method, on the appropriateness of the scope of the PSFs, on the quantitative relationship between the identified PSFs and HEP estimates, and on how analysts interpret and apply a given method, such as the decomposition into sub-tasks. The method's/analyst's ability to identify both cognitive and execution demands under various conditions, to incorporate them into the methods' underlying human behavior model, and to accurately use the methods' quantification process to estimate HEPs are issues that need to be addressed at different levels by all methods.

Furthermore, adherence to HRA good practices, such as performing a thorough qualitative analysis, checking the reasonableness of produced HEPs, and ensuring whether a method used as a "screening" tool will produce conservative bounding values, is critical in applications whose objective is to reduce human error and improve plant capability to deal with initiating events.

As noted, this study is the result of collaboration between a large number of organizations that are taking advantage of its results to improve methods, tools, and practices in their applications. The NRC is taking full advantage of the results of this study: for example, efforts are underway to improve the SPAR-H method currently used in several regulatory applications. The NRC also conducted a follow-on study referred to as the U.S. HRA Empirical Study [23] [24]. This is a smaller-scale study addressing such issues as the use of plant visits to collect HRA information and inter-analyst reliability, which were not within the scope of the international study. The U.S. Empirical HRA Study is taking full advantage of the methodological tools developed here: the experimental design focusing on evaluating HRA methods; the methodology for collecting crew data; and the methodology for data-to-method comparisons, which were major achievements of this study. The study is also supported by several of the organizations that participated in the design and execution of the international study, including the Halden Reactor Project, the Paul Scherrer Institute, and EPRI, thus capitalizing on the expertise developed from this international activity.

Finally, the NRC and EPRI are pursuing the development of a hybrid method called the Integrated Decision-tree Human Event Analysis System (IDHEAS) [25]. The objective of IDHEAS is to address, to the extent possible, the issue of variability in HRA results. These NRC activities exemplify the ways in which the results of this study work to improve the robustness of HRA used to support regulatory decision making.

ACKNOWLEDGMENTS

This study is a collaborative effort between the Joint Programme of the Organization for Economic Co-operation and Development (OECD) Halden Reactor Project, particularly Halden's signatory organizations, which provided the analysis teams, the U.S. Nuclear Regulatory Commission (NRC), the Swiss Federal Nuclear Inspectorate (DIS-Vertrag Nr. 82610), and the U.S. Electric Power Research Institute (EPRI).

The authors gratefully acknowledge Bertil Johansson for invaluable assistance in analyzing the empirical data from the Halden Human-Machine Laboratory (HAMMLAB).

The work of the human reliability analysis (HRA) teams, 13 teams providing 14 analyses, has been invaluable. The authors of this report want to give a warm thanks to all of the teams who have contributed. We have tried to evaluate their analyses as objectively as possible. In addition, the views expressed by numerous participants in the preparatory workshops provided essential inputs to the design of the study. The following HRA teams contributed analyses of the loss of feedwater (LOFW) scenarios to this part of the study:

ASEP (UNAM)

- Pamela Nelson, Universidad Nacional Autónoma de México, Mexico
- Teresa Ruiz-Sánchez, Universidad Nacional Autónoma de México, Mexico
- Manuel González-Cuesta, Universidad Nacional Autónoma de México, Mexico

ASEP/THERP (NRC)

- Y. James Chang, U.S. NRC, Office of Nuclear Regulatory Research, USA
- Mary Drouin, U.S. NRC, Office of Nuclear Regulatory Research, USA
- Stacey Hendrickson, Sandia National Laboratories, USA
- Rick Grantom, C.R.Grantom P.E. & Assoc. Inc., USA

ATHEANA (NRC)

- Susan E. Cooper, U.S. NRC, Office of Nuclear Regulatory Research, USA
- Dennis C. Bley, Buttonwood Consulting, USA
- Mark King, U.S. NRC, Office of Nuclear Reactor Regulation, USA
- Michael A. Junge, U.S. NRC, Office of New Reactors, USA
- John E. Thorp, U.S. NRC, Office of Nuclear Reactor Regulation, USA

CBDT + THERP (EPRI)

- Jan Grobbelaar, Scientech, USA
- Kaydee Kohlhepp, Scientech, USA

CESA_Q (PSI)

- Luca Podofillini, Paul Scherrer Institut, Switzerland
- Bernhard Reer, Swiss Federal Nuclear Safety Inspectorate – ENSI, Switzerland

CREAM (NRI)

- Jaroslav Holy, Nuclear Research Institute, Czech Republic
- Jan Kubicek, Nuclear Research Institute, Czech Republic

Decision Trees + ASEP (NRI)

- Jaroslav Holy, Nuclear Research Institute, Czech Republic
- Jan Kubicek, Nuclear Research Institute, Czech Republic

Enhanced Bayesian THERP (VTT)

- Jan-Erik Holmberg, Technical Research Centre of Finland, Finland
- Kent Bladh, Vattenfall Power Consultant, Sweden
- Johanna Oxstrand, Ringhals AB, Sweden
- Pekka Pyy, Teollisuuden Voima Oy, Finland

HEART (Ringhals)

- Johanna Oxstrand, Ringhals AB, Sweden
- Kent Bladh, Vattenfall Power Consultant, Sweden
- Steve Collier, OECD Halden Reactor Project, Norway

K-HRA (KAERI)

- Wondea Jung, KAERI, Korea
- Jinkyun Park, KAERI, Korea

MERMOS (EDF)

- Pierre Le-Bot, Electricité de France, France
- Hélène Pesme, Electricité de France, France
- Bastien Brocard, Electricité de France, France
- Patrick Meyer, Electricité de France, France

PANAME (IRSN)

- Véronique Fauchille, Institut de Radioprotection et de Sûreté Nucléaire, France
- Vincent Ridard, Institut de Radioprotection et de Sûreté Nucléaire, France
- Manuel Lambert, Institut de Radioprotection et de Sûreté Nucléaire, France

SPAR-H (INL)

- April Whaley, Idaho National Laboratory, USA
- Harold Blackman, Idaho National Laboratory, USA

SPAR-H (NRC)

- Gary M. DeMoss, U.S. NRC, Office of Nuclear Regulatory Research, USA
- Bruce B. Mrowca, Nuclear Systems Analysis Division, ISL, Inc., USA
- Chris Hunter, U.S. NRC, Office of Nuclear Regulatory Research, USA

Three teams have utilized the scenario descriptions or data as input for their simulation models in order to test the applicability of these methods. These teams have participated by giving general input to the study, but the results have not been analyzed or compared to the HAMMLAB data.

ECAT, Discrete event simulation, MicroSaint (NRC/Sandia/Alion)

- Beth M Plott, Alion Science, USA

IDAC (University of Maryland)

- Kevin Coyne, University of Maryland, USA
- Ali Mosleh, University of Maryland, USA

QUEST-HP (Risø, DTU)

- Igor Kozine, Technical University of Denmark, DTU (Earlier Risø National Lab)

Another team has used the data to test a selection algorithm for fuzzy classification of results.

Fuzzy data classification (Politecnico di Milano)

- Piero Baraldi, Politecnico di Milano, Italy
- Massimo Librizzi, Politecnico di Milano, Italy
- Enrico Zio, Politecnico di Milano, Italy

The views expressed in this report do not necessarily represent the views of the U.S. NRC, or of the organizations affiliated with the Halden Reactor Project. This report was prepared as an account of work sponsored in part by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights.

Sandia National Laboratories is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Idaho National Laboratory is a multiprogram laboratory operated by Battelle Energy Alliance LLC for the United States Department of Energy under contract DE-AC07-05ID14517.

ACRONYMS AND ABBREVIATIONS

AFW	auxiliary feedwater
AFWS	auxiliary feedwater system
ARO	assisting reactor operator
ASEP	Accident Sequence Evaluation Program
ATHEANA	A Technique for Human Event Analysis
B&F	bleed and feed
BIT	boron injection tank
CBDT	Cause-Based Decision Tree
CD	core damage
CESA-Q	Commission Errors Search and Assessment – Quantification
CPC	common performance condition
CREAM	Cognitive Reliability Error Analysis Method
EDF	Electricité de France
EFC	error-forcing context
EOC	error of commission
EOO	error of omission
EOP	emergency operating procedure
EPC	error-producing condition
EPRI	Electric Power Research Institute
ERG	emergency response guideline
FW	feedwater
GTT	generic task type
HAMMLAB	Halden Human-Machine Laboratory
HCR/ORE	human cognitive reliability/operator reliability experiments
HEART	Human Error Assessment and Reduction Technique
HEP	human error probability
HFE	human failure event
HMI	human-machine interface
HRA	human reliability analysis
HRP	Halden Reactor Project
IE	initiating event
INL	Idaho National Laboratory
IRSN	Institut de Radioprotection et de Sûreté Nucléaire
K-HRA	Korean Human Reliability Analysis
KAERI	Korea Atomic Energy Research Institute
LOFW	loss of feedwater
MERMOS	Méthode d'Evaluation de la Réalisacion des Missions Operateur la Sûreté
MSIV	main steamline isolation valve
NPP	nuclear power plant
NRC	U.S. Nuclear Regulatory Commission
NRI	Nuclear Research Institute
OECD	Organization for Economic Co-operation and Development
PANAME	New Action Plan for the Improvement of the Human Reliability Analysis Model
PORV	power-operated relief valve
PRA	probabilistic risk analysis
PRT	pressurized relief tank

PRZ	pressurizer
PSA	probablistic safety analysis
PSF	performance-shaping factor
PSI	Paul Scherrer Institute
PWR	pressurized water reactor
RCP	reactor coolant pump
RCS	reactor coolant system
RO	reactor operator
RWST	refueling water storage tank
SG	steam generator
SGTR	steam generator tube rupture
SI	safety injection
SPAR-H	Standardized Plant Analysis Risk-Human Reliability Analysis method
SS	shift supervisor
THERP	Technique for Human Error Rate Prediction
TRC	time/reliability correlation
UA	unsafe act
UNAM	Universidad Nacional Autónoma de México
VTT	Technical Research Centre of Finland
WR	wide range

1. INTRODUCTION

1.1 Intended readers

This report should be useful for human reliability analysis (HRA) method developers and HRA practitioners, as well as those who want a better understanding of HRA in general. The study has provided insights that can be used by nuclear power plant staff tasked with improving procedures, training, and general safety. This report partially describes these results, as well as referencing earlier, more detailed reports. For HRA method developers, the sections on specific conclusions for each method should be particularly important. There is a lot of advice regarding the strengths and weaknesses of the method, and recommendations as to which parts of the method's documentation need improvement. For HRA analysts using various methods, both the overall conclusions and the method-specific conclusions should be relevant. Some of the overall conclusions may be seen as advice on best practices within HRA. Lessons learned in benchmarking studies are included, as are suggestions for future HRA studies.

1.2 Background and motivation (aims)

Diverse HRA methods are currently available to treat human failure in probabilistic risk assessments (PRAs). Given the differences in the scopes of the methods and their underlying models, there is substantial interest in assessing HRA methods, and, ultimately, in validating the approaches and models underlying them. Such a validation is warranted to assess the credibility of HRA results in risk-informed decisions.

In the International HRA Empirical Study, a diverse set of HRA methods was assessed based on reference data obtained in a dedicated simulator study. The benchmarking and assessment of each method involved comparing the empirical data with the predictions made under the method. The comparisons examined both the qualitative and quantitative methods' predictions. Qualitative predictions included the aspects of the scenario or task that were identified as the driving factors of human performance or as leading to difficulties, while the quantitative comparisons took into account the estimated failure probabilities of the defined human failure events (HFEs) and their ranking by failure probability within each set. Both types of predictions were compared with their empirical counterparts, which were produced by the data analysis of the emergencies simulated at the Halden Human-Machine Laboratory (HAMMLAB).

Overall, the HRA methods were primarily assessed based on their qualitative predictive power. This was an anticipated consequence of the study design. Although the simulations employed up to 14 crews and a total of 48 scenario runs, the sample of HFEs with a low failure probability is small for the purpose of estimating a reference probability. Even more importantly, the limitations on the statistical analysis and quantitative reference data are in many ways inherent to HRA: human performance is known to be situation-specific, and HRA data and analysis must consider not only average performance (aggregating data from different contexts), but also the impacts of the situational context factors on performance in specific scenarios. Consequently, the study provides a stronger test of the qualitative insights than of the quantitative results: that is, it is more of a test of the methods' ability to identify the performance issues in the scenarios and their capacity to use this information to produce human error probabilities (HEPs) that reflect the difficulty of the associated tasks, rather than a test of the methods' accuracy in matching "empirical HEPs." However, the quantitative results still played an important role in the comparison process, and in the overall evaluation of the methods.

1.3 Scope

This study has addressed the use of HRA for HFEs similar to those that would be found in a probabilistic risk/safety analysis (PRA/PSA) with a scope corresponding to a Level 1 (core damage) for internal initiating events during full-power operation. Given an initiating event (IE), we studied how the IEs were handled in the HAMMLAB simulator by licensed operating crews. The scenarios typically lasted for one to two hours, and stopped when data on the tasks of interest (HFEs) were collected. Similarly, the HRA teams were told to analyze these events and calculate the HEP for failure of a set of predefined HFEs. Two scenario types with two variants each, with or without complicating factors were addressed. The first scenario type was a steam generator tube rupture (SGTR), while the second was a loss of feedwater (LOFW).

We studied control room actions associated with the crew response to these internal initiating events, rather than field operator actions. All interactions with field operators were done by roleplay in HAMMLAB, and these actions had a predefined level of difficulty, which was planned for the various scenarios (e.g., restoration of failed equipment was determined beforehand as part of the scenario design). Most of the HFEs represented individual operator actions, with the exception of two feed and bleed cases, where the joint probability, representing failure of both an early action and a late action, was modeled.

The results documented in this report are thus mainly valid for control room actions within level 1 PRA. As the scenarios reflect typical use of emergency operating procedures (EOPs) after an initiating event, we believe that the results may be generalized to other IEs, such as LOCA. Some of the crews' tasks in the complex scenarios of this study were very difficult (for instance, due to time requirements and/or instrumentation failures). By masking key indications of plant states, we designed the scenarios to challenge the operating crews' progress in the emergency operating procedures. In one of the variants of the SGTR scenario, for instance, all radiation indications were masked. In another scenario, two out of three steam generators had level measurements misscalibrated or stuck. This enabled us to study how crews might solve very difficult situations, and to evaluate the HRA methods' predictions of these scenarios. In spite of the challenges, the operating crews maintained control of the reactor, and were able to redirect the procedures when their goals were not achieved. However, we observed a rather large range of performance effectiveness among the crews, as was expected, given the challenging nature of the scenarios.

Control room diagnoses, decisions, communications, and actions in emergency scenarios may be different from their equivalent in maintenance work. This study did not address the latter class (e.g., manual task work in the field during maintenance), and the results should not be directly used for pre-initiators (PRA study of actions leading to or becoming latent errors for initiating events).

Using simulators to study difficult scenarios in nuclear power plants may not provide an appropriate degree of realism; for instance, one may question whether factors like stress are present in simulated emergencies, or whether roleplaying actions outside the control room oversimplifies the external world. We should emphasize that the threats to a simulation's validity, as well as their countermeasures, are well understood (NEA/CSNI/R(98)1, 1998, p. 164). It is important to point out that simulations' ecological validity issues are among the factors that make it difficult to incorporate failure data obtained in simulator studies directly into PRA. This study further supports the idea that HRA is needed to account for the specificities of situations, plants, and crews.

This study delves into the work of operating crews, throwing light on central aspects of HRA practice (e.g., task analysis, treatment of different scenarios' evolutions, treatment of complicating factors, treatment of the use of operating procedures, etc). Information from emergency simulations is as important to risk analysis as the use of simulators is to operator training and qualification.

1.4 Overview of the study design, tasks

The International HRA Empirical Study focuses on the HRA of the control room personnel actions required in response to PRA initiating events. This focus was motivated by the widespread use of HRA methods within industry PRA/PSA, as well as by the significant research and development efforts on HRA methods addressing the issue of errors of commission and decision making performance, as surveyed, for instance, in [4]. An overview of the study, which consists of the four high-level tasks listed below, is presented in Figure 1-1.

- **Task 1**. Define the scenarios and the HFEs to be analyzed and compile the inputs for the HRA analyses.
- **Task 2.** Produce the empirical or reference data for the comparison, starting from the collection of raw data in simulator experiments conducted in HAMMLAB, and analyze it.
- **Task 3.** Analyze the HFEs with HRA methods, which produces the predicted outcomes.
- **Task 4.** Compare the predicted outcomes against the empirical data and develop insights for improving HRA methods and practices.

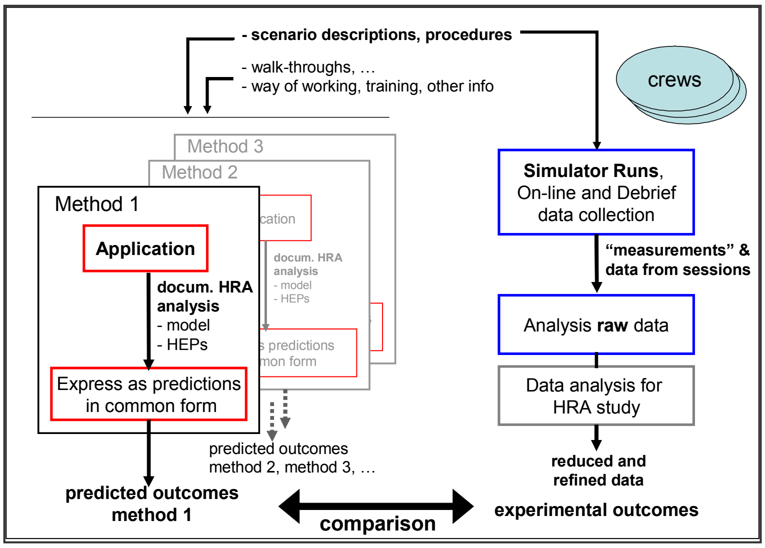
- scenario descriptions, procedures
- walk-throughs, …
- way of working, training, other info
crews
Method 3
Method 2
Method 1
Application
cation
docum. HRA analysis
- model
- HEPs
docum. HRA analysis
- model
- HEPs
Express as predictions in common form
s predictions
mon form
predicted outcomes method 2, method 3, …
predicted outcomes method 1
Simulator Runs, On-line and Debrief data collection
"measurements" & data from sessions
Analysis raw data
Data analysis for HRA study
reduced and refined data
experimental outcomes
comparison

Figure 1-1 Overview of the HRA Empirical Study

Task 1 is the compilation of the inputs for the HRA analyses. As shown at the top of Figure 1, these inputs include not only the descriptions of the scenarios and of the HFEs to be analyzed, but also information on the relevant procedures, the training of the operators, their way of working, the human-machine interface, and other aspects of the performance context. The performance of the predictive HRA analyses (Task 3) is shown on the left. The production of the empirical data, Task 2 (right-hand side of Figure 1-1), consisted of three subtasks: (1) the simulator experiment itself, in which the operator crews responded to the scenarios while observations and other data were collected; (2) an initial data analysis stage aimed at producing an understanding of individual crew performances; (3) an HRA-oriented data analysis, which aggregated the set of crew performances in order to characterize the overall performance level related to each HFE and the drivers of performance. Task 4 was the comparison of the predicted outcomes to the empirical outcomes, and required the predicted and observed outcomes to be expressed in a compatible format.

1.5 Study organization, participants, and roles

There were four sets of study participants:

- **Halden experimental staff** (Tasks 1, 2): The simulator sessions were conducted in the Organization for Economic Co-operation and Development (OECD) Halden Reactor Project's HAMMLAB research simulator facility. The Halden staff was responsible for collecting and analyzing the experimental data.

- **Operator crews** (Task 2): A set of licensed operator crews responded to a series of scenarios in the HAMMLAB simulator. Each crew responded to four scenarios consisting of a base and a "complex" variant of two scenario types.

- **HRA teams** (Task 3): Each team applied an HRA method to obtain predictions for the HFEs in the scenarios defined for the study. Organizations representing industry, regulators, and the research community have participated.

- **Assessment group** (Overall organization and Tasks 1, 4): This group had the overall responsibility for the organization and implementation of the study. In the early stages of the study, it prepared the information package (analysis inputs) for the HRA teams and answered their subsequent requests for additional information and questions concerning ambiguities in the instructions and assumptions. After the HRA teams delivered their analyses, the group reviewed and summarized the predicted outcomes before performing the actual comparison.

1.6 Phases of the Empirical Study

The Empirical Study has been structured in three phases, as shown in Table 1-1. The focus of Phase 1 was to test the study methodology. The HRA teams analyzed nine HFEs in a first set of scenarios, two variants of SGTR scenarios. In Phase 1, the data analysis and a qualitative comparison were performed for the first two of these HFEs, and the results were reported in HWR-844/NUREG/IA-0216 Volume 1 [1]. The remaining HFEs of the SGTR scenarios and the quantitative comparison are addressed in HWR-915/NUREG/IA-0216 Volume 2 [2], which covers Phase 2. HWR-951/NUREG/IA-0216 Volume 3 [3] documents Phase 3, which covers the second set of scenarios, the two variants of LOFW scenarios. The three phases were designed to allow the study participants (Halden, the assessment/evaluation group, and the HRA teams) to review the study methodology and the initial results, and, in particular, to allow the HRA teams to provide feedback on the methodology. Workshops on all phases were conducted, in which all HRA teams participated and discussed empirical results and preliminary comparison results with both the assessment and the experimental groups. This report documents the final conclusions from the study and presents the overall results.

Table 1-1 Phases of the Empirical Study.

Phase 1 (2007-2008) Pilot study	HRA teams analyzed SGTR scenarios based on information package Used data from two HFEs in SGTR scenarios Established the methodology and reached some preliminary results on HRA methods Workshop on phase 1 results, October 2007 Issued HWR-844/NUREG/IA-0216 Vol. 1, results of phase 1 (two HFEs from SGTR)
Phase 2 (2008-2010)	Data analysis and comparison of remaining seven HFEs in SGTR scenarios; refined methodology, including quantitative issues Workshop on phase 2 (SGTR results), March 2009 Issued HWR-915/NUREG/IA-0216 Vol. 2, study results of phase 2, all HFEs from SGTR
Phase 3 (2009-2011)	HRA teams analyzed LOFW scenarios based on information package and knowledge of crews based on phase 1 discussions and report Data analysis and comparison of LOFW scenarios Workshop on phase 3 (LOFW results), December 2009 Issued HWR-951/NUREG/IA-0216 Vol. 3, study results of phase 3, LOFW

Phases 2 and 3 partly overlap in time because the HRA teams performed predictive analyses for the LOFW scenarios while the SGTR data and predictions were being analyzed.

A particular facet of this work was the workshop of each phase in which all the HRA teams participated. In these workshops the results of observed crew performance as well as the comparisons of analytical to empirical results were discussed. The workshops contributed to the success of the study for many reasons. The in-depth discussions about crew performance and underlying performance drivers were important to the HRA teams; they were given the opportunity to understand how crews could perform and how and to what extent their analysis/methods could handle the observed performance. The in-depth discussions about the comparisons of each submitted analysis to empirical data and the opportunity afforded to the HRA teams to express their views on their analyses strengthened the outcomes of this study addressing important issues in HRA practices. An independent review was performed of the first phase of the study and its methodology. This was presented and discussed at the second workshop.

This report presents the overall conclusions from the whole study, both at the general HRA level and for each HRA method.

2. METHODOLOGY

This chapter describes the methodology that was developed and applied in the study. The first section describes the data collection performed at the Halden Human-Machine Laboratory (HAMMLAB) facility and the analysis of these data to derive the empirical (reference) data. The second section describes the methodology for the comparison between the empirical HAMMLAB data and the human reliability analysis (HRA) predictions from the HRA teams, and the assessment method that was used for the HRA methods. The third section contains a description of the input that the HRA teams received before their analyses, as well as their required reporting. Scenarios and human failure event (HFE) definitions are described at a general level in the fourth section, and in more detail, including event trees, in Appendix A.

This section is organized as follows:

- **Section 2.1**. Task 1 – The simulation design, including the definition of the scenarios and HFEs to be analyzed, and compilation of the inputs for the HRA analyses.

- **Section 2.2.** Task 2 – The production of the empirical or reference data for the comparison, starting from the collection of raw data in simulator experiments conducted in HAMMLAB and followed by the analysis of this data.

- **Section 2.3.** Task 3 – The analysis of the HFEs with HRA methods, which produced the predicted outcomes.

- **Section 2.4.** Task 4 – The comparison of the predicted outcomes against the empirical data and development of insights for improving HRA methods and practices.

2.1 Simulation design

2.1.1 Data Collection and Participants

Fourteen crews of licensed pressurized water reactor (PWR) operators participated in the study. The HAMMLAB PWR simulator, called FRESH, is a full-scope simulator of a three-loop Westinghouse French plant (CP0 series). HAMMLAB uses a computerized human-machine interface for the PWR simulator, and the procedures are based on those used at the participating operators' home plant, adapted to the simulated PWR and the HAMMLAB interface. The home plant uses the Emergency Response Guidelines (ERGs, Revision 2) developed by the Westinghouse Owners Group.

The crews' home plant has conventional control rooms with panels and alarm tiles, while the HAMMLAB PWR simulator is based on digital instrumentation and control. There are also a few differences between the systems/equipment in the crews' home plant and those simulated in the Halden PWR simulator (e.g., the power-operated relief valves (PORVs) are different). Therefore, prior to the experimental scenarios, the crews were trained in the use of the screen-based interface, and on the differences between their home plant and the simulator. For more details on the training, see [1], [2], and [3].

The crew staffing for the study was reduced from the normal staffing at the plant. Each crew consisted only of a shift supervisor, a reactor operator, and an assisting reactor operator. The staffing expected in emergencies like those simulated for this study would also include a balance of plant operator, two or more field operators, and, for most crews, an extra operator or shift supervisor. In the current experiment, the assisting operator did the initial checks for turbine trip and then acted as an assisting reactor operator. The interactions with the field operator(s), the safety engineer, and plant management were simulated via roleplay. An operations expert situated at the gallery in HAMMLAB played all of these roles, as needed, by answering phone calls from the control room. The crew was instructed to interact with the organizational environment as they would in the plant or in a training simulator session. For more details on the crews' operation performance, both in their home plant and in HAMMLAB, please see HWR-844, HWR-915, and HWR-951 [1], [2], and [3].

All fourteen crews from the home plant participated in the study. Due to simulator problems, only 10 of the 14 loss of feedwater (LOFW) scenarios were analyzed. During the seven-week data collection period of the study, two crews participated in the experiment per week. Each crew stayed in Halden for three days, starting either on Monday or on Wednesday.

To control for confounding effects caused by learning due to the order of presentation of treatment level (base case or complex case, that is, degree of complexity for the "treatment" or manipulation of the independent variable in the experiment) and scenario type, the experiments were organized under a combination of theoretical and combinatorial considerations. This included, for example, avoiding combining consecutive presentations of the same scenario type on the same day, and avoiding a contiguous scenario type between day one and day two. It was also assumed that there was symmetrical learning between scenario types (i.e., learning related only to increasing simulator experience).

2.1.2 Defining Human Failure Events

To ensure that all HRA teams would produce predictions for identically defined HFEs, this study predefined all HFEs for the HRA analysts. These definitions were based on the definitions of similar HFEs from real plant PRAs and were defined on a functional level (i.e., "fails to perform X before Y" or "fails to perform X within *t* minutes"). In some cases the HFEs were defined with stricter success criteria than a standard PRA would use. The reason for this is that while the HFE success criteria used in the study should relate to those commonly used in PRA/HRA, they also have to be clearly observable in the simulated scenarios.

Failure for some actions might result in a near-immediate plant state change, such as onset of core damage. This type of action (similar to the "bleed and feed" (B&F) HFEs in the LOFW scenarios) could be defined in this study as it would be in a PRA. However, some other actions may not have such a near-immediate and/or irreversible plant state change, but might instead have a lesser effect, such as delaying the overall timing of the operators' response or influencing the evolution of the scenario (e.g., system pressure ends up at a level that is not preferred), without necessarily affecting the final outcome of interest (e.g., core damage). Therefore, for actions without immediate and/or irreversible plant state changes, the corresponding HFEs were defined by superimposing timing criteria.[1]

In some cases, PRA event trees model accident sequences by defining several HFEs that are carried out in sequence, as was the case in our referenced PRA for the steam generator tube rupture (SGTR) event. Failure of any specific HFE within the sequence, or even all HFEs observable in a typical simulation session of one to two hours, might not cause core damage. In the SGTR event, for instance, the HFEs include identification and isolation of the affected

[1] Timing criteria might be also introduced for actions that relate directly to irreversible plant states, but that have time windows so large that typically no crew failures are observed in simulated events due to high performance reliability and high likelihood of recovery.

steam generator, forced cooldown of the reactor coolant system, depressurization of the system, and termination of safety injection when appropriate. These actions collectively stop the leakage past the tubes, significantly decrease the radiation being released to the secondary side of the plant, and allow for control of the plant to eventually achieve safe shutdown. Failure in any or even all of these actions does not necessarily cause core damage.

On the other hand, since these actions should be conducted expeditiously in order to stop the primary-to-secondary leakage, decrease the radiation release, and avoid core damage without resorting to less desirable means (e.g., bleed and feed), it is reasonable to associate failure of any single action with a time criterion. These definitions impose a time by which the actions should be taken, as well as the precise steps that have to be taken to implement the desired action (e.g., close main steam line isolation valves as part of the overall isolation action). This allows analysts to observe whether the actions are performed within the desired time frames, and whether all critical steps are indeed carried out.

Another reason to adopt stricter failure criteria is that, when assessing the effect of performance-shaping factors (PSFs), human performance should ideally be treated at a fine level of granularity. As there can be several human actions that are carried out in a particular sequence to achieve one specific goal, the contexts or PSFs for these actions vary. The procedure might for example give guidance of varying clarity, the complexity of the sub-tasks can also vary, and, if the series of tasks is extended in time, stress might increase with it.

The timing criteria used in the definitions of the HFEs were obtained from expectations of operators' performance and/or inferred from plant thermal-hydraulic expectations for the scenarios. For a few of the HFE definitions, we included time limitations based on trainers' expectations, rather than more common PRA criteria (e.g., "25 minutes from rupture" rather than "overfill and opening of the steam generator (SG) relief valve" as failure criteria for the HFE "SG identification and isolation"). It is recognized that failure criteria based on training expectations and past experiences with the participating crews could lead to somewhat artificial definitions of failure, as training expectations and past experience may be loosely defined.

Despite this, another argument to include timing criteria involves the expectations of the HRA methods to be applied: for instance, many methods use a time reliability function to estimate the diagnosis failure probability portion of the HEP. Use of this function requires a definitive time by which the desired action must be performed (typically referred to as the "allowable time"). Then, based on the time estimated for implementing the action, the analyst can block out a time by which the diagnosis must occur. Using the time reliability function to determine the diagnosis time yields the diagnosis failure probability. To use such methods, the HFEs had to be defined in terms of time, although we recognize that the relation to the plant thermal-hydraulic expectations for the scenarios could have been made more explicit in the example above.

2.1.3 Scenarios and HFE definitions

The HFEs analyzed in this study occur in two versions each of an SGTR and an LOFW scenario, a base case and a complex case.

All 14 crews ran all four scenarios of the experiment:

- SGTR base
- SGTR complex
- LOFW base
- LOFW complex

All HFEs denoted as “B” are HFEs in the complex scenarios, while “A” represents HFEs in the base scenarios.

The HFEs in the SGTR scenario were summarized as follows:

- HFE-1A: Failure of the crew to identify and isolate the ruptured SG in the base SGTR
- HFE-1B: Failure of the crew to identify and isolate the ruptured SG in the complex SGTR
- HFE-2A: Failure of the crew to cool down the reactor coolant system (RCS) expeditiously in the base SGTR
- HFE-2B: Failure of the crew to cool down the RCS expeditiously in the complex SGTR
- HFE-3A: Failure of the crew to depressurize the RCS expeditiously in the base SGTR
- HFE-3B: Failure of the crew to depressurize the RCS expeditiously in the complex SGTR
- HFE-4A: Failure of the crew to stop the safety injection (SI) in the base SGTR
    - HFE-4A applies to the base scenario only.
- HFE-5B1: Failure of the crew to give a closing order to the PORV block valve associated with the partially open PORV within five minutes of closing the PORV used to depressurize the RCS (but it remains partially open, allowing ~6% flow). The PORV position indication shows “closed.”
- HFE-5B2: Failure of the crew to give a closing order to the PORV block valve associated with the partially open PORV within five minutes of closing the PORV used to depressurize the RCS (but it remains partially open, allowing ~6% flow). The PORV position indication shows “open.”

HFEs 5B1 and 5B2 apply only to the SGTR complex scenario, and are two different versions of HFE #5B. Half of the crews were in a group analyzed per HFE-5B1, and the other half were in a group analyzed per HFE-5B2.

The HFEs in the LOFW scenario were summarized as follows:

- X4 = Initiation of Primary Bleed and Feed = Establish/Initiate B&F before SG dryout. SG dryout occurs when there is no water left in the SGs, indicated by 0% wide range (WR) SG level.

- X4L = Late Recovery Before Core Damage = Establish/Initiate B&F within 25 minutes of SG dryout. This HFE is conditional on X4 (failure of B&F before dryout).

The HFEs to be estimated by the HRA teams are coded as follows:

HFE-1A:	X4 in the base case
HFE-2A:	X4L in the base case (given failure of HFE-1A)
HFE-1A1:	X4*X4L in the base case (given failure of HFEs 1A and 2A)
HFE-1B:	X4 in the complex case
HFE-2B:	X4L in the complex case (given failure of HFE-1B)
HFE-1B1:	X4*X4L in the complex case (given failure of HFEs 1B and 2B)

See Appendix A for detailed scenario descriptions and HFE definitions.

2.2 Empirical Data Analysis Methodology

The methodology used to obtain the reference data for comparison consists of the following phases:

1. Collection of raw data for the scenarios listed in Section 2.1.3

2. Crew-level analysis

3. Determination of the number of HFE failures

4. Aggregate level analysis: writing the operational descriptions (summaries of how the crews performed under the various HFEs) and derivation and rating of the PSFs

5. Assessment of the relative difficulty of the HFEs and their ranking

2.2.1 Raw data

The data collection for the experiment included:

- Logs of all crews' activities in the simulator, log of plant process parameters, and all events in the simulated process.

- Audio/videos: Two fixed cameras behind the operators and two head-mounted cameras, one each for the shift supervisor and the reactor operator, were employed. All operators were equipped with wireless microphones.

- Crew interviews: After each scenario, the crew participated in an interview focusing sequentially on phases of the scenario.

- Performance ratings: After each scenario run, a process expert rated the crews' performance of the main tasks on a scale from 1 to 5, where 1 is poor performance, 3 is average performance, and 5 is excellent performance. The process expert was a former shift supervisor who has also worked as a nuclear power plant simulator instructor, and this rating should correspond to a performance assessment from a trainer's point of view.

The performance measures utilized extensive information on the various phases of the scenario, which correspond to the defined HFEs. However, for the present study, audio/video recordings, coupled with simulator log data, constituted the fundamental information sources for writing narratives about crew performance in the HFEs, for deriving the PSFs, and, in general, for allowing detailed understanding of what the crews did, when they did it, and why. This process is described below.

2.2.2 Crew-level analysis

The method-to-data comparison's strong focus on qualitative aspects of HRA predictions required the analysts to investigate crews' performance at a detailed operational level. This included such aspects as the identification of specific execution conditions resulting from dynamic crew-system interactions and an understanding of the decision processes involved in observed procedural activities.

The cornerstone of the in-depth qualitative analysis was the review of the audio-video recordings, coupled with data logs of process parameters and simulator and operator activities. The analysis was interdisciplinary, requiring a great deal of nuclear power plant (NPP) process expertise, in addition to human factors knowledge. Two analysts with backgrounds in control room operation viewed the videos and transcribed key communications and events. They also wrote explanatory comments about salient aspects of crew performance. The accuracy and validity of the reviews were enhanced by online graphical access to log data, which allowed the analysts to reconstruct plant conditions at any given time.

2.2.3 Crew performance and HFE failures

By combining the information contained in the crews'/scenarios' reviews with quantitative performance data (e.g., performance times, SG level), we determined the crew performances corresponding to HFE successes and failures.

The HFEs are defined at a functional level, such as "failure to perform X before Y" or "failure to perform X within t minutes after Y," where X represents a set of crew actions (e.g., starting pumps, opening valves) and Y represents a state of the process. As both actions and process states are available from the simulator log data, the determination of the number of failures could be considered straightforward. It should be remembered, however, that the time t (25 minutes) after SG dryout for late recovery before core damage is an average estimate of the start of core damage. The actual time would depend on previous events and other scenario conditions that were not equalized for all crews (e.g., SG pressure). Since the simulator does not accurately model core damage states, it is not possible to infer whether the success/failure of the study HFE for "late recovery" (HFE 2B) would correspond to success/failure at the reference plant ("real" action failure).

2.2.4 Operational descriptions

The operational descriptions summarize the individual crews' performances with respect to the various HFEs, and are built by reviewing the DVD recordings and focusing on how the crew handled a set of tasks identified before the data analysis: for instance, for the task "start B&F," the focus is on how the crews monitored the SG levels, and how they worked to depressurize the SGs in the complex scenario version.

Based on the individual operational descriptions, the crews that acted or reasoned in a similar way for the task “start bleed and feed” were divided into operational sub-groups called “operational modes.” The operational modes describe the way one or several crews handled the task, including the inferred motivation for their actions.

2.2.5 PSF assessment: Observational and HRA ratings

During the DVD review the analysts evaluated the effects of a set of PSFs on HFE success or failure. Some PSFs, such as training, experience, and human-machine interface (HMI), were evaluated before the video review because they were equal for all crews. The rest were assessed when deemed to present: in this case the analysts commented on them, rated them as positive or negative, and weighted them based on their assumed effect on HFE success (small, big, or no effect). The assessed PSFs assumed the working definitions adopted by the assessment team and reported in [2]. The PSF definitions were:

Table 2-1 Performance-shaping factors – definitions.

Factor	Definition
Adequacy of time	The adequacy of time relates to the difference between available time and the required time. The available time is estimated based on an expected evolution of the scenario, which defines when performing the action modeled by the HFE can no longer be effective in reaching the success criteria. The required time is an estimate of the time needed by the crews to perform the cognitive and execution components of the task. The adequacy of time affects the assessment of the HEP simply because there may be a shortage of time to get the actions done as well as by allowing opportunities for checking the performance of the action, detecting errors, and correcting these errors.
Time pressure	Time pressure refers to the crews' subjective perception that there is a limited amount or shortage of time available to accomplish the required tasks. In many methods (and in NUREG-1792 [5]), time pressure is addressed as a component of or contributor to stress. The crew's perception of the available time can differ from the time actually available in the scenario. Consequently, the crews may experience or report time pressure when the adequacy of time is good; conversely, they may not feel time pressure although the adequacy of time is poor.
Stress	Effect of high workload, perceived time pressure, urgency, perceived threat on performance, perceived severity of consequences, perception/effect of losing overview and control over the situation.
Scenario complexity	Difficulty of situation assessment and diagnosis. Related to ambiguous situations (e.g. masking), diagnosis complexity, and the need to decipher and combine numerous indications, alarms, and other sources of information in order to assess the situation. The number of simultaneous goals influences both scenario complexity and execution complexity. If it involves prioritization, it probably should be listed under "Scenario Complexity," which deals with decision-making, planning, etc. If it involves the management and coordination of tasks, it should probably be listed under "Execution Complexity." In many PSF frameworks, it relates to the indications of conditions (availability of cues, ease of perceiving these cues, the difficulty of interpreting these indications.)
Indications of conditions	Availability and clarity of key indications and/or alarms. This is affected by the availability of instrumentation and, given that the instrumentation is available, the salience of cues, signal-to-noise, ambiguity of cues. In some cases, also the availability of system feedback for execution. This factor is often treated in "scenario complexity," although the latter has a larger scope.

Factor	Definition
Execution complexity	Difficulty of performance (implementation) of the task (not including situation assessment, diagnosis etc.). The number of steps to be performed, whether the task is associated with a single variable or multiple variables, non-linear response of the system, so that you need “to have a feel” in order to adequately control, and whether special sequencing or coordination of multiple performers is required are features of a task that increase the execution complexity. The number of simultaneous goals influences both scenario complexity and execution complexity. If it involves prioritization, it probably should be listed under “Scenario Complexity,” which deals with decision-making, planning, etc. If it involves the management and coordination of tasks, it should probably be listed under “Execution Complexity.”
Training	The degree of familiarity with the scenario and the actions to be performed that can be expected based on the training of the crews. Includes both “theory”/knowledge (classroom) and practice (e.g. in training simulator). The factor should consider not only the amount or general quality of training but also the applicability of the training in the specific scenario, i.e. how helpful the training received will be in the scenario. (In rare cases, the training may even be counterproductive.) Note: HRA analyses deal primarily with training as it concerns the behavior of the NPP and the appropriate situation-specific response. In data analysis, training also includes training on how to solve problems in general, etc. In predictive analysis, it is frequently combined with experience.
Experience	Familiarity and practice of the personnel with the task being analyzed. Although correlated, it is not equivalent to the amount of experience of the crew (e.g. number of years in position). Like training, in rare cases, experience may be counterproductive.
Procedural guidance	Support provided by the procedure for performing the situation assessment (decision-making) and execution of the specific task being analyzed. In the context of the scenario of interest, steps that are ambiguous, unclear (including layout), or not detailed and situations where the way to proceed through the procedure is unclear contribute to a poor rating for this factor.
Human-Machine Interface	Ergonomics, including the presentation and labeling of process parameters, the availability of feedback following an action on a component or system, and the interface for acting on components or systems.

Factor	Definition
Work processes	Refers to the way of working and mechanics of work, e.g. the care taken in reading procedures and generally in performing the task work. Task work is referred to as the work directly with the process as opposed to teamwork work which is about the collaborative aspects of work. Task work can be analyzed at a more individual level than teamwork. In a predictive analysis, this factor indicates how well the expected work processes match the given scenario and how sensitive the task may be to work practices. In analyzing an actual performance, this factor is rated poor if individual work is not thorough, and if the general handling of the procedures is less than adequate. Note that in fast-moving scenarios, “good” work processes may have a negative effect on task success. In the given study RO and ARO sometimes perform process work together as a close unit. In these cases this is analyzed as work process and not teamwork.
Communication	In a predictive analysis, this factor refers to (a) the impact of the environment, e.g. noise, and the hardware used for communication, e.g. an intercom, on task success, as well as (b) the “communication requirements” of the task. These requirements may be viewed as contributing to scenario complexity or task complexity (depending on whether the communication is about situation assessment or what to do). In analyzing an actual performance, refers to the successful exchange of information (e.g. failure to provide information or feedback) and the adherence to communication practices and protocols (e.g. repeat-back, communicate parameter values *and* trends).
Team Dynamics	This factor is often labeled teamwork. It relates to the management of the team, e.g. the adequacy of leadership and support, coordination, sharing of information, proactive communication, questioning attitudes, treatment of suggestions, and sharing and allocation of tasks and responsibilities. In analyzing actual performance, this factor is rated poor also when assessments and decisions are made without review, e.g. without following meeting practices. In a predictive analysis, this factor represents the requirements of the task in terms of good team dynamics and how the expected teamwork matches the requirements, i.e. how sensitive the task may be to the quality of team dynamics.

All PSFs are evaluated only in view of the given HFEs (e.g., the task “start bleed and feed”). For instance, if training, procedures, or indications are poor for the previous and partly concurrent task “re-establish FW,” they will not be included in the PSF evaluation of the HFE “start feed and bleed.” On the other hand, problems that the crews might have experienced in other tasks, such as difficulty depressurizing SGs due to poor procedure guidance or training, are taken into account if they had a direct impact on establishing feed and bleed (e.g., they are included in the assessment of scenario complexity if they increase the total workload of the crew members).

We used the following procedure to aggregate individual crews' PSF ratings into similar-performing crews' PSFs, and, finally, into overall-HFE observed PSF ratings (i.e., ratings for all crews):

1. Crew by crew ratings: After observing each crew's performance of the HFEs, we created a table evaluating whether each PSF was present, and, if so, whether it had a small, large, or null effect on the fulfillment of the HFE success criteria.

2. Grouping of crews: Based on the quality of performance (failures, near misses, operational problems), the crews were assigned to groups, normally well-performing vs. less well-performing crews.

3. PSF aggregation for groups of crews (well- and less well-performing): The crews within each group typically showed consistent PSF configurations (e.g., less well-performing crews had negative team dynamics, whereas well-performing crews had positive dynamics).

4. Contrast analysis, overall observed PSF rating for each HFE: PSF aggregations for well-performing crews were contrasted with aggregations for less well-performing crews to produce the overall observed PSF rating for each HFE, and, if any, the "secondary effect" (i.e., the different effect of the factor on the less well-performing crews).

For example, for HFE 1A, the majority of the crews belonged to the well-performing group; thus, the majority group dominated the main effect evaluation of the final PSFs. If both groups had the same sign on a given PSF (e.g., good communication), the final rating had the same sign, and the weight was assigned based on the number and weights of the observations and the number of total crews. If the two groups had different signs (e.g., team dynamics was positive for the well-performing crews and negative for the others), then a secondary effect was singled out and used as the rating for the minority group.

2.2.6 HRA PSF ratings

When integrating crew-level PSFs into overall-HFE observed PSF ratings, we did not try to use a single "orthogonal" set of PSFs. Some of the PSFs were recognized as partly overlapping, and judgments were made to ensure consistency within each HFE and across HFEs. The overall observed PSFs were translated into a format appropriate to HRA, that is, in terms of factors familiar to the HRA community and consistent with the general assumptions of HRA (e.g., nominal conditions are good). The observational ratings were mapped on the following scale for HRA ratings:

- MND = Main negative driver
- ND = Negative driver
- 0 = Not a driver
- N/P = Nominal/Positive (i.e., contributes to the overall assessment of the HEP being small; note that some methods use the term "Nominal" to denote a default set of positive circumstances, and our use of the N rating is consistent with that terminology)

We used the following rules to translate overall observational PSF ratings into HRA PSF ratings:

1. If there is a secondary negative effect (i.e., the PSF causes problems to some crews), then the HRA rating is negative, even when the observed main effect is positive.

2. If a factor has no observational effect and all crews are consistent on that factor, then the HRA rating is nominal/positive (N/P).

3. If a factor has no observational effect but the crews differ on that factor, then the HRA rating is 0 (no effect).

4. If stress and time pressure have no observed effects, then the HRA rating is 0 (no stress or time pressure).

5. The crew factors (team dynamics, communication, and work practices) are rated as nominal when the observational rating is positive.

In addition to these "rules of translation," one PSF is identified as a main factor (MND) for some HFEs, meaning that, although it might be rated no stronger than other PSFs, it had a larger effect on the performance of the HFE, or may even have caused other PSFs to assume non-nominal, non-zero values.

2.2.7 Difficulty and ranking of the HFEs

The HFEs were ranked relative to their difficulty, as shown in the empirical data. This evaluation was made by considering all available information on the simulator crews' observed performance of the tasks comprising the HFEs. This implies that the HFE ranking is not based on a mere counting of "failing crews," but took the following into account:

1. The number of observed "failing" crews, according to the HFE criteria.

2. Observed "near misses," such as depressurizations that missed the target by a few bars but were not considered failures.

3. Observed difficulty in operational terms, including the difficulties faced by teams that succeeded with suboptimal performances.

In determining the ranking, all available information on the observed performance of the crews in the simulator in the tasks making up the HFEs, was considered. Thus, the HFE ranking is not based on mere counting of "failing crews."

The final ranking was reached by group consensus, in which both experimentalists and the assessment group participated. This empirically based ranking was used in the comparisons with the predicted outcomes from the HRA methods. The ranking was only defined by the empirical data, and was not adjusted in any way during the comparison process (i.e., the ranking has not been influenced by the HRA method predictions).

2.3 HRA predictive analyses performed for the study

2.3.1 HRA analysis inputs

The HFE definitions described in Section 2.1.3 were provided to the HRA prediction teams for analysis. A prerequisite for HRA analysis in a PRA is the analysts' familiarity with the background, training, and experience of the operators (the crews); the performance conditions (e.g., human-machine interface and job aids, such as procedures); and the PRA and the plant response (e.g., thermal-hydraulic plots). In the Empirical Study, however, it was not possible to allow the HRA teams to perform the full scope of familiarization tasks, such as a plant visit, observations of the crews, walkthroughs of the tasks, and interviews with crews or training personnel, due to logistical considerations and the need to ensure that all HRA teams obtained consistent information. To compensate, the information package compiled by the Halden staff and the assessment group, documented as much of information as possible. Furthermore, the HRA teams requested and received additional information in a question-and-answer process, the responses from which were provided to all teams.

It is noted that the HRA teams became more familiar with the crews' behavior, as well as with the HAMMLAB setting, when they were performing the LOFW analyses because they had the opportunity to see the results of the SGTR scenarios and discuss them with the experimental staff during the Phase 1 three-day workshop.

2.3.2 Reporting of HRA analyses and predicted outcomes

HRA methods differ in terms of the underlying human performance and human reliability models, the number of performance-shaping factors, the definition of their scope, and terminology. Additionally, HRA analysis documentation in PRA is typically oriented to tracing how the information on the performance conditions (obtained in the qualitative analysis) has been incorporated into the estimation of the HFE failure probability, rather than to predicting specific outcomes in terms of behaviors and actions. To address the terminological differences and provide predicted outcomes that could be compared with the outcomes obtained in the simulator study, the HRA teams were asked to deliver the following:

- Predictions for each HFE in a three-part "open-form" questionnaire (Form A), where the teams reported (1) the human error probability (HEP), (2) the PSFs, and (3) the "operational expressions" (see below).

- The "normal" documentation of their HRA analysis and quantification, as in a PRA.

2.4 Methodology for assessment of HRA methods

The methodology developed to assess the HRA methods in this study includes multiple criteria. Assessments of the methods' qualitative and quantitative predictive power are based on comparisons between each method's predictions and the reference data obtained in the HAMMLAB simulator. Assessments with other criteria (traceability, guidance, and insights for error reduction) are mainly based on examination of the submitted HRA analyses.

2.4.1 Assessment criteria

The criteria include:

- predictive power:
    - qualitative predictive power in terms of driving factors (drivers) of performance
    - qualitative predictive power in terms of operational expressions
    - quantitative predictive power (to the extent that this can be assessed in light of the limitations of the reference data)
- traceability of the qualitative analysis and quantification process
- adequacy of the guidance provided by each method for the qualitative analysis and for quantification of an HFE
- usefulness of the HRA results in human error reduction

The repeatability of the HRA predictive analysis, including both qualitative analysis and quantification, is not addressed in this study's method assessment. Both traceability and adequacy of the method guidance relate to the HRA analyses' repeatability, consistency, and reviewability. In our concept of repeatability, we include consistency when the same analyst repeats an analysis after some time and when the same analyst analyzes two HFEs with similar levels of difficulty (intra-analyst reliability), as well as when multiple analysts analyze one HFE (inter-analyst reliability). Although there are some indications from the study on the methods' repeatability, a comprehensive assessment of method repeatability would require a different study design, particularly one involving multiple HRA analysis teams using the same method (for inter-analyst reliability). In this study, this was the case for only one method, the Standardized Plant Analysis Risk-Human Reliability Analysis method (SPAR-H), which was used by two HRA teams. A follow-up study is currently being performed on a U.S. training simulator, and includes several HRA teams, per method guidance. That study will incorporate more on this topic.

The assessment for each method addresses each of these criteria, in statements that provide a qualitative rating from poor to good (on a five-point scale) for the individual criteria and include the main aspects of how the method performed against each criterion. This assessment accounts for all of the HFEs in the scenarios. The five points represent "poor," "moderately poor," "fair," "moderately good," and "good."

2.4.2 Structure of summary assessment of each method

The assessment for each method addresses the criteria introduced in Section 2.4.1. The specific aspects considered in assessing each criterion are discussed further in Sections 2.4.4 – 2.4.7. The categories used in the assessment are shown in Table 2.1.

An overall judgment of the predictive power of each method in this application is provided, and is based on the assessment of the predictive power of the qualitative and quantitative analyses, as described below. A single overall assessment summing the assessment of all of the separate criteria (e.g., including the guidance, traceability, and insights for error reduction criteria) is not included, since it would to some extent mix dissimilar criteria. At various points, the summary assessment may include some discussion of the strengths and weaknesses of each method.

The process for assessment and comparison was described in detail in [2] and [3], and is outlined below. The qualitative predictions made by the HRA analyses were summarized in terms of negative drivers for the HFEs and the associated failure mechanisms or modes, in the form of operational expressions. This was done with a common set of definitions for all HRA methods, which allowed for a coherent representation of the analyses from the various methods. The method's qualitative and quantitative predictions were then compared to the empirical data. This was performed per HFE, according to the established criteria. HFE by HFE comparisons served as the basis for the overall assessments. Note that while summaries of the strengths and weaknesses of the methods are provided in Chapter 5 below, the details of the results of the assessments of predictive power are not provided in this report, but are presented in [2] and [3].

Table 2-2 Structure of assessment summary for each HRA method.

Criterion	Process step	Criteria
Predictive Power	Overall predictive power	The overall predictive power is assessed, based on the comparisons between the predictions for each HFE and the reference data. See 2.4.2 for discussion.
	Qualitative predictive power - comparison of drivers	Assessment of: • How well the method predicted the specific performance issues and drivers identified in the reference data • Whether the method predicted factors and issues that were not supported by the reference data See 2.4.4 for a discussion of specific aspects of comparison and assessment.
	Qualitative predictive power - comparison of operational expressions	• Assessment of how well the method predicted the failure mechanisms (in operational terms) observed in the reference data See 2.4.5 for a discussion of specific aspects of comparison and assessment.
	Quantitative predictive power - Comparison of the quantitative method predictions with the empirical data	Listed from highest to lowest priority: 1. Potential optimism for the most difficult HFEs 2. Consistency of the HFE (by predicted HEP) with the reference difficulty ranking 3. Predicted HEPs relative to the confidence/uncertainty bounds of the reference data 4. Quantitative differentiation of the HFEs by HEP See 2.4.6 for discussion of specific aspects of comparison and assessment.
Traceability	Assessment of traceability	• Traceability of the basis for quantification inputs • Traceability of quantification See 2.4.7 for discussion.
Guidance	Assessment of guidance	• Guidance for the qualitative analysis • Guidance for modelling the HFE and decomposition (if applicable) • Guidance for the quantification See 2.4.8 for discussion.
Error reduction	Insights for error reduction	See 2.4.9 for discussion.

2.4.3 Comparison of method's qualitative predictions

The qualitative predictive power considered three aspects in terms of drivers and operational expressions.

Drivers:

- How well did the method predict the specific performance issues and drivers identified in the reference data?

- Did the method predict factors and issues that were not supported by the reference data?

Operational expressions:

- How well did the method predict failure mechanisms in operational terms that were identified in the reference data?

These aspects are discussed in the next two sections.

2.4.4 Comparison of method's qualitative predictions in terms of drivers

- *Prediction of the drivers identified in the empirical data, including the associated performance issues.* Did the method identify the correct task performance issues? In addition to identifying a driving factor, did the method's explanation of why the predicted driver contributes negatively to HFE performance correspond to the empirical data? Given some of the differences in factor definitions among the methods, this emphasis on the drivers in operational terms and in terms of specific issues bypasses possible ambiguities with the assignment of issues to specific PSFs (the "translation" problem). Some methods may not identify specific performance issues, but may identify the correct drivers. Such methods would be ranked lower with respect to this criterion than methods that did identify the performance issues.

- *Predicted factors and issues that were not supported by the reference data.* In contrast to the preceding subcriterion, this one starts with the factors and issues predicted by the HRA analysis. Did the HRA method predict drivers and performance issues that were not observed in the simulator or shown not to be a performance issue for the crews? The assessors accounted for the fact that crew performance tends to be fairly high (i.e., low HEPs), and that there may be issues that are correctly predicted but not observed, given the sample size. In contrast, if a driver was confirmed in the small sample of observations, then it is likely to be a significant driver. Such drivers are addressed by the previous subcriterion.

2.4.5 Comparison of method's qualitative predictions in terms of operational expressions

- *Prediction of failure mechanisms in operational terms.* Although HRA analysts need to understand how crews will approach a given task in order to predict the HEP, some methods rely strongly on these operational aspects, and many predict specific modes or mechanisms of failure. This subcriterion deals with the accuracy of these predictions. Did the HRA analysis correctly characterize how the crews would fail, or where they would have problems? It can be seen that the “driving factors and issues” subcriteria focus on the problematic performance conditions, while this subcriterion focuses on the manifestation of degraded or failed performance.

2.4.6 Comparison of quantitative predictions (including ranking)

The comparison of the method’s quantitative predictions with the reference data addressed both the absolute values (HFE by HFE) and the ranking of the HFEs based on the HEPs (across the HFEs). First, the small sample of observations results in large uncertainties in the reference HEPs, so the accuracy of the HEPs is difficult to assess. Secondly, in many PRA applications, the relative values of the HEPs (i.e., the ranking of the HFEs) are sufficient to draw conclusions and derive safety insights. The subcriteria identified below are listed in order of decreasing priority.

- HFEs in which several failures were observed in the empirical data can be regarded as very difficult tasks that should have correspondingly high HEPs. If an HRA method produced low HEPs for such HFEs, the submission was examined in more detail in order to identify indications of systematic method optimism.

- Consistency of the ranking of the HFEs (by predicted HEP) with the reference difficulty ranking. In the analysis of the simulator observations, the HFEs were ranked in terms of difficulty (i.e., the assessors produced a rating/ranking of the likelihood of failure in the HFE tasks while documenting the crew performance). Despite the large confidence interval for the reference HEPs (in terms of the empirical error rate), it was possible to obtain a strong consensus on which HFEs appeared to be more difficult, with the expectation that the probability of failure was higher.

- Predicted HEPs relative to the confidence/uncertainty bounds of the reference data. Were the HEPs within the bounds, which in this study have been estimated by a Bayesian update that uses the observed performances as evidence (see Section 4.5 for the derivation of the confidence intervals for the empirical HEPs)? Note that the uncertainty bounds predicted by the HRA teams for each HEP are not utilized in the current comparison.

- Quantitative differentiation of the HFEs by HEP. Were the predicted HEPs for the most difficult HFEs significantly larger than those predicted for the least difficult HFEs? The quantitative predictive power of the method is judged to be reduced if the HEPs predicted for HFEs with a wide range of difficulty fall within a narrow band.

As noted above, the predicted ranking of the HFEs is based solely on the HEPs from the HRA analyses. On the other hand, the reference or empirical ranking of the HFEs is not solely based on the empirical HEPs, but is instead based on an overall, partly subjective assessment of the relative difficulty (a relative failure likelihood) that combines the Bayesian HEP results with qualitative considerations of the performance. The qualitative considerations accounted not only for the failure counts, but also for other objective evidence from the experiment, such as the performance as measured by plant parameters, the amount by which the success criteria were missed (in terms of the time windows defined for the HFEs or the plant parameters), and the difficulties experienced by the crews (even if these difficulties were overcome) during the tasks associated with the HFE.

2.4.7 Assessment of traceability

The assessment of traceability examines:

- The basis for the quantification inputs obtained in the application of the HRA method. For instance, it examines the derivation of the PSF ratings from the qualitative analysis, or the identification of the failure mechanisms associated with operational narratives. In both cases, the assessment looks at how the HRA method and the documentation of the application of the method (of the HRA analysis) establish the link between the qualitative analysis and the quantification inputs (the probabilistic safety analysis (PSA) ratings). How did the issues and factors identified as relevant and important to HFE failure translate into PSF ratings or identified failure mechanisms?

- The documentation of the quantification of each HFE. This part of the assessment of traceability looks at the link between the quantification inputs and the HEP values. Is expert judgment involved in deriving the HEPs from the quantification inputs? If so, how large is the role of expert judgment? Alternatively, is the quantification based on a mathematical, fully repeatable algorithm?

2.4.8 Assessment of adequacy of method guidance

The assessment of method guidance examines:

- The guidance for the qualitative analysis. Relevant questions include: (1) To what extent does the method provide guidance for performing the qualitative analysis, and how does this guidance contribute to a comprehensive assessment of the performance-shaping factors or contextual factors in terms of how they may affect the probability of HFE failure? (2) Does the method guidance clearly describe the required or expected scope of the qualitative analysis? (3) To what extent does the guidance for the qualitative analysis appear to support inter-analyst consistency? (This last question is also related to repeatability; see the remarks on Assessment Criteria in the conclusion of Section 3.1.)

- The guidance for HFE modelling and decomposition (if applicable).

- The guidance for the quantification. For those methods where factor ratings are used to translate the qualitative analysis into quantification, what guidance is available to support the rating of the factors? For those methods where quantification includes expert judgment, what guidance or aids are available to support the expert judgment process and its consistency?

2.4.9 Insights for error reduction

This assessment addresses the degree to which the qualitative analysis and evaluation of performance influences addressed by the HRA method provide information that would allow insights into reducing error. In other words, do the analysis of driving factors and the understanding of potential failure mechanisms support the identification of potential fixes in areas where errors might occur (e.g., procedural or training improvements)? The overall ability of the method to produce this information was judged.

3. EMPIRICAL HAMMLAB RESULTS

This chapter presents the overall results of the empirical analysis, as well as a few examples of detailed results. The overall results are expressed in terms of the human failure event (HFE) ranking, which is based on difficulties, success and failure, and failure bounds. The detailed results are exemplified by operational descriptions and performance-shaping factor (PSF) evaluations of two sample HFEs. The empirical reference data correspond to the types of predictions requested of the human reliability analysis (HRA) teams. For details on the steam generator tube rupture (SGTR) and loss of feedwater (LOFW) empirical results, please see [2] and [3], respectively.

3.1 Variability of performance

Some scenarios and/or parts of the scenarios in this study were designed to be very challenging for the crews. In general, the crews managed according to expectations, despite the reduced team size (see [1], [2], and [3]). However, large degrees of variability in the performance quality were observed.

One rationale for designing a base and a complex version of the basic scenarios was to add extra HFEs, as well as to produce performance outcome variability (i.e., to create the possibility of observing both failure and success). One way to increase the complexity of some HFEs was to define them by setting time limitations on the performance of their tasks. Another strategy was to mask key indications of plant conditions, either by concurrent failures that affected the main problem or by failures in instrumentation. Both cases resulted in mismatches with important steps within the emergency operating procedures (e.g., procedure transition conditions not met or delayed).

The second purpose of devising complex scenarios was to create the possibility of observing performance quality variability. Process variability beyond outcome variability (i.e., variability in performance quality beyond the mere failure counts) was necessary to rate the difficulty of the HFEs and create a rank order to be compared with human error probability (HEP) rankings of the HRA predictions. It was also important to create a broader spectrum of performance issues and test the methods' capabilities to identify these.

The empirical data analysis showed a large degree of variability in the ways in which different crews responded to the scenarios and the HFEs (i.e., crew-to-crew variability) in both the SGTR and the LOFW data. This variability made the evaluation of the overall difficulty of the HFEs and the overall estimation of the PSFs for the given HFEs all but trivial, since these are normally averaged between all crews. Some variability arises from differences in initial crew actions in the scenarios, which leads to cascading differences in plant configurations for the upcoming HFEs. Other sources of variability relate to differences in the teams' internal functioning (i.e., how the crews members interact with each other).

An example of the first type of variability is the different procedures' observed progressions, the fact that different crews moved differently though the procedures set. Different progression times, different transfer criteria, different procedures, differences in looping and in parallel following affected the scenario dynamics, and, thus, the timing and character of the plant information available to the crews.

The second type of variability relates to the team's internal functioning, for instance, to the allocation of tasks, to how information is communicated and sought, or to the decision making process (teamwork factors). This type of variability is generally not treated within the scope of most HRAs; however, the empirical observations confirmed that the team's functioning was an important underlying cause of differences in crew performance.

The comprehensive emergency response guidelines did not restrict crew responses within strict boundaries,[2] simply because the emergency procedures did not cover the situational variations created or prompted by some of the study's scenarios in enough detail. Instead, the procedures largely required the operators to make autonomous judgments. The cognitive demands associated with such autonomous judgments and with following the emergency procedures in these types of scenarios, were not always recognized or satisfactorily accounted for by all HRA analyses.

Overall, the study suggests that HRA analysts may not give sufficient attention to variability of scenario development, which is caused by complicating factors, cognitive requirements of procedure-following, and differences in teamwork and expertise.

3.2 HFE difficulty ranking

The difficulty ranking of the HFEs is a qualitative assessment of their observed difficulty. Two rankings were obtained, corresponding to the two main phases of the Empirical Study. All HFEs in the two SGTR scenarios were compared and ranked for phase 2; all HFEs in the two LOFW scenarios were compared and ranked for phase 3. The HFEs in the SGTR are not compared to the HFEs in the LOFW scenarios. The methodology for the ranking is described in Section 2.2.7. In short, the ranking accounts for (1) the number of observed failures, (2) observed near misses, and (3) observed difficulty in operational terms.

3.2.1 SGTR difficulty ranking of HFEs

The HFEs of the SGTR scenarios are ranked as follows (from difficult to easy):

5B1 > 1B > 3B > 3A > [1A ~ 2A ~2B] > 5B2 > 4A

The ranking of the HFEs in the SGTR scenarios is shown in Table 3-1, in descending order of difficulty. The number of near misses is included in the column "Crews with operational problems," while the operational difficulties are explained in "Comment on difficulty." For a more in-depth description of the SGTR scenarios and the operational results, see [2]. For a description of the scenarios and the HFE definitions see section 2.1.3.

[2] At least with regard to the individual HFEs' performance times and level of analysis.

Table 3-1 Summary table of ranked HFEs and their difficulties in SGTR.

HFE, most to least difficult	# of failing crews	# of crews with operational problems[1]	Comment on difficulty	Difficulty rating
5B1	7/7	7/7	This HFE required the crews to detect a power-operated relief valve (PORV) leakage within five minutes of closing it when concluding the depressurization. Given this time limit, it was very unlikely that the crews would focus on the PORV status beyond checking its indication (e.g., by checking pressurized relief tank (PRT) pressure and level), as the steps following depressurization would lead them to continue the procedure. The reactor coolant system (RCS) pressure for five out of six crews was increasing when applying E-3, step 18 ("Check RCS pressure – increasing," the step directly after the end of depressurization), or at least stable when applying step 19 ("Check if safety injection (SI) flow should be terminated"). After the HFE time window, clearer indications of RCS leakage would appear to the crew. This was the most difficult HFE of this set.	Very difficult
1B	7/14	7/14	The crews showed difficulties in identifying the presence of an SGTR, due to the concomitant steam line (SL) break and absence of radiation indications. The majority of the crews did not transfer to E-3 (SGTR procedure) to follow a transfer condition in the procedure set, but instead diagnosed the situation by interpreting the available indications on the plant status, with a rising SG1 level as the primary cue. Eventually all crews identified and isolated the ruptured steam generator.	Difficult
3B	2/14	2/14	Same issues as in HFE-3A, with the addition of a reactor coolant pump	Somewhat difficult

			(RCP)/spray problem. The latter distracted two crews, with one exceeding the 15-minute criterion as a result. In both cases the task requirement for teamwork, in this case especially for leadership, led the crews towards poor outcomes (one too late, one too far from the target). Also, more cases of execution complexity in 3B than in 3A (seven cases of RCS pressure not exactly lower than ruptured steam generator (SG) pressure) and generally inferior teamwork could indicate more stress during depressurization in this scenario.	
3A	1/14	3/14	This task is well trained and covered by the E-3 procedure; however, three crews had problems with concluding the depressurization (stopping too early and/or for the wrong reason) as a consequence of the task, implying some execution complexity (high speed of depressurization, several stop conditions to monitor) and requiring coordination and supervision in controlling and verifying the outcome. There were also several cases of crews not strictly meeting the depressurization end criteria (RCS pressure should reduced to “less than” ruptured SG pressure).	Somewhat difficult
1A	1/14	1/14	All crews identified and isolated the ruptured steam generator. However, there were several occasions of excessive time consumption: evaluation of initial conditions and which procedure to take (i.e., AOP-3 or E-0), transfer to E-3 and the option to hold an evaluation meeting, and complex build-up of isolation step (3) in E-3. As a result, one crew exceeded the time criterion, and four others were less than two minutes away from trespassing it.	Easy to somewhat difficult
2A	1/14	3/14	This task is well trained and covered by the E-3 procedure. As a result, all crews cooled down and	Easy to somewhat difficult

			maintained the RCS temperature under the right table value. However, four out of fourteen crews activated an automatic protection system, which isolates the steam lines, by using full dump while having a large SG-RCS pressure difference (i.e., one activation condition). Three of these crews did not immediately recognize what had happened, and used extra time to complete the cooldown (and typically did so less than optimally). It seems that all crews that used dump (including those who did not activate the protection system and did not have a large SG-RCS pressure difference) forgot to do so with care, instead following the procedure instructing them to use dump at maximum.	
2B	0/14	4/14	All crews cooled down and maintained the RCS temperature under the right table value. The fact that the task had to be performed with previous SL isolation caused two crews some problems in understanding the situation. Execution problems were observed in two other crews using the SG PORVs (not opening them completely, setting set-points upon completion). Furthermore, three other crews wasted some time by waiting for the completion of the local actions for isolation (this condition is not fully captured by the HFE definition, which has its starting point at the cooldown step, rather than at the end of the previous HFE). Stress carried over from the previous HFE in this (complex) scenario could have caused the higher rate of small execution problems observed for HFE-2B, as compared to HFE-2A. In comparison to HFE-2A, however, the crews had only one cooldown modality available (SG PORVs), and thus could not get the SL	Easy to somewhat difficult

			isolation problems.	
5B2	0/7	0/7	The crews train twice a year in the E-3 (SGTR procedure), and they always check the isolation valve before using the PORV. If the PORV is not closing fast enough, they will close the isolation valve. Furthermore, the procedure step for depressurization with PORV (step 17) points to closing the isolation valve if the PORV cannot be closed. The only complicating issue here is the five-minute time limit.	Easy
4A	0/14	0/14	This task is well trained, well described in the procedure, and involves control room actions only. Furthermore, the HFE-4A definition does not specify a time limit for accomplishing the required actions. This is the easiest HFE of this set.	Very easy

[1] Including near misses and failing crews. “Operational problems” refers to the crews’ distinctive actions that brought them closer to failing the HFE.

3.2.2 LOFW difficulty ranking of HFEs

There were six defined HFEs in the LOFW scenarios (see Section 2.1.3). Two of these, 1A1 and 1B1, were joint HFEs. These were used in the predictions by the HRA teams, but, empirically speaking, they cannot be separated from the outcome of the latter of the two combined HFEs; thus, in the empirical results, only four HFEs are described.

The four HFEs of the LOFW scenarios are ranked according to level of difficulty, from difficult to easy:

1B > 2B > 1A > 2A

The ranking of the HFEs in the LOFW scenarios is shown in Table 3-2, in descending order of difficulty. The number of near misses is included in the column “Crews with operational problems,” while the operational difficulties are explained in “Comment on difficulty.” Below the table, some more explanation is given for the HFEs. For a more in-depth description of the LOFW scenarios and the operational results, see [3].

Table 3-2 Summary table of ranked HFEs and their difficulties in LOFW.

HFE, most to least difficult	# of failing crews	# of crews with operational problems[1]	Comment on difficulty	Difficulty rating
1B	7/10	7/10*		Very difficult
2B	0/7	1/7	Weaker success evidence by counts, with one almost failed (by one minute).	Somewhat difficult – Difficult
1A	0/10	1/10	Stronger evidence for success by counts, but one (crew L) almost failed (they reached 5% NR level before acting)**. Unlike the other crews, they were not actively monitoring SG levels because they entered the procedure instructing them to do so late.	Easy - Somewhat difficult
2A	N/A	N/A	No crews observed in this condition. As difficult as 1A at most, but most likely easier.	

[1] Including near misses and failing crews. “Operational problems” refers to crews’ distinctive actions that brought them closer to failing the HFE.
*Of the seven failures, two performed rather well. On the other hand, of the three that succeeded, two had aggravated the scenario (B avoided a transfer to ES-0.1 and C didn’t depressurize SGs according to procedure, causing a high RCS pressure criterion for B&F).
** Crew L SG wide range indications: in the base scenario, the dryout rate was 7% (8% to 1%) in five minutes (between 30 and 35 minutes).

HFE-1B

In case of total loss of heat sink, the emergency procedures instruct the operators to start bleed and feed (B&F) on high RCS pressure, or when the levels in two SGs fall below 12%, as indicated by the wide range (WR) measurements. In order to act according to the procedural criterion after failing to meet the RCS pressure criterion, the crews needed (1) to identify the anomaly in the SG level measurements, and (2) to recognize that two SG level indicators were wrongly indicating that the levels in the associated SG were greater than 12%.

To identify the anomaly in the SG level measurements, the crews needed to monitor the SG WR levels over time, and infer that there was no reason for two out of three SGs to stop emptying. This identification and inference would have been complicated if the crew had relied only on point readings, rather than consulting the trend displays. In the study, the assisting reactor operator (ARO), who usually monitors the levels, had to take on the tasks of the balance of plant operator, who was absent from the experimental crew set-up. This increased the crews’ workload and reduced their overall capacity for monitoring and diagnosing the SG level measurement anomaly. The difficulty of identifying and diagnosing the anomaly was further complicated by concurrent and competing goals: the crews were instructed by the procedures to reestablish feed flow from condensate, and attempted to do so, unaware that the specifics of the scenario design made this impossible. The establishment of condensate flow was a difficult task in itself because of an insufficient procedure step (FR-H.1 step 7), which also added to the crew workload, and, therefore, to the HFE-1B complexity (training for LOFW and the start of B&F is only held every six years, and normally without failing SG levels).

HFE-2B

The diagnosis was still difficult because the SG WR levels indicated more than 12% in two SGs; thus, this B&F criterion would never be met. However, the extra 25 minutes reduced the masking, as the SG level trends after dryout showed only straight lines at different levels. This was a more salient indication of instrument failure, which facilitated the diagnosis of the real SG levels. Furthermore, the condensate pump was tripped for the crews that succeeded in establishing flow, thereby removing the concurrent and competing goal.

HFE-1A

If the crew works as expected, the procedures will guide them to the start of B&F, as the SG WR levels are unambiguous. Still, training for LOFW is seldom held.

HFE-2A

Training for LOFW is seldom held, but for this HFE crews have a lot of time.

3.3 Operational descriptions and PSF assessments

This section presents the operational descriptions and the PSF assessments for two selected HFEs observed in the study, HFE-1B from the LOFW scenario and HFE-3B from the SGTR.

3.3.1 LOFW HFE-1B (B&F before dryout in complex scenario)

LOFW HFE-1B = Initiation of B&F before SG dryout in the complex LOFW scenario. SG dryout occurs when there is no water left in the SGs, indicated by 0% WR SG level.

In the complex LOFW scenario, three out of ten crews started B&F before the SGs were empty. Based on how the crews either made or failed to make the decision to start B&F, we have identified five operational modes.

Table 3-3 Operational descriptions of the crews in LOFW HFE-1B.

	Operational mode	Crews*	Result	Comment
1	The crew **identified and diagnosed** the failing SG level measurement.	G	B&F started when the SG level was 3%.	The reactor operator (RO) detected the abnormal SG levels and concluded that the real levels were below the 12% criterion. He then persuaded the shift supervisor (SS) to start B&F. The crew had problems depressurizing quickly, which gave them more time for the diagnosis.
2	The crews **unwillingly caused the RCS pressure to increase** and started B&F based on the high RCS pressure criterion.	B, C	B&F started based on the RCS pressure criterion, when SG levels were 18% (B) and 12% (C).	- Crew B manually actuated SI earlier in the scenario, as they hesitated over which procedure to use. - Crew C tried to establish condensate flow but failed to depressurize the SGs, causing the high RCS pressure.
3	The crews **identified** the abnormal WR SG levels **but focused on achieving the concurrent goal** of establishing condensate flow.	**M, I**	B&F started four (M) and seven (I) minutes after empty SGs.	The crews suspected that the SG level measurements were incorrect, but worked hard to establish a feed flow from condensate. - In crew M, the SS was aware of the suspect SG levels. - In crew I, the operators suspected that the levels were wrong but the SS did not agree.
4a	The crews **did not identify** the abnormal SG levels and did not monitor the SG level trends.	**N, K, F**	B&F started 6 (N), 9 (K), and 17 (F) minutes after empty SGs.	The crews relied on reading instant values of SG WR levels without displaying trends. This, in combination with the concurrent work in depressurizing the SGs, prevented the crews from detecting the abnormal indications and diagnosing the real SG levels.
4b	The crews **did not identify** the abnormal SG levels, although they monitored the SG level trends.	**L, J**	B&F started 15 (L) and 24 (J) minutes after empty SGs.	Even though SG WR level trends were displayed, the crews, absorbed by the procedure work of restoring feedwater (FW) to the SGs, did not react or stop to analyze the SG level situation before dryout, and thus did not diagnose the real levels in the SGs.
*Bold: Failing crews				

There were three successful crews for HFE-1B (start B&F before empty SGs in the complex case), and two of them started B&F on high RCS pressure. The RCS pressure normally does not exceed the high-pressure criterion until after the SGs are empty. Crew B manually actuated SI because they felt that would take them to a more controlled situation. The pressure rose, and the crew correctly started B&F on the high RCS pressure criterion after the pressurizer (PRZ) PORVs had opened. They were given a good score (4) on the expert rating, based on the fact that SI actuation would increase safety margins. However, manually actuating SI falls outside of procedural guidance, and starting B&F before the SG level criterion is met could imply that the crew missed the opportunity to restore normal feed flow (in case of recovery of failed equipment), which is the solution path embodied by the emergency procedures before the low SG level criterion is reached. If the crew had understood that starting SI (which was not actually needed when it was performed) had caused the RCS pressure to reach the B&F criterion, they could have considered stopping SI again before they decided to establish B&F. Crew C had problems depressurizing the SGs according to procedure FR-H.1. The RCS pressure increased, and they started B&F on this criterion. Since the crew failed to control the RCS pressure, they were rated low (2) on the expert rating. Starting B&F on high pressure in this scenario made these two crews succeed in the HFE, though spuriously. Had the HFE been defined as establishing condensate flow, they would have failed.

The third crew that succeeded in starting B&F before empty SGs had more time. They depressurized slowly, increasing the time to empty the SGs. This result shows the importance of situational dynamics in analyzing performance. This crew was given a low expert rating (1) in spite of their success, because they waited too long to start B&F after identifying the failing SG levels.

Overall, the three crews that achieved “success” (in different operational ways) exhibited operational difficulties or non-standard behaviors. This fact made it difficult to identify clear patterns for certain PSFs in this HFE: for example, work practices and team dynamics may be rated poorly regardless of HFE outcome.

Independent of the HFE definition, all crews except one (the crew that started SI early in the scenario) seemed to be distracted by the task of starting condensate, and had some problems identifying the failing SG level measurements. Given their simultaneous occurrence, it is impossible to assess which of the two circumstances had the greater influence on crew performance.

Table 3-4 Overall PSF evaluation for LOFW HFE-1B.

Overall PSF evaluation for HFE-1B			
HRA	**Observational***	**PSF**	**Comment**
ND	-1	Adequacy of time	If the criterion to start B&F is detected, there is adequate time to start it before the SGs empty out. However, depressurizing the SGs in procedure FR-H.1 step 7 will reduce the time for them to empty out.
0	0	Time pressure	No observations of time pressure for starting B&F in most crews.
0	0	Stress	Differences in stress did not systematically differentiate HFE success or failure. Stress was observed in all successful crews, although two out of three successes were “spurious” (i.e., crews who wrongly caused a second condition for performing B&F). Stress was observed in both failing and successful crews: in two

			crews (one successful and one failing), the SS had no overview and seemed stressed. In four crews (two successful and two failing), the ROs were stressed and had problems with their procedure work, and, in one case, with finding information in the human-machine interface (HMI). In crew G (the only “real” success), the SS was stressed, which might have delayed B&F because the SS did not seem to take in the RO’s warning that the levels were low. Given the reduced crew staffing in the experiment, see Scenario Complexity for workload related to the task of reestablishing FW.
MND	-2	Scenario complexity	The task of depressurizing SGs preoccupied the crews and took focus away from analyzing the SG WR levels (which are seldom used during normal operation). The procedure-directed task of SG depressurization made the detection and analysis of failing SG level measurements more difficult, as did the fact that the crews were working towards the concurrent goal of establishing condensate flow. In some cases this focus on a concurrent goal made the crew members ignore any doubts about the SG level measure. Two out of three WR SG levels were failing, making it difficult to meet the criterion to start B&F at the right time (this criterion would literally never be met, as two SG WR levels would always be above 12%). Complexity was even higher for crews that did not display trends on SG WR levels. The scenario complexity was further increased by the higher-than-normal workload, given the reduced staffing of the crews in the experiment (the balance of plant operator was absent, and the other crew members had to organize themselves in order to perform these tasks). This reduced the crew’s capacity for identification, diagnosis, and communication of the SG level condition.
MND	-2	Indication of conditions	The criterion to start B&F on SG levels is masked by the failure of two of three SG levels: the criterion of two SG WR levels falling below 12% will literally never be met, as the scenario design prevents this from happening.
N/P	0	Execution complexity	The three crews that started B&F before empty SGs did not have any difficulty in executing B&F.
0	1	Training	LOFW with the start of B&F is simulated every six years. Training on failing SG levels is not offered during LOFW, but the crews are trained on failing measurements in other scenarios.

0	0	Experience	Differences in experience did not differentiate the crews' performance. All crews had some experienced operators. In six crews, two well-performing and two not well-performing, all operators were experienced.
ND	-1	Procedural guidance	The procedure doesn't cover the failing level measurements, and consequently will not guide the crew to start B&F on SG levels. The B&F start criteria are presented in a warning in FR-H.1 step 2, and must be monitored continuously. Furthermore, the procedure assumes that the condensate pumps give normal pressure, and depressurization strictly following FR-H.1 was not enough in the scenario, which adds to the mismatch. On the other hand, the procedure FR-H.1 guided the two crews that caused high RCS pressure (C and B) to start B&F.
N/P	0	HMI	The HMI is screen-based, unlike the conventional HMI at the crews' home plant. The crews had training before the experiment, and we did not observe any difficulties with the HMI in this scenario. One of the unsuccessful crews (F) had trouble finding the indications necessary to depressurize, but figured it out. This difficulty had no observed direct impact on starting B&F.
ND	-1	Work processes	High requirements for work processes, particularly in terms of careful monitoring. The aggregate evaluation of the effects of this PSF on this HFE is complex, as two of the three crews who started B&F before dryout did it on the "wrong grounds." Overall, because not well-performing crews (F, J, L) had a prevalence of negative work processes, and because this was also noticeable within the majority of the remaining crews, the main effect of this PSF is negative. One less well-performing crew (L) had major problems related to work processes: the RO was not careful in following the procedure, made mistakes in procedure reading, and tried to anticipate the procedure without understanding it well enough. This slowed down the work and prevented the crew from working systematically to understand and control the situation. The SS was too busy controlling the RO's poor work, and lost the overview. One of the crews that started B&F based on high RCS pressure (C) also showed negative work processes in terms of problems with understanding FR-H.1 step 7, which caused the RCS pressure to increase (and caused the crew to start B&F on high RCS pressure criterion). General examples of observed negative work processes include poor procedure reading (not moving forward, not reading foldout page and warnings), not

			working in given roles (SS too involved in details), and poor monitoring of the SG levels (e.g., not displaying trends). We also observed positive work processes, such as good procedure reading, good division of work, and good monitoring of SG levels. In the only “real” successful crew (G), the RO focused on the goals and carefully monitored the SG levels, which led to successful identification of the failing measurements and the start of B&F before dryout. Additionally, three crews, two successful and one unsuccessful, had both positive and negative work processes.
0	0	Communication	Examples of both positive and negative exchanges of information, but without consistent or significant effects on the HFE.
ND	-1.5	Team dynamics	High requirements for the SS in maintaining overview, guiding and leading, and effectively using crew resources. We observed a prevalence of these negative team dynamics in the not well-performing crews (F, J, L), and one of the crews starting B&F on high RCS pressure also showed negative team dynamics: they started a meeting that never ended, which prevented them from getting to the right procedure and doing timely work (this actually caused them to start SI when they could not find the transfer to FR-H.1 and the high RCS pressure). Positive team dynamics were also observed.
* Main observed effect and secondary effect (i.e., the effect of this factor on the crews that had operational problems).			

3.3.2 SGTR HFE-3B (depressurization in complex scenario)

Performing the depressurization in the complex scenario was different than in the base scenario. In HFE-3B, an extra malfunction to one RCP pump was set, strongly reducing the effectiveness of the spray (one train was still available).

The crews started reading step 16 in E-3 about 21 minutes after transferring to E-3 (the same as in the base scenario) and about 37 minutes after the tube rupture (with a range from 14:41 to 26:29 and 28:38 to 59:25, respectively). Seven crews stopped the depressurization without the RCS pressure falling below the ruptured SG pressure (the procedure instructs the crew to depressurize the “less than” ruptured SG), although only three cases had a pressure difference greater than 2 bar. One crew exceeded the time criteria for depressurization. No crew had to stop the depressurization because of the PRZ level exceeding 75% or because of losing subcooling.

The average time to stop depressurization after entering step 16 was 5:50, with a range from 2:22 to 16:26.

Table 3-5 Operational modes observed in the SGTR HFE-3B.

	Operational mode	Crews*	Result	Deviation/Comment
1	These crews followed the procedure using spray and then PORV.	F, H, L, N	RCS pressure *below* SG pressure by 1.1 (F) to 4.7 (N).	
		B, **E**, G	RCS pressure *slightly above* SG1 pressure by 0.8 (G) to 2.8 (B) bar.	- Crew E spent five minutes on a meeting discussing the RCP problem without mentioning the PORV option. It took five more minutes for the RO to transfer to step 17 (depressurization with PORVs). Total time for depressurization for crew E was 16:26.
2	These crews stopped the PORV before the RCS pressure fell below the ruptured SG pressure, and reopened spray to complete.	A, **C**	RCS pressure above SG1 pressure by 3.9 (A) 4.2 (C).	- Crew A: While the ARO was communicating that he was using spray after closing the PORV, the SS reported that the PRZ level was approaching the criterion for stopping depressurization (75%). The ARO stopped the spray. Although the crew thought that they were using the spray, they never got it to work. - Crew C: After closing the PORV at about 78 bars, the SG1 pressure decreased quickly from 71.5 (the crew had cold RCS and a large RCS-SG pressure difference). When the RCS was depressurized with spray to 72 bar, the SG1 pressure fell to 68 bar.
3	These crews used PORV only, which they decided before starting spray (in step 16).	I, K, M	RCS pressure *below* SG pressure by 3.3 (I) to 6.3 (K).	- In crew K, the SS stopped the RO from starting the spray, and they changed to PORV very quickly.
		D, J	RCS pressure *slightly above* SG1 pressure by 0.3 (D) to 1.44 (J).	

* **Bold**: Crew E exceeded the allotted time for depressurization, Crew C could not meet the "less than" condition.

Table 3-6 Overall PSF assessment for SGTR HFE-3B

Overall PSF assessment for HFE-3B			
HRA	**Observational***	**PSF**	**Comment**
0	0	Time pressure	Normally not, but four crews pointed to the need for quick work.
ND	0 (-1)	Stress	Indications of stress for less well-performing crews (possibly carried over from the difficulty in identifying the SGTR). The fast depressurization rate with PORV, when three stopping conditions had to be monitored at the same time, could also have caused many crews to stop the depressurization too early. Two crews planned to “fine-tune” the final pressure with spray outside of both procedural guidance and standard practice, which could also be a sign of stress.
ND	0 (-1.5)	Scenario complexity	Two crews were distracted from the main task of fast depressurization by the minor RCP problem. Most other crews had a good understanding of the situation.
N/P	0	Indication of conditions	
ND	-1	Execution complexity	Seven crews stopped the depressurization too early, without waiting for the SG pressure to drop. The depressurization went quickly, and the crew needed to continuously follow several parameters. Tendency to set target to SG pressure and not below SG pressure. Some crews might have expected more of a delay between closing order and actual closing of the PORV. There were multiple stopping conditions for depressurization, including monitoring subcooling margins and the fast-moving PRZ level.
N/P	1	Training	Well-trained task.
0	0	Experience	Experience level did not affect performance levels.
N/P	1	Procedural guidance	The procedure guided/supported the crews during depressurization. No observations of problems in the procedural guidance.
N/P	0	HMI	
0	0	Work processes	Several crews did not follow the transition between steps 16 and17 correctly, and some did not read notes and warnings, although this didn’t affect the HFE.
N/P	1	Communication	Normally good communication.

<table>
<tr><th colspan="4">Overall PSF assessment for HFE-3B Continued</th></tr>
<tr><td>ND</td><td>1 (-1.5)</td><td>Team dynamics</td><td>Lack of coordination and leadership for less well-performing crews (as well as instances in other crews). Otherwise normally good coordination and supervision in well-performing crews.</td></tr>
<tr><td colspan="4">* Main observed effect and secondary effect (i.e., effect of this factor on the crews that had operational problems).</td></tr>
</table>

4. OVERALL QUANTITATIVE RESULTS

The methodology for comparing the human reliability analysis (HRA) method predictions with the empirical Halden Human-Machine Laboratory (HAMMLAB) results is outlined in Section 2, and both quantitative and qualitative comparisons were performed. The various types of quantitative comparisons and the criteria for rating quantitative predictive power are described in Section 2.4.6. In these comparisons, the mean human error probabilities (HEPs) from the HRA methods are used and compared against the 90% confidence bounds of the reference empirical HEPs, which were obtained in a Bayesian update using the HAMMLAB data as evidence. In the present chapter, the quantitative predictions of all methods are presented, the empirical HEPs and their derivation are discussed, and an overall comparison of the method predictions against the empirical HEPs is presented. The detailed comparisons and assessments for each individual method are given in [2] and [3], while in this report the final results and conclusions are presented.

4.1 Role of the quantitative data

Drawing definitive conclusions from the quantitative results is limited because of the small set of observations. The empirical HEPs are derived from observations of 14 crews for the steam generator tube rupture (SGTR) scenarios and 10 crews for the loss of feedwater (LOFW) scenarios.[3] Considering the HEPs' expected range of values, particularly for those response actions where the HEPs would be expected to be low, this is a small set of observations. A Bayesian update was performed to calculate the empirical HEPs (90th percentile confidence bounds of the HEPs). Two human failure events (HFEs) will have the same empirical confidence interval for the HEP if they have the same sample size and number of failure counts. Note that there were only data for three of the HFEs in the LOFW scenarios; all crews succeeded in HFE-1A (they managed the bleed and feed (B&F) in the base scenario), so there is no data for HFE-2A, which is conditional on the failure of 1A.

Although the qualitative data from the simulator could help to distinguish between HFEs with the same failure counts, the empirical Bayesian HEPs do not incorporate such information, since they only use the failure count as input. Consequently, we also produced an HFE difficulty ranking, which considered not only the empirical HEPs but also the qualitative observations. Thus, this ranking accounts for both quantitative (failure counts) and expert assessment of the observations made by subject matter experts on the HFEs and the related crew performances. In determining empirical difficulty, the expert assessment accounted for a number of performance issues, potential delays, crew situation awareness, etc. In the difficulty ranking for LOFW, HFE-2A is included in a comparison of the hypothetical performance conditions, given the failure of HFE-1A. HFEs 1A1 and 1B1 are not included in this ranking in this way, but are handled separately because their definitions overlap those of 1A and 2A, and 1B and 2B, respectively (ranking part vs. joint HFEs).

The ranking of the HFEs, based on the Bayesian HEPs and the difficulty ranking incorporating the qualitative evidence, is closely correlated but not identical. The latter is considered the reference ranking, and is referred to as the empirical difficulty ranking, because it is more informative. It accounts for all available empirical data and represents the consensus of all analysts who reviewed the empirical data. The difficulty ranking used as the X-axis in the figures in this chapter is the empirical difficulty ranking. In the rank comparisons, the empirical difficulty ranking is compared to the predicted ranking of the HFEs by each HRA team. The

[3] 14 crews participated, as in the SGTR runs, but, due to simulator problems, only 10 crews were analyzed.

predicted ranking used in the comparisons with the reference empirical ranking is always based on the HEPs obtained by the HRA team for the HFEs.

As can be seen in Table 2-1 and discussed further in Section 2.4.6, concerning the HRA method assessment criteria, the assessment of the quantitative predictive power of the methods weighted the prediction of the HEPs for the difficult HFEs, with observed failures and the more narrow confidence bounds being weighted the most heavily, followed by the comparison of the predicted HFE rankings to the empirical difficulty ranking. The comparison of the HEP predictions to the bounds was given a low weight, together with the quantitative differentiation between the most difficult and least difficult HFEs.

The quantitative comparisons supplement the qualitative comparisons and insights. Generally, the quantitative empirical data and comparisons give a very good starting point for assessing the qualitative predictions of the methods by prioritizing these qualitative findings and providing a measure of the significance of the predicted or observed performance issues. Thus, the overall evaluation of the HRA methods is based on both qualitative and quantitative insights. Due to limitations in the quantitative data, however, the qualitative comparison results and insights are weighted more strongly in the overall evaluation of the methods.

4.2 LOFW, Overall quantitative results from HRA method predictions

Figure 4-1 shows the range of predicted mean HEPs from all of the HRA methods in the study, and for all of the HFEs in the LOFW scenarios. On the X-axis, the HFEs are ordered by the empirical difficulty ranking. As stated in Section 3.2.2, they were ranked as follows:

1B > 2B > 1A > 2A (from difficult to easy)

For each HFE, boxes are drawn around a range, from which one maximum value and one minimum value are excluded. When outliers are excluded or censored in this way, it can be seen that the method-to-method variability for each HFE is one order of magnitude, or slightly more. This is less than in the SGTR scenarios, in which the variability for each HFE was on average two orders of magnitude (see Section 4.3). HFEs 1A1 and 1B1 are not included in this consideration, since these are joint HFEs and need special consideration.

Because the HFEs are ordered by difficulty, a comparison against difficulty ranking can be made (for methods in the aggregate). Compared to the difficulty ranking (horizontal axis), it can be seen that the HFEs in the complex scenario are predicted as more difficult than the base scenario HFEs, which corresponds with the data. However, one interesting feature of the results is that the HEPs predicted for HFE-2A tend to be larger than for HFE-1A, as do the HEPs for HFE-2B, which were predicted to be larger than those for 1B (see the “curve” in Figure 4-1, which takes the shape of a saw tooth). This is not consistent with the empirical data, in which the empirical difficulty ranking states that HFE-2 should be easier than HFE-1 for both the base and the complex cases. This is in part due to the treatment of dependency in some of the methods; a summary of the differences in this part of the analysis is provided in Section 6.2.

It should be noted that the predictions of individual HRA methods were not consistently placed within each box: in other words, the highest probabilities in the boxed ranges were not, as a rule, produced by the same methods (as is also the case for the lowest probabilities). In some cases, a given method would produce some of the highest HEPs for some HFEs (relative to other methods) while predicting some of the lowest for others.

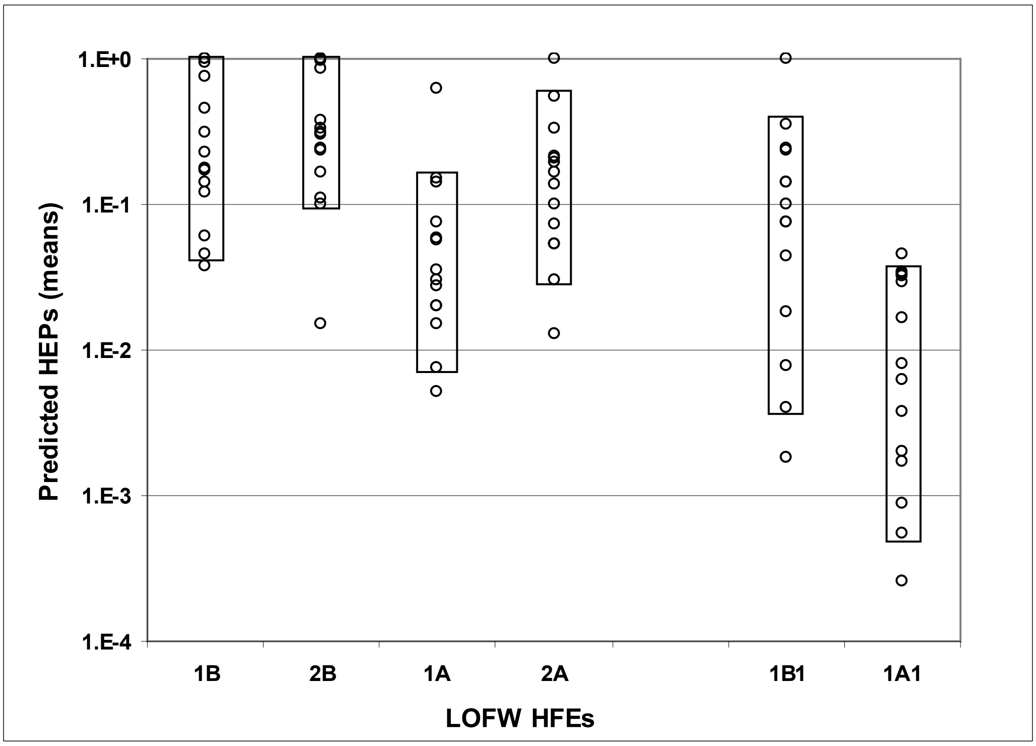

Figure 4-1 LOFW, range of predicted mean HEPs for the HRA methods, in decreasing order of difficulty

4.3 SGTR, Overall quantitative results from HRA method predictions

Figure 4-2 shows the range of predicted mean HEPs from all the HRA methods in the study, for all the HFEs. On the X-axis, the HFEs are ordered by the empirical difficulty ranking (see Section 3.2.1), which was as follows:

5B1 > 1B > 3B > 3A > [1A, 2A, 2B] > 5B2 > 4A (from difficult to easy)

For each HFE, boxes are drawn around a range, from which one maximum value and one minimum value are excluded. Some outlier estimates are explainable based on the analysts' interpretation of the information provided, or the assumptions they made to address missing or incomplete information. When outliers are excluded or censored in this way, it can be seen that the method-to-method variability for each HFE is two orders of magnitude or less. Furthermore, with the exception of the three outliers circled in the figure, the remaining outliers are relatively close to the boxed range. At least one of the extreme outliers (the three circled values) was caused by an incorrect assumption.

Because the HFEs were ordered by difficulty, a comparison of predictions against difficulty ranking (horizontal axis) can be made for the methods. As indicated in Figure 4-2, the predicted HEPs for the most difficult HFEs (i.e, the first four HFEs from left to right) at the aggregate level are consistent with the empirical evidence of decreasing difficulty. With the predicted HEPs for the less difficult HFEs (i.e., for the last five HFEs), the methods did not make clear distinctions in the HEPs. Also the empirical data did not make clear distinctions for these HFEs; no ranking was established between 1A, 2A, and 2B, but clear ranking was established for 5B2 and 4A. Thus, the HRA predictions mostly correlate with the empirical difficulty. However, it should be noted that the predictions of individual HRA methods were not consistently placed within each box. In other words, the highest probabilities in the boxed ranges were not, as a rule, produced by the same methods (as is also the case for the lowest probabilities). In some cases, a given method would produce some of the highest HEPs for some HFEs (relative to other methods) while predicting some of the lowest HEPs for others.

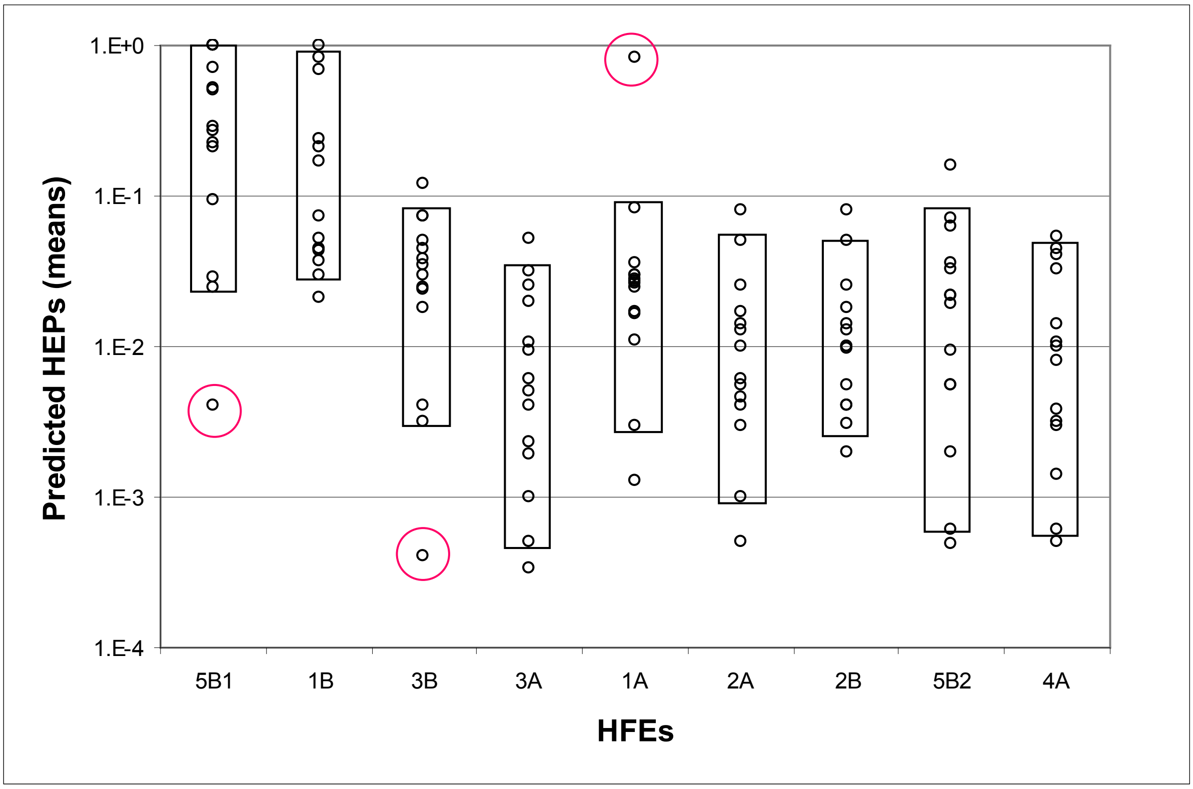

Figure 4-2 SGTR, range of predicted mean HEPs of the HRA methods

4.4 General discussion of the quantitative results of the HRA predictions

Despite of the care taken to provide a detailed description of the scenarios and to define the HFEs, the HEPs provided by the HRA teams show significant method-to-method variability.

- The variability was present for both the easy (i.e., those with expected low HEPs, such as HFE-4A in SGTR) and the difficult (i.e., those with expected high HEPs, such as HFE-1B and HFE-5B1 in SGTR) HFEs.

- The variability is not correlated across the HFEs in the sense that the same HRA analysis did not consistently produce the highest (or the lowest) HEP for the set of HFEs. In other words, none of the methods was systematically more conservative or optimistic than the other methods. In addition, the ranking of the HEPs was not consistent from method to method.

- Some method applications did not exhibit much variation among the HEPs; the range of HEPs for the set of HFEs was rather narrow, in some cases, less than an order of magnitude. One possible explanation is that this is a reflection of the discriminating power of the method. Methods with more degrees of freedom in choosing the HEPs can, in principle, provide a wider range of possible values. However, even if a method has many degrees of freedom (e.g., different numbers and levels of performance-shaping factors (PSFs)), this may not necessarily be exercised, and the focus of the analysis may be on a narrow set of PSFs. This has not been explored in detail at this time.

4.5 The empirical HEPs (Bayesian results) (LOFW part)

As noted, we performed a Bayesian update to obtain the reference or empirical HEPs because of the small sample size for each HFE. In contrast to the SGTR phases of the study, in which a "minimally-informed" prior was used [2], the LOFW data analysis used a non-informative prior distribution, the Jeffrey's prior ([6] and [7]). For this type of evidence, the Jeffrey's prior is a beta distribution with the parameters 0.5, 0.5.

Figure 4-3 shows the posterior distributions obtained in the Bayesian update for the LOFW data. When no failures are observed, the confidence interval for the posteriors is large (thin dashed lines). The interval spans about three orders of magnitude. In contrast, the large proportion of failures in the case of HFE-1B is strong evidence, and results in a narrow confidence interval (thick line on right of figure), a factor of 2 between the 5th and 95th percentile bounds. The hypothetical case of a single failure observed in 10 trials, shown for illustration only, is intermediate in terms of strength of evidence; correspondingly, the range of its confidence interval, one order of magnitude, falls between the previous cases.

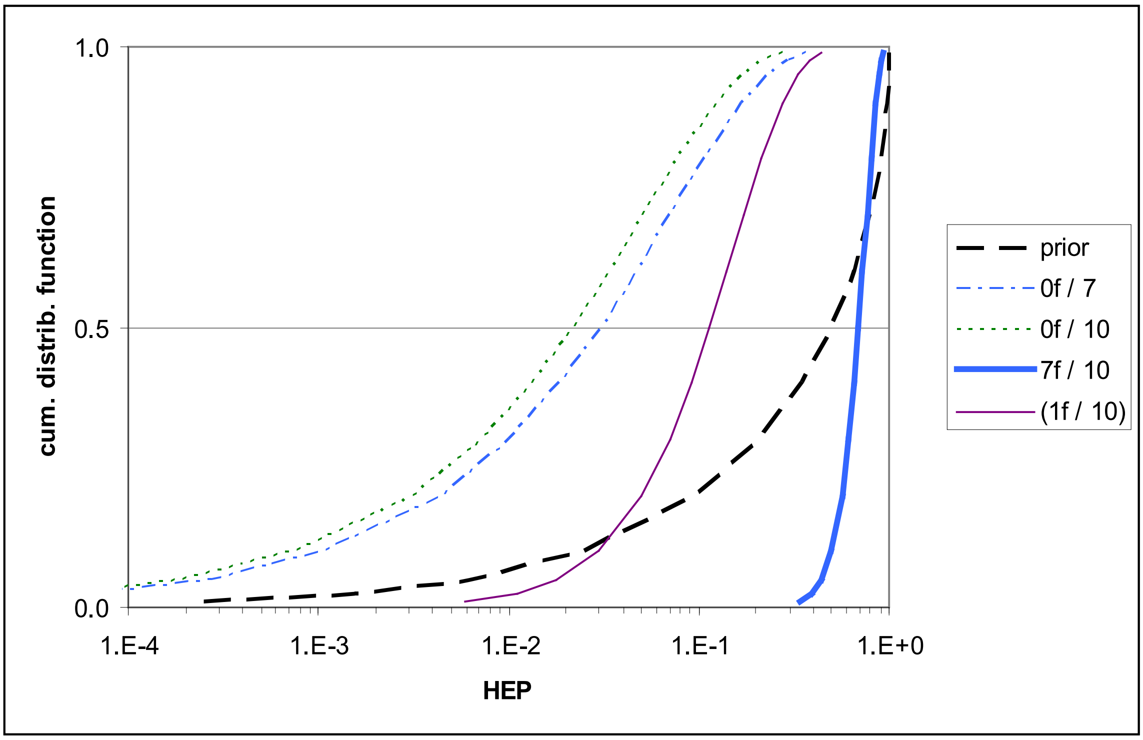

Figure 4-3 Bayesian posterior distributions resulting from update of Jeffrey's prior

Because of the large confidence intervals, a comparison with the mean HEP of the posterior distributions suggests in most cases an unwarranted accuracy. As a result, the empirical HEP mean value was not considered when comparing the HEPs predicted by the HRA teams with the empirical HEPs. The comparisons with the empirical HEPs focused instead on the relationship between the predicted mean HEP values and the 90% confidence bounds.

4.6 LOFW, Predicted HEPs vs. empirical HEPs (Bayesian results)

Figure 4-4 superimposes the 5th and 95th percentile Bayesian bounds for the empirical HEPs (dotted lines) on the plot of the predicted HEPs from all of the HRA methods in the study.

The breadth of these bounds is acute for the zero-failure cases (HFE-2B, 0/7; and HFE-1A, 0/10). This illustrates the limitations of quantitative comparisons based solely on failure counts. If the failure counts alone are used to define the empirical reference data without accounting for qualitative observations of the performances and identifying issues short of failure, the resulting reference data is “consistent” with predictions that are different by orders of magnitude. In addition, such reference data is practically unable to differentiate between HFEs with zero observed failures: for instance, for observations of 0 failures in 14 trials, the reference HEP has a mean of 0.03 with 90% confidence bounds from 1E-4 up to 1.3E-1 (for 0 failures in 10 trials, the reference is 0.046 and bounds from 2E-4 to 1.7E-1). On the other hand, for a given HFE, an observation of just one failure in 10 trials yields a mean failure probability of 0.136, with a much narrower confidence bounds; the lower and upper bounds span about one order of magnitude (e.g., from 1.8E-2 to 1.6E-1).

Thus, as mentioned earlier, the assessment of quantitative predictive power weighted the comparison of the ranking of the HFEs based on the predicted HEPs against the empirical difficulty ranking rather strongly. The HEPs predicted for the least difficult HFEs (particularly those with no observed failures and large confidence bounds) were considered primarily in terms of the obtained ranking rather than against the confidence bounds.

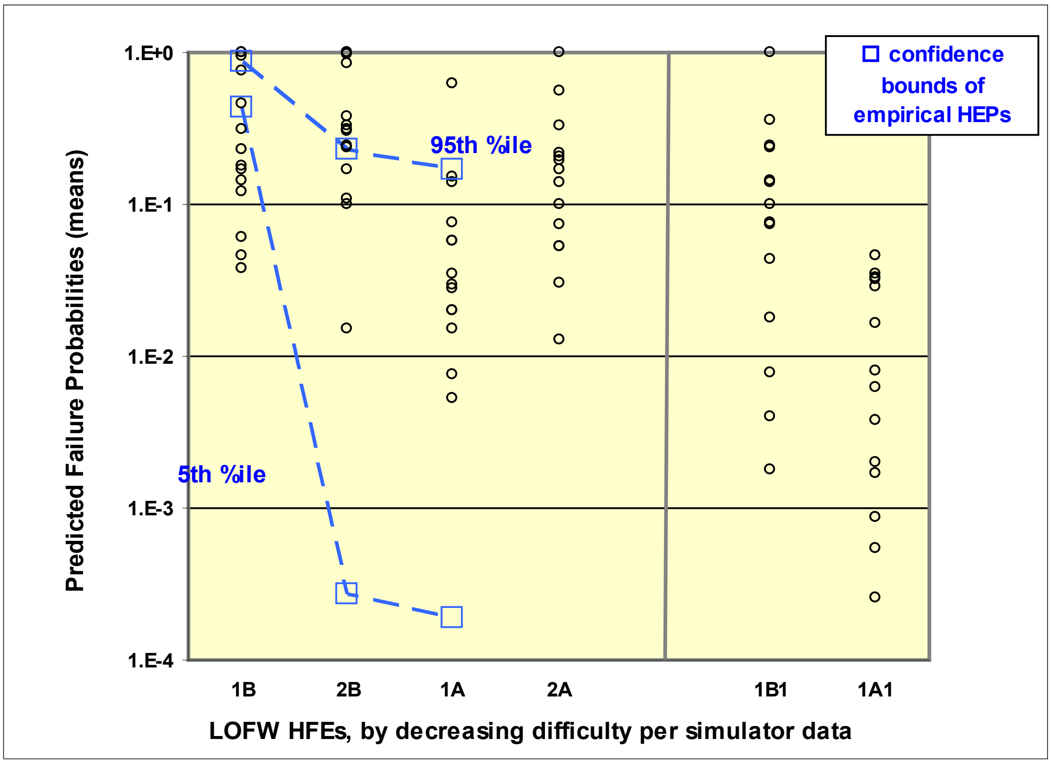

Figure 4-4 Bayesian confidence bounds of the LOFW empirical HEPs vs. all predicted HEPs

As can be seen in the plot, many methods underestimated the HEP for the most difficult HFE, 1B. This seems to be fairly systematic. Nevertheless, it should be noted that the majority of the predictions were above 0.1, consistent with a high expectation of failure.

At the same time, many methods overestimated the HEP for 2B. This is mainly due to the modeling of dependency (see Section 6). For HFE-1A, most of the methods had reasonable HEPs. There is no data for 2A (the conditional HFE), since all crews succeeded in 1A.

The joint HFEs, 1A1 and 1B1, were not used as extensively in the comparisons. The simulator observations, interpreted as failure counts for the joint HFEs, resulted in 0 failures in 10 observations for both joint HFEs. The corresponding confidence bounds for 1A1 and 1B1 would be the same as for HFE 1A, that is, broad and therefore limited in providing insights, except to suggest some pessimism (if the method produces a mean value above the 95th percentile value of 0.17 for these joint HFEs). Secondly, the empirical bounds for these HFEs do not discriminate between 1A1 and 1B1. On the other hand, the difficulty of 1B1 relative to 1A1, considering when B&F is implemented relative to the procedural criteria and qualitative considerations, is unambiguous.

4.7 SGTR, predicted HEPs vs. empirical HEPs (Bayesian results)

Figure 4-5 shows, as does Figure 4-2, all of the SGTR HEPs predicted by the HRA methods. It also shows the 5th and 95th percentile Bayesian bounds for the SGTR empirical HEPs (dotted lines).

The empirical Bayesian distributions have large bounds due to the small sample size (14 crews).

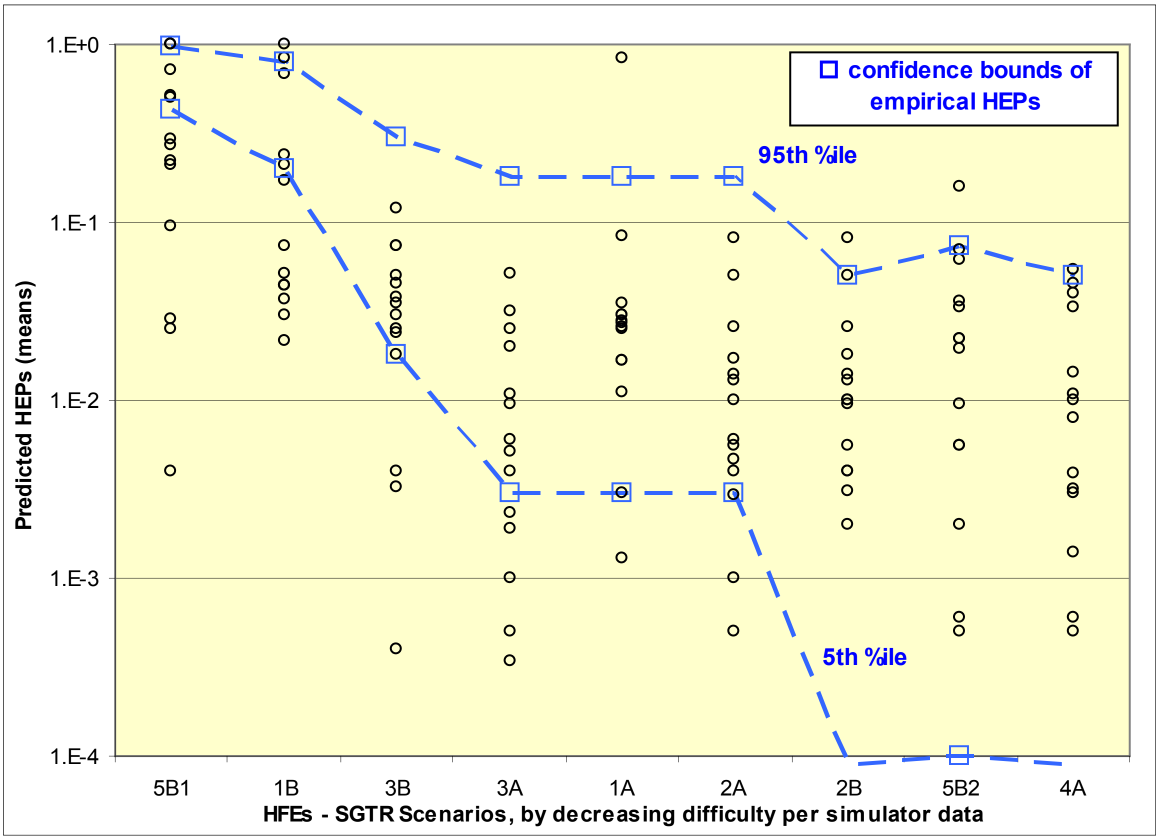

Figure 4-5 Bayesian confidence bounds of the SGTR empirical HEPs vs. all predicted HEPs

The plot shows that many methods underestimated the HEPs for the most difficult HFEs (5B1 and 1B). This seems to be fairly systematic. For the rest of the HFEs, nearly all predictions (mean values) fall within the Bayesian bounds; however, these bounds are very broad.

Figure 4-5 also shows the limitations of the empirical HEPs, in comparison to the predicted HEPs. The detailed qualitative analysis suggests that these empirical distributions, which are based solely on the failure counts in the number of runs, are not as informative as the difficulty ranking. As stated in Section 3.2.1, the difficulty ranking was as follows:

5B1 > 1B > 3B > 3A > [1A, 2A, 2B] > 5B2 > 4A (from difficult to easy)

1A, 2A, and 2B were all considered equally difficult. This is in contrast to the empirical HEPs, in which 2B, 5B2, and 4A were all zero-failure cases. In HFE-5B2 only seven crews participated, while all 14 crews participated in the other HFEs.

Overall, the qualitative findings (identification of issues, driving factors, etc.) are weighted more heavily in the evaluation than the quantitative performance.

5. HRA METHOD ASSESSMENTS

This chapter comprises findings on the individual human reliability analysis (HRA) methods as applied in this study, including their strengths and weaknesses, and provides recommendations on improving method guidance, development, and use. All assessments are based on an overall evaluation of each HRA team's analyses of both the steam generator tube rupture (SGTR) and the loss of feedwater (LOFW) scenarios. The findings, presented in terms of strengths and weaknesses (and sometimes in a neutral discussion of method features), are based on assessments of each team's analysis. Note that in many cases, the same feature might be a strength in some ways and a weakness in others. Thus, the discussions under the "Strengths" header may include mention of weaknesses, and vice versa.

While there is no empirically-based method to clearly separate method effects from analyst effects, by examining each method's guidance and the documentation provided by the analysts on their results, the assessors were usually able to identify the different methods' key strengths and weaknesses and judge those instances where the analysts may have gone beyond the methods or deviated from the guidance. A more empirically-based approach for separating method effects from analyst effects, which involves the use of multiple teams on the same method, is the topic of a follow-up study.

Please refer to HWR-844/NUREG/IA-0216, Volume 1 [1], for a one-page description of all of the methods applied in the study, and for the evaluation of the first two SGTR human failure events (HFEs). The remaining HFEs in the SGTR scenarios and the quantitative comparison are addressed in HWR-915/NUREG/IA-0216, Volume 2 [2]. HWR-951/NUREG/IA-0216, Volume 3 [3] documents the two variants of LOFW scenarios. A summary of the overall insights related to each method is presented below. Specific details related to the evaluation of these methods are provided in HWR-844/NUREG/IA-0216, Volume 1 [1], HWR-915/NUREG/IA-0216, Volume 2 [2], and HWR-951/NUREG/IA-0216, Volume 3 [3].

5.1 Overall Assessment of ASEP (UNAM)

The Accident Sequence Evaluation Program Human Reliability Analysis Procedure (ASEP) is, as described in NUREG/CR-4772 [8], intended to be a less resource-intensive version of the THERP (Technique for Human Error Rate Prediction) method described in NUREG/CR-1278 (THERP Handbook) [9]. ASEP also extends THERP in several ways, particularly with respect to the treatment of pre-initiators.

5.1.1 Strengths

Simplicity

One strength of ASEP is ease of use, given its simplicity: its human performance model was simplified by separating diagnosis from post-diagnosis actions, it estimates the diagnosis human error probability (HEP) using only the diagnosis time reliability curve with performance-shaping factor (PSF) adjustment, and it focuses on the major procedural steps. On the one hand, these simplifications make the method easy to use; on the other hand, they contribute to the weaknesses discussed below. The developer has justified this simple analysis by claiming that conservative HEPs will generally be obtained. However, apparent optimism due to the method's weaknesses was seen in some cases (e.g., HFEs 1B and 5B1 in the SGTR scenarios; see discussion below). The implication is that the trade-offs between simplicity and thorough analysis need to be weighed before the method is applied.

Traceability

Another strength of the method is its traceability. The estimation of allowable diagnosis time and allowable post-diagnosis time, the derivation of the HEP within the method, and the identification of what is important to performance given the factors considered are generally traceable, and the method for weighting various factors in calculating the final HEP can be determined. However, determining how the analysts might bias or alter the rating or level of the factors considered in applying the method, based on other identified information that is not covered by the method, could be difficult if the analysts do not document their decision process well.

5.1.2 Weaknesses

Insufficient guidance on when to include or exclude modeling of the diagnosis phase in performing the analysis

It is interesting to note that although the analyses for the LOFW and SGTR scenarios were performed by the same HRA team, the predictive power of the analyses for the LOFW scenarios is considered to be better than that for the SGTR scenarios. Although it could be argued that there might be a scenario and/or learning effect (HRAs for LOFW scenarios were conducted after the HRA team saw the study results for the SGTR scenarios), the relatively poorer predictive power for the SGTR scenarios appeared to be caused by the HRA team's assumption of a successful diagnosis once the crews entered symptom-based procedures. Comparison across the two categories of scenarios can shed light on the consequences of such an assumption.

By dividing the total time available for coping with an abnormal event into two independent parts, allowable diagnosis time and allowable post-diagnosis time, ASEP provides an option to explicitly include and quantify diagnosis. However, based on the results from this study, either the guidance on when to include or exclude diagnosis is insufficient, or the option needs to be taken out of the methodology. As shown in the empirical data, excluding the diagnosis from the SGTR scenarios was inappropriate, as the analysts then failed to recognize that operators had to assess the situations and/or make new response plans while the scenarios progressed. The decision to skip the diagnosis part of the crew response apparently precluded the opportunity to address operators' cognitive activities, examine any difficult conditions that the operators would be facing, or identify important factors influencing performance. Consequently, the HRA team only obtained a partial picture of the dynamic nature of the accident scenarios, and failed to consider the most relevant factors: for instance, by focusing mainly on crews working through the procedures, the HRA team did not really register the strong difference between the conditions for HFEs 5B1 and 5B2 in the SGTR scenarios. Additionally, except for the easiest HFEs, 5B2 and 4A, where there seems to be a good agreement between the predicted drivers and those identified from the crew data, the predicted negative drivers rarely matched those identified from the crew data in the SGTR scenarios. In contrast, the HRA team identified many of the important drivers that would influence performance in the LOFW scenarios. Although this seems to be partly due to the HRA team's experience (from participating in the study) and an effort that went beyond ASEP guidance, addressing the diagnosis in terms of the ASEP diagnosis curve did lead the team to a good understanding of what would be going on in the scenarios, and to consider the potential impact of available time on the diagnosis.

Quantitatively, the final HEP in ASEP is the sum of the diagnosis and execution HEPs. Under the assumption of a successful diagnosis, the final HEP is only determined by the probability of making an error in executing post-diagnosis actions, and thus can be optimistically estimated. Although it is difficult to estimate the true HEPs, given the limited data, the optimism in the HEPs of the SGTR scenarios is well illustrated in the HEP pattern. For the most difficult HFEs, 5B1 and 1B, the HEPs fall below the lower Bayesian uncertainty bound. In particular, the HEP for HFE-5B1 is 0.025, which shows a large disconnection from the fact that all crews failed that HFE. In addition, the HEPs for HFEs 3B (0.025) and 3A (0.004) appear to be smaller than the actual crew failure rates (2 out of 14 crews failed in HFE-3B, and 1 out of 14 crews failed in HFE 3A).

Limited guidance for estimating time requirements

The above trend towards optimism is interesting in that ASEP claims to provide generally conservative HEP values. However, where diagnosis was addressed, the HEPs for the LOFW scenarios do seem to suggest conservatism. In particular, the HEP for HFE 2B (0.312) not only falls above the upper Bayesian uncertainty bound, it also seems to be more conservative than appropriate, given the zero-failure rate. In this case, the main contributor to the conservatism seems to be the conservative assumption about the allowable diagnosis time (i.e., the time needed to determine whether bleed and feed (B&F) is needed), in conjunction with the use of the ASEP diagnosis curve. It appears that more guidance on estimating time requirements and appropriately considering factors that could influence time requirements (e.g., concurrent activities, demands of working through procedures) would strengthen the method.

Inadequate set of factors covered

It is likely that the HRA team's analysis of the SGTR scenarios would have been improved if the HRA team had explicitly addressed diagnosis. However, even if diagnosis is explicitly included, the method is still unable to guide analysts to examine an adequate set of factors that would influence crew behavior; for instance, the guidance to address diagnosis/cognitive tasks is minimal, and the method relies heavily on its diagnosis curve, with adjustments for a few PSFs. Even with the guidance provided, it seems that the analysts would already have to have an idea of what they are looking for (i.e., they would need a good background in information needs for HRA in the context of probabilistic risk analysis (PRA)) in order to perform an appropriate analysis. In the LOFW scenarios, as mentioned above, the better predictive power seemed to be partly due to the HRA team's experience and their effort, which went beyond ASEP guidance. Furthermore, in the SGTR scenarios, it appears that it became more difficult for the HRA team to predict performance drivers as the scenarios became more complicated. This suggests that improved methodology and more guidance are needed to help analysts address critical tasks at a more cognitive level, and it is necessary to include additional performance drivers to address complicated scenarios.

Inadequate guidance for distinguishing step-by-step vs. dynamic post-diagnosis tasks[4]

When addressing post-diagnosis actions, decisions are made by analysts regarding stress levels and whether an action was step-by-step or dynamic (i.e., execution complexity). The guidance on those decisions is limited, which may explain why the team's decisions on those

[4] ASEP defines step-by-step tasks and dynamics tasks as follows:
Step-by-step task: A routine, procedurally guided set of steps performed one step at a time, without any requirement to divide one's attention between the task in question and other tasks. With high levels of skill and practice, a step-by-step task may be performed reliably without recourse to written procedures.
Dynamic task: One that requires a higher degree of interaction between the people and the equipment in a system than is required by routine, procedurally guided tasks.

factors did not appear to correspond well to the factors and conditions observed from crew performance data.

Another limitation of ASEP is its focus on procedural steps at a high level (e.g., identification of the initiating event and entry into the appropriate emergency operating procedure (EOP)), rather than on the diagnosis and cognitive activities involved in following and responding to the steps in the EOPs: that is, lower-level cognitive activities, such as interpreting the plant status in the context of the step-by-step procedures and associated time-limiting conditions, need more attention than is given in evaluating post-diagnosis tasks. Consequently, HRA predictions are likely to be limited to the crew's interaction with the main procedural steps and to lead to optimistic HRA results by ignoring the difficulties that operators would face at the sub-step level.

Limited insight for error reduction

In general, ASEP relies heavily on its diagnosis curve and a few PSF adjustments to address diagnosis. This approach limits the method's ability to discover cognitive mechanisms that would lead to human failures, thus limiting its ability to offer insights into error reduction.

5.2 Overall Assessment of ASEP/THERP (NRC)

Although both ASEP [8] and THERP [9] were used in the analysis, the HRA team basically followed the guidance in ASEP. THERP is allowed in ASEP, where appropriate, to support quantification. In most cases, the HRA team determined that there was sufficient information for a task analysis; thus, THERP was used to estimate HEPs of post-diagnosis actions, per Item 2 in ASEP Table 8.5. In some cases where there was insufficient information, rules in ASEP Table 8.5 were used. The strengths and weaknesses of ASEP and some aspects of THERP are discussed below.

5.2.1 Strengths

Simplicity

One strength of ASEP is ease of use, given its simplicity: its human performance model was simplified by separating diagnosis from post-diagnosis actions, it estimates the diagnosis HEP using only the diagnosis time reliability curve and a few PSF adjustments, and it focuses only on the major procedural steps without examining potential complexities in the sub-steps, given the conditions of the scenario. On the one hand, these simplifications make the method easy to use; on the other hand, they seem to contribute to the weaknesses discussed below. The developer has justified this simple analysis by claiming that conservative HEPs will generally be obtained; however, apparent optimism due to the method's weaknesses was seen in the study (e.g., HFE-1B in the SGTR scenarios; see discussion below). The implication is that the trade-offs between simplicity and thorough analysis need to be weighed before the method is applied.

Traceability

As noted in Section 5.1.1, a strength of the method is the traceability of the quantification. The estimation of allowable diagnosis time and allowable post-diagnosis time, the derivation of the HEP within the method, and the identification of what is important to performance given the factors considered are generally traceable, and the method for weighting various factors in calculating the final HEP can be determined. However, determining how the analysts might bias or alter the rating or level of the factors considered in applying the method, based on other information identified that is not covered by the method, could be difficult if the analysts do not document their decision process well.

5.2.2 Weaknesses

The disconnection between the ASEP/THERP HRA team's predictions and the crew data seemed to stem largely from the method's insufficient guidance, which failed to lead the team to fully understand the nature of the scenarios that the crews would face. The insufficient guidance is manifested in the following aspects.

Insufficient guidance on when to include or exclude modeling of the diagnosis phase in performing the analysis

By dividing the total time available for coping with an abnormal event into two independent parts, allowable diagnosis time and allowable post-diagnosis time (i.e., related to response execution time), ASEP provides an option to explicitly include and quantify diagnosis. However, insufficient guidance is provided as to when to include or exclude diagnosis. The HRA team assumed that no diagnosis was required once the crews entered symptom-based procedures in all SGTR and LOFW scenarios. As shown in the study, such an assumption was inappropriate, as crews had to assess the situations and/or make new response plans while the scenarios progressed. Failure to address diagnosis seemed to be a major contributing factor to the team's predictions, which were inconsistent with the empirical data in terms of performance drivers and operational stories. The decision to skip the diagnosis part of crew response may have precluded the opportunity to address operators' cognitive activities, examine the difficult conditions operators would be facing, and identify some important factors influencing performance. As a result, the HRA team only obtained a partial picture of the dynamic nature of the accident scenarios, which is well illustrated by the fact that, without counting for diagnosis, the HRA team concluded that the analysis for HFE-2B should be the same as that for HFE-2A in the LOFW scenarios, while the empirical data shows they were not similar. By equating the analysis for HFE-2B with that for HFE-2A, the HRA team failed to consider the difficulties involved in identifying the misleading steam generator (SG) water level indications and the degraded performance of one running condensate pump in HFE-2B.

Quantitatively, the final HEP in ASEP is the sum of the diagnosis and execution HEPs. Since skipping diagnosis implies a zero probability of failure for that part of the scenario, it can lead to an optimistically estimated final HEP. For example, the HEP (0.02) for HFE-1B in the SGTR scenarios appears to be optimistic, compared to the 50% crew failure rate.

Inadequate set of factors covered

It could be argued that the HRA team's analysis might have been improved if the team had explicitly addressed diagnosis. However, even if diagnosis is explicitly included, the method is still unable to guide analysts to examine an adequate set of factors that could influence crew behavior. For example, the guidance to evaluate diagnosis/cognitive activities is minimal, and the method relies heavily on its diagnosis time reliability curve, with adjustments for only a few PSFs.

Inadequate guidance for distinguishing step-by-step vs. dynamic post-diagnosis tasks

When addressing post-diagnosis actions, whether using ASEP or THERP (as mentioned above, THERP was used in most cases in quantification with respect to post-diagnosis actions, per ASEP instruction), decisions needed to be made regarding stress levels and whether an action was step-by-step or dynamic (i.e., execution complexity). The guidance on those decisions appears to be limited for some situations, which may explain why the team's decisions on those factors did not correspond well with the factors and conditions observed from crew performance data.

Insufficient guidance to examine lower-level cognitive activities

Another limitation of ASEP is its focus on procedural steps at a high level (e.g., identification of the initiating event and entry into the appropriate EOP), rather than the diagnosis and cognitive activities involved in following and responding to the steps in the EOPs. That is, lower-level cognitive activities, such as interpreting the plant status in the context of the step-by-step procedures and associated time-limiting conditions, need more attention than is given in this analysis when evaluating post-diagnosis tasks. As a consequence, HRA predictions are likely to be limited to the crew's interaction with the main procedural steps, and to lead to optimistic HRA results by ignoring the difficulties that the operators would face at the sub-step level.

Limited insight for error reduction

In general, ASEP relies heavily on its diagnosis curve with a few PSF adjustments to address diagnosis. This approach limits the method's ability to discover cognitive mechanisms that would lead to human failures, thus limiting its ability to offer insights into error reduction.

5.3 Overall assessment of ATHEANA (NRC)

For many of the HFEs, the analyses performed with ATHEANA (A Technique for Human Event Analysis, NUREG-1624, Rev. 1 [10], NUREG-1880 [11]) identified many of the important drivers (key, driving PSFs). Although PSFs are not at the center of ATHEANA analyses, the failure scenarios predicted in these analyses could be interpreted in terms of the associated PSFs and compared against the drivers derived from the empirical observations. The operational expressions in the ATHEANA analyses also often encompassed the failures actually found in crew performance, although several failure paths were identified that were not observed among the crews in the study, which is not necessarily unreasonable, given the limited sample size. This particular benchmark exercise did not fully test a major feature of performing an ATHEANA analysis, which is the search for a range of error-forcing contexts (EFCs) and unsafe acts (UAs) (i.e., deviation scenarios) that are consistent with the PRA definition of the HFE. It could be argued that much of the value of performing an ATHEANA analysis has not been tested by this exercise, because the EFCs and UAs were essentially predefined. However, it was still possible for the scenarios to evolve in somewhat different ways (particularly from crew actions, timing of actions, etc.), so that the ATHEANA analysis could, in principle, have identified some deviations that would have affected performance on HFEs; and, in fact, some of the operational stories did reflect such variations. In addition, even within the constraint of predefined HFEs, the method's approach of searching for error modes or mechanisms has been shown to provide some valid predictions, particularly when the EFC is strong.

This search is one way in which the ATHEANA process can handle some aspects of crew-to-crew variability, which was an important aspect of the empirical data. However, to do this reliably would require considerably more knowledge about specific crew characteristics than was available prior to in this exercise. The ATHEANA approach of providing a framework for evaluating the impact of context on HEPs by considering potential failure modes is most valuable when there is an identifiable EFC, or EFCs. Compared to other methods, this is less of an advantage when the tasks are straightforward, the EFC is weak, and success is expected.

The ATHEANA team would also typically include operations experts from the plant being analyzed, but such experts were not available to join the analysis team. The ATHEANA team used information provided by Halden about the crews and the procedures they used; enlisted operational experience on the team by including former trainers from other plants, inspectors, and HRA experts; obtained additional documents (such as additional procedures) from U.S. plant operations, as needed; and developed insights to adjust the U.S. nuclear power plant (NPP) operating experience for the non-U.S. operating crews. Regarding the latter point, it was noted by the ATHEANA team that the crews in the Halden study tended to move more quickly through the procedure steps, and there was more variability in performance than would be expected for U.S. crews; however, it is impossible to determine the veracity of this observation without formally comparing the U.S. and non-U.S. crews. Additional insights into Halden crew performance were gained from the ATHEANA team's experience completing the analysis for the SGTR scenarios and subsequent comparison of their analysis to the actual crew performance. While the ATHEANA analysis did not in most cases produce a good match to the SGTR crew performance in terms of quantitative predictive power, the LOFW analysis was calibrated to Halden crew performance based on feedback in the SGTR round of the study, and the subsequent quantification of the LOFW scenario proved a very close match to the performance data.

5.3.1 Strengths

Qualitative predictive power

The ATHEANA method is not, strictly speaking, a PSF-based approach. Analysts consider a full but informal complement of PSFs in the analysis, and the weighting of the drivers varies, based on the context. The EFC—a somewhat unique element of ATHEANA—is generally more important than the individual PSFs or drivers. The ATHEANA analysis provided a number of possible outcomes for each HFE, which successfully identified the sources of failure. For most of the more challenging HFEs, the ATHEANA team's qualitative discussion matched the observations well. For the less challenging HFEs, particularly in the SGTR scenario, the qualitative analysis in terms of operational expressions was generally mixed. The ATHEANA analysis did not identify the different strategies observed when the crews were completing the actions; however, the analysts' focus was on identifying ways in which the time-based success criteria for the HFEs might be exceeded. Furthermore, it is not clear that it would have been easy to predict the different approaches taken without having considerably more information on the crews and their training. The analysis team did identify some of the observed crew behaviors that could lead to a delay in completing the tasks in the SGTR scenario, and also identified most issues in the LOFW scenario, once it was calibrated to the crews. The success of the method clearly hinged on the amount of information available to the analysis team.

Traceability of the analysis

The traceability of the analysis was good, due to the plentiful supporting documentation. The question remains, however, as to whether all ATHEANA analyses are equally traceable, and, for that matter, whether another analysis by a different team would replicate the present findings. The extensive supporting documentation provided by the present team is essential to understanding ATHEANA. More so than most other HRA methods, ATHEANA relies less on templates and forms and more on the skill of the analysts in documenting their decision process. Without such documentation, the traceability of ATHEANA would be negligible. This was found to be the case during the assessment of the ATHEANA analysis for the LOFW scenario, when a key piece of documentation had inadvertently been excluded from the ATHEANA analysis. Without this piece of documentation, it was initially very difficult for the assessor to trace through the analysis. Without thorough and extensive supporting documentation, an ATHEANA analysis would likely be neither navigable nor traceable.

Insights for error reduction

A strength of ATHEANA is the search process, which identifies EFCs. In the present study, it was not possible for the ATHEANA team to complete the search process, which would have required the team to discuss the scenario with crews, or with operations experts familiar with the specific crews and control room featured in the analysis. Such a process might have resulted in less reliance on the skill of the individual ATHEANA team to identify the relevant EFCs and more reliance on the quality of the process presented in ATHEANA for identifying errors.

The ATHEANA search strategy is useful in identifying ways in which errors occur, and it lends itself to use for error reduction. The search strategy may be generous in terms of identifying more failure paths than would be expected in reality. The large number of failure paths is, however, addressed during quantification, during which the most risk-significant failure paths are clearly identified. As an HRA method, ATHEANA uniquely provides a comprehensive search process that is invaluable in predicting failure paths, although at present the identification of failure paths is to some extent a byproduct of the search for EFCs, UAs, and deviation scenarios (also see weaknesses below). The method does not provide equally extensive discussion on applying this process to error reduction.

5.3.2 Weaknesses

Potentially poor consistency for quantitative predictive power

The traceability to the quantification is not clear in this application. The quantification is based on expert judgment. While there is a discussion of the factors that can influence particular failures, it was not clear how these were taken into account by the contributing experts. This is most obvious when the error-forcing conditions are not strong, and the failure modes are slips and lapses. In these cases, the ATHEANA method seems to provide little advantage over other methods.

The quantification, relying as it does on expert elicitation, needs to be much more clearly documented. Even though the driving factors were identified, it was not possible to determine their relative weighting or importance. The quantification could be very difficult to reproduce in that a different set of experts might provide very different assessments. This concern may not be unique to ATHEANA, but the reliance on expert elicitation poses a particular concern in terms of the replicability of the quantitative analysis.

Note that a member of the analysis team was a previous trainer, who provided quantification estimates in the form of “We saw this happen in crews maybe 2 out of the 1,000 times we ran this type of scenario.” These insights provided a degree of informal operational data that matched the actual observed crew performance very closely. The quality of the quantification in ATHEANA seems highly dependent on the quality of the expert panel. The panel used for the analysis featured trainers, former reactor operators, and human reliability experts with considerable operations experience that fed directly into the quantification. It is not clear that the ATHEANA method would replicate such findings with a different or less qualified panel of experts.

Improved guidance is needed

The ATHEANA analysis followed a logical method of identifying failure paths and quantifying the likelihood of those paths. Still, this process seemed to deviate from the process outlined in the accepted standard documentation for ATHEANA, namely NUREG-1624, Rev. 1 or NUREG-1880. The guidance for ATHEANA is indeed extensive, but it is also diffuse thinly spread, particularly in the areas used in this exercise, namely the understanding of the use of the procedures, the identification of failure modes and failure paths, and the influence of PSFs. Overall, the guidance appears to be too diffuse and complex, and relies heavily on the analysts to collect and use the information in an appropriate manner. A more structured approach would be beneficial, as would specific, systematic guidance for searching for failure paths, especially for less expert analysts.

5.4 Overall Assessment of CBDT+THERP (EPRI)

The method referred to as “EPRI/CBDT” in this study is the Electric Power Research Institute (EPRI) HRA approach, which is applied using the EPRI HRA Calculator [12] and refers to a combination of CBDT (Cause-Based Decision Tree [13]) and THERP, supplemented by human cognitive reliability/operator reliability experiments (HCR/ORE) for time-critical actions [13]. The EPRI HRA Calculator includes quantification methods that were not used to quantify the HFEs. As applied in this experiment, several potential strengths and weaknesses were identified in the methodology.

5.4.1 Strengths

Identified some factors important to performance

The method did in many cases identify factors that were important contributors to the crews’ performances. As discussed below, while the method did not appear to cover an adequate range of important factors, it did appear to reliably identify some of the driving factors.

Traceability and structured approach

Another strength of the method was its traceability (at least in one aspect of traceability). The derivation of the HEPs within the method and the identification of which factors contributed to the HEPs are generally traceable. How the various factors are weighted in determining the final HEP can be determined by examining the contributions of various factors from the decision trees. However, the ability to trace the basis for the judgments regarding the branch points in the trees will rely on the analysts' documentation. Similarly, if analysts attempt to incorporate factors or expected effects not directly addressed by the decision trees, good documentation of the rationale will be necessary to allow traceability. Nevertheless, a structured, generally traceable approach for obtaining the HEPs, given a set of assumptions, is a strength of the method.

5.4.2 Weaknesses

Inadequate treatment of diagnosis in some cases

For the SGTR scenarios, the CBDT method was generally used to quantify the diagnosis portion of the HFE, and the THERP method was used to quantify the execution portion; however, for several HFEs in the SGTR scenarios, the analysts assumed that, based on the identified conditions, there would not be any additional diagnoses for some of the HFEs. They argued that after the initial diagnosis of the SGTR event and the presence of straightforward cues for the actions, the crews would simply follow the procedures, and limited diagnosis would be involved. For these HFEs, the HEP was quantified solely on the assessment of response execution using THERP. This decision, which in general appears to have been a modeler's choice and not a function of the software tool (EPRI HRA Calculator) or the associated methods (specifically the CDBT and HCR/ORE approach described in EPRI TR-100259), meant that for some HFEs the analysts did not investigate potential negative diagnosis factors that could influence performances based on the CBDT decision trees. In some cases, disregarding the crews' cognitive activities and related failure mechanisms while they were following procedures apparently led to a failure to identify some important negative drivers, which in turn led to apparent underestimations of HEPs.

Use of HCR/ORE time reliability model led to apparent overly conservative HEPs

Based on the results from their analysis of the SGTR scenarios (in which the HRA team arbitrarily limited themselves to applying CBDT), the analysts chose to use both the CBDT quantification approach and the time-reliability correlation from HCR/ORE to quantify diagnosis for the LOFW scenarios (summing of the results from each model). This was done because the CBDT is known to produce relatively low HEPs when time pressure is a relevant driver, and limiting the approach to CBDT caused the team to miss important information in the SGTR scenarios; however, this solution led to overly conservative HEPs for the LOFW HFEs, as suggested by the crew data. That is, the use of the HCR/ORE model caused the time available to become a driving factor in some cases where it did not have a detectable impact on the crew data.

Modeling options

The implication of the above findings is that the option to not explicitly address diagnosis in applying the CBDT/HRA Calculator may lead analysts to miss important information in some situations. Additional guidance is needed to determine when (if ever) this should be a viable option. Similarly, the findings suggest that the HCR/ORE time/reliability correlation (TRC) may produce overly conservative HEPs in some cases; thus, blanket application of this approach does not appear to be warranted. Summing the results from the HCR/ORE model and the CBDT approach to obtain an estimate of the HEP for diagnosis may also create problems with the relative accuracy of the HEP, although the calculator also allows the user to select the maximum of the two HEPs. While summing the HEPs allows both the potential impact of time and various PSFs in CBDT to be included, it is not clear that doing so is consistent with the intent of the methods, and it may result in unrealistic HEPs.

Inadequate set of factors covered

Another important potential CBDT limitation or weakness identified in the study is that the factors addressed or covered by the CBDT model (and, more generally, the HRA Calculator) may not always be adequate to identify important driving factors that influence crew performance (i.e., the model did not always lead the crews to address significant aspects of the scenario). Similarly, even if analysts identify operational conditions in the scenario that could be a problem, the model may not provide a direct means to incorporate this information. This was evinced to some extent by the fact that a good operational story developed by the analysts and consistent with the data did not always translate into "appropriate" HEPs (at least as suggested by the data). It also appears that in some cases, the approach may identify some PSFs as important contributors that inappropriately lead to higher HEPs: that is, the PSFs are judged to be at a level that should lead to increased HEPs, when in fact that they have no impact on crew performance. This effect may be due to (1) some PSFs not being relevant factors, given the conditions; (2) the analysts misjudging the level of the factor that is present, given the context; or (3) insufficient data available to detect the effect of the factor (not enough crews were run through the scenarios). Taken together, these potential issues with the methodology may have contributed to the lack of differentiation that was seen between some of the HEPs where significant differences in error rates were obtained in the crew data. However, it should be noted that while there was little differentiation between HEPs in some cases, there was often good general correspondence between the difficulty ranking of the HFEs and the corresponding HEPs: that is, while the analysis sometimes failed to reflect the degree of the differences in the difficulty of some HFEs and did not always detect when error rates would be very high or very low, at least in this study the application of the method appeared to show some sensitivity to the relative difficulty of the HFEs. Whether this is an inherent characteristic of the method or somewhat of a coincidence could not be determined by this study.

5.5 Overall Assessment of CESA-Q

It should be noted that the Commission Errors Search and Assessment – Quantification (CESA-Q [14]) method was developed for errors of commission (EOCs), and was being adjusted for use in this application. Thus, the guidance had not been developed to the level it might be in the future. In addition, since the EOC-focused method was intended to be used alongside an error of omission (EOO) approach, the method itself did not explicitly address how to treat execution issues or the execution part of the HFEs. ASEP or THERP would generally be used, but were not explicitly used in this application.

5.5.1 Strengths

PSFs covered

The method appears to provide a reasonable set of situational factors to represent important factors in the scenario being analyzed (at least in terms of decision making errors), but some additional ones may be needed for most scenarios. Although the CESA-Q method did not explicitly address some of the PSFs used to represent the driving factors for the crew data (at least in terms of using the same terminology), the factors addressed by the method appear to get at many of the same general issues. In other words, even though different terminology might be used, many of the important factors identified in the crew data are still addressed when determining HEPs, and they could be represented in the table of driving factors in the comparisons with the actual data in the study. Although not initially included, a factor to address the impact of a shortage of time to complete the necessary actions (as opposed to the impact of time pressure on crew judgment) was added after recognizing the need based on the analysis of the "pilot" data. This strengthened the method's coverage of relevant factors. A factor that did not seem to be explicitly addressed in CESA-Q, but that was used in assessing the crew data, was Team Dynamics, but the HRA teams were not given sufficient information to address this factor (i.e., they were not able to observe or interview crews), and this is a difficult factor to assess for PRA/HRA purposes anyway. As applied, the method also did not address or weight the impact of the execution difficulties seen in some of the HFEs.

The CESA-Q analysis often identified the main negative drivers reflected in the crew data. In some cases they identified PSFs as negative drivers that either did not have an impact or whose impact could not be determined, but, for the most part, they were fairly consistent with those identified in the crew data. In a couple of cases, minor negative factors were predicted that matched the negative factors for the crew data, but the basis for the effects of the factors differed. The match between the positive factors identified by the method and in the crew data was usually reasonably consistent. However, the method application in this study benefitted from a good task/qualitative analysis that appeared to be more of a function of a knowledgeable analysis team, rather than of the method.

Traceability (to some extent)

The derivation of the HEPs within the method and the factors important to performance are generally traceable, but the weighting of the various situational factors in determining the final HEP is not yet traceable: that is, a strength of the method is that the analysts' judgments in applying the method and obtaining the HEPs are traceable in the sense that the analysts' decisions on the ratings of the situational factors in terms of whether they are error- or success-forcing is fed directly into the quantification process. A potential shortcoming from an "understanding" perspective lies in how these decisions are weighted relative to one another in obtaining the final HEPs. The underlying basis for the final HEPs (underlying data) is not explicit either. Apparently a systematic quantification was done to obtain the final HEPs in the underlying data from which the method HEPs are obtained, but the catalogued analyses need a validation of some sort.

Insights for error reduction

In conjunction with a good task analysis, the PSFs and situational factors included in the CESA-Q method should allow insights into improving safety and reducing errors. The method examines aspects that, when identified as problematic, could be improved to facilitate error reduction; however, this will depend heavily on the strength of the judgments made about the different potential situational factors and the underlying qualitative analysis, for which additional guidance is needed. It is likely that some additions to the situational factors or to the scope of the current ones will improve insights into error reduction.

5.5.2 Weaknesses

Guidance for scaling PSFs

The approach for quantification and the factors addressed are somewhat untraditional in the sense that the assessments/questions asked would probably not be considered classic HRA (e.g., they examine aspects like verification hint, verification means, and whether there is an adverse exception). Without additional guidance on how to make the relevant decisions and which factors to consider, it is not clear that the method would produce consistent results. Some of the results from the study (e.g., HRA team weighting some factors identified as contributing to HFEs) indicated that additional guidance for scaling the PSFs or situational factors was needed. In the SGTR scenarios, for example, the contributing factors were not always weighted negatively enough, and the weighting for mild EFC cases was difficult, although some of these effects may have been caused by the lack of explicit treatment for execution issues (see below).

Guidance for performing qualitative analysis

The underlying qualitative analysis performed for this study (e.g., developing/understanding the operational story, identifying key decision points, evaluating the EFC, and ultimately identifying the driving factors) was generally good, but it is not clear whether the basis for the assessment of the situational factors addressed explicitly by the method would normally be adequate without strong analysts to develop such a base (the analysis for the study was performed by the method developers, who were very experienced in PRA/HRA). In addition to guidance on scaling the PSFs or situational factors, guidance on developing the basis for selecting the key situational factors and judging them is also needed.

Treatment of response execution issues

As acknowledged above, since the EOC-focused method was intended to be applied in addition to an EOO approach, the method itself did not explicitly address how to treat execution issues or the execution part of HFEs. In the SGTR scenarios in particular, some crews had some response execution issues in some HFEs. Although the CESA-Q method developers note that ASEP or THERP would generally be used to address response execution, it is not clear (needs more investigation) whether those methods would adequately address the issues identified in the study (see NUREG/IA-0216, Volumes 1-3). Thus, an improved treatment of response issues may also be needed.

5.6 Overall Assessment of CREAM (NRI)

The Cognitive Reliability Error Analysis Method (CREAM [15]) was developed for general applications in the HRA field. CREAM has a well-defined analysis process, classification scheme, and model of cognition. The classification scheme gives the analyst a template in which to describe the details of the event, as well as to identify probable underlying causes. A model of cognition serves as a basis to organize or link different classification categories in an antecedent (i.e., genotype) and consequence (i.e., phenotype) model of human action. CREAM also provides the analyst with a clear stop-rule that indicates when an analysis has been completed. The CREAM method provides both basic (screening or scoping analysis) and extended (detailed analysis) processes. As implemented in the international HRA empirical study, only the extended process was used.

5.6.1 Strengths

Qualitative predictive power in terms of drivers

The greatest strength of the CREAM method, as applied to the analyses, is its ability to anticipate certain error types. The cognitive function failure types cause the analyst to consider the types of errors that might occur for each action. This approach is inherently conservative, and may overestimate certain types of errors. However, at the possibilistic level, this process holds great potential to anticipate certain errors that might be overlooked in other HRA methods. Selecting the dominant failure type holds promise for prioritizing likely error types. The CREAM quantification process does not, however, adequately distinguish probable from possible failure types.

The Extended CREAM method employed in this analysis did a good job of predicting cognitive failure types and identifying positive influences on behavior. The CREAM analysis identified four positive drivers: Procedural Guidance, human-machine interface (HMI), Training, and Experience. However, it also featured negative drivers that were not found in the empirical data.

Insight for error reduction

The CREAM method uses failure types, which categorize errors cognitively, offering a good basis for mitigation or error reduction. The CREAM documentation does not guide the selection of dominant failure types, and no explicit guidance is provided on using failure types for error reduction. Because the failure types used for HFEs are fairly generic, they may over-identify errors by being too broad in scope, thereby limiting their usefulness in error reduction.

5.6.2 Weaknesses

Lack of nuanced qualitative predictive power in terms of operational expressions

The CREAM qualitative insights are found primarily in the four cognitive functions: Observation, Interpretation, Planning, and Execution. These encompass a set of possible failure types, which included, in most cases, actually occurring error types. The failure types may be considered conservative in that they posit errors that may not actually occur, and there is some failure probability associated with them.

The CREAM analysis used the same three probable failure types for all HFEs:

- Delayed interpretation of symptoms
- Wrong planning after insufficient diagnosis
- Action performed too late

These failure types generally accounted for the errors actually seen in crew performance, and the CREAM analysis team did a fair job of predicting the operational expressions. Since the same failure types were used across all HFEs, there was a bit of a “one size fits all” approach in the analysis, and it would appear that the CREAM failure types are not sufficiently nuanced for this analysis application.

The CREAM analysis team conducted a thorough review and qualitative pre-analysis of the scenarios prior to encoding them into a CREAM-specific analysis. This process is compatible with CREAM, but it is not clear if the CREAM analysis benefitted from or was otherwise influenced by the analysis team’s pre-analysis classification of errors.

Overly uniform quantification results

The main weakness of the Extended CREAM method concerns the assignment of failure types. The assignment of generic error types, which serve as nominal HEPs, is subjective, and the process of determining the dominant failure type is complicated. For the effort required to complete this part of the analysis, the result is a list of highly similar nominal HEPs that do not appear to be conservative. The CREAM analysts in the study chose to forego the standard method of completing quantification in CREAM by not downselecting a single, dominant failure type. Instead, they considered all failure types for each analysis. This modified process may have inflated HEP values over those typical for a CREAM analysis, but the analysts saw this as a fair compromise to ensure reasonably conservative values in CREAM (and probably more realistic values).

Most HEP values are very similar across the HFEs, and represent similar assignments of failure types and common performance condition drivers. The selection of the failure type is the largest influence on the HEP. While this process is complicated, the value of differentiating the generic failure types is diminished by the large number of overlapping nominal HEP values. For example, while five generic failure types are provided for Execution (action of wrong type, action at wrong time, action on wrong object, action out of sequence, or missed action), all but one (action on wrong object) feature the same nominal HEP of 3.0E-3. The lower and upper bounds do vary, but the importance of selecting among these types has, in most cases, virtually no impact on the HEP. Thus, the method did not seem designed to discriminate between the HEPs in this study.

The effect of the common performance condition (similar to PSFs) multipliers on increasing or decreasing the nominal HEP may be negligible in a surprisingly large number of cases. Of the 29 levels or permutations possible across the nine common performance conditions for each of the four contextual control models (Observation, Interpretation, Planning, and Execution—29 levels x 4 failure types = 116 total), over half (62 of 116 total) have a value equal to 1, which does not change the nominal HEP. Another 13% (15 of 116) of the multipliers serve to modify (increase or decrease) the nominal HEP by 20% (i.e., multiplier equal to 0.8 or 1.2). The reviewer does not wish to refute the validity of this reliability distribution, but it should be noted that the multipliers tend to keep the values anchored close to the nominal HEPs. Only Adequacy of Time, Training, Experience, Procedural Guidance, and HMI (as represented in CREAM's common performance conditions) can singularly have a large effect on increasing or decreasing the HEP (by a factor of five). These may be seen as the dominant drivers on the HEP in the method, but they were not always adequate to account for what impacted the crew behavior.

Guidance and traceability gaps

While CREAM is well documented in the book of the same name [15], the analysis revealed a number of areas where the guidance and traceability could be improved. Foremost is the fact that the drivers (as designated by the common performance conditions (CPCs) in CREAM) do not have a strong effect on the quantification. Over half of the CPCs, even when denoting negative or positive influences, feature a multiplier of one, meaning that the heart of the qualitative analysis does not adequately feed into the quantification. This effect is magnified in terms of the relatively low HEPs produced by the method in standard practice. In the CREAM application featured in this comparison, the analysts deliberately circumvented the standard CREAM approach in order to drive up the HEPs to what the analysts felt were more realistic

HEPs. Furthermore, while most HRA methods account for dependency, there is no guidance on dependency in the standard CREAM documentation.

The Extended CREAM method is complicated by a lack of guidance to disambiguate the generic failure types amid the cognitive function failure types. The terminology can be confusing, and CREAM features more steps in quantification than do most HRA methods. In many cases, the selection of a specific failure type was not traceable beyond examples provided by the analysts. The reviewer does not mean to critique the analysts' specific assignments, but rather to point out that the selection of one generic failure type over another can seem arbitrary, and it appears that the method does not provide adequate guidance. The process in CREAM can introduce opportunities for subjective differences of opinion between analysts. Moreover, the selection of a single dominant failure type omits potentially valuable information about errors that could occur for that task. Most tasks, especially the HFEs modeled in these analyses, which spanned several minutes, must reasonably be seen as having Observation, Interpretation, Planning, and Execution components. All failure types should manifest, and it would be difficult to select a dominant one. It is to the analysts' credit that they have included every failure type in their analysis.

A detailed CREAM analysis features many steps not found in other methods, but that ultimately do not appear to produce a richer analysis than simpler methods. Several steps unique to CREAM include:

- Cognitive Function Failure Types
- Common Performance Conditions
- Critical Cognitive Activities
- Contextual Control Model (CoCoM)
- Probability Control Mode

While the concepts may prove useful to analysts, the proliferation of steps ultimately makes it difficult to compare a CREAM analysis to other methods, or to incorporate the products of CREAM analyses into the framework of a standard PRA.

5.7 Overall Assessment of Decision Trees (DTs) + ASEP (NRI)

According to the analysts who used the approach, the DT + ASEP method [1] represents a combination of two HRA principles: the decision tree approach, which reflects the work of EPRI (i.e., it was based to some extent on the DTs used in the CBDT method), and a modified ASEP approach. According to this approach (as used by the analysts in this study), the probability of failure consists of three main contributors:

- failure of information processing
- failure of diagnosis
- failure of manipulation

For each mechanism that could cause a failure in information processing, one decision tree has been developed to obtain the failure probability of information processing. The time reliability curve from ASEP is used to estimate the failure probability of diagnosis. For failure probability of manipulation, ASEP is used.

5.7.1 Strengths

Traceability

A strength of the method is that the judgments made by the analysts in applying the method and obtaining the HEPs are clearly traceable, in the sense that the decisions made by the analysts on the branches of the DT are fed directly into the quantification process and can be traced through the trees, based on the end points. With adequate documentation, the basis for judgments regarding which branches to take in the decision trees is traceable. The method for weighting the various factors in calculating the final HEP can be determined by examining the contributions of various factors to the overall HEP.

Insight for error reduction

Once the factors included in the method are correctly evaluated, with a good task and qualitative analysis, the method can provide guidance to facilitate error reduction.

5.7.2 Weaknesses

Inadequate guidance in addressing critical aspects of scenarios

It seems that the method's predictive power tends to degrade as the scenario complexity increases. The method's major limitation appears to be its inability to address complex diagnosis situations. The guidance, influencing factors, and specific questions asked during application of the method do not always seem to be adequate to identify critical issues at a more scenario-specific level, particularly with respect to the cognitive aspects. For example, although the crew had much more difficulty with HFE-1B than with 1A in the SGTR scenarios, the HRA team seemed to underestimate the severity of the issues in HFE-1B, as the method guidance did not guide them to a full understanding of the critical aspects of HFE-1B.

In some cases, a factor can be identified as a driver with a rationale that is inconsistent with crew data, and it seems that the qualitative analysis depends more on analysts' experience and expertise than the method guidance. For scenarios where the factors that would affect crews' performance are subtle and not obvious, additional guidance and/or inclusion of additional performance drivers seems to be necessary to correctly assess influencing factors and understand failure mechanisms.

Inadequate sensitivity in HEPs

Quantitatively, one apparent limitation of the method seems to be the lack of power (or sensitivity) in terms of HEPs to differentiate HFEs. For both scenario categories, it has been observed that the HEPs did not exhibit as much differentiation as was reflected in the crew failure rates. An apparent cause is an insufficient set of PSFs and inadequate guidance to identify important factors. The correspondence between the difficulty rankings based on HEPs and crew data was considered to be poor for the SGTR scenarios. Although the correspondence between the difficulty rankings was considered to be fair for the LOFW scenarios, the somewhat better correspondence may have been an effect of having only four data points (lack of sensitivity) for the LOFW scenarios (there were nine data points for the SGTR scenarios).

Inadequate guidance for dependency modeling

When compared to the crew failure rates and the upper Bayesian uncertainty bound, the HEPs for the LOFW scenarios (particularly for HFEs 2B and 1A) tend to be conservative. The conservatism for HFE-2B seems largely attributable to the use of the THERP dependency model, even though the HRA team accounted for potential negative dependency to avoid excessive conservatism. This suggests that improved methodology is needed to better model dependency.

Inadequate guidance to drive quantitative analysis with qualitative analysis

In contrast to the LOFW scenarios, the HEPs for the SGTR scenarios tended to be optimistic, particularly for difficult HFEs, such as 5B1 and 1B. For HFE-5B1 and, to some extent, 1B, the optimism seemed to be caused by a disconnection between qualitative analysis and quantitative analysis. Although many factors that the crew would face with these HFEs were identified, the method needs improvement to be able to better reflect the appropriate impact of such factors on the final HEPs. In addition, for HFE-1B, as discussed above, failure to fully understand the difficulties in the scenario also contributed to the optimistic HEP.

The disconnection between qualitative analysis and quantitative analysis was also observed in the LOFW scenarios. For example, although the qualitative analysis for HFE-2B was very accurate and several positive aspects that contributed to crew performance were identified, the analysis did not lead to an appropriately low HEP consistent with the data.

5.8 Overall Assessment of ENHANCED BAYESIAN THERP (VTT)

This assessment is based on an overall evaluation of the Technical Research Centre of Finland (VTT) team's analyses of both the SGTR and the LOFW scenarios, using the Enhanced Bayesian THERP method [1].

It should be noted that in the quantitative analysis of the SGTR scenarios, the Enhanced Bayesian THERP did fairly well. The HEP values were generally within the empirical error limits, and the analysis identified the relative difficulty of the different tasks (i.e., the ranking of the HEPs was consistent with the empirical difficulty rankings). However, the analysis did not provide a strong differentiation between the easy and difficult HFEs, with only one order of magnitude difference between the most difficult and the easiest HFE, and substantially less when the most difficult HFE is ignored. The collective evidence from the empirical data (the Bayesian bounds and the qualitative judgments leading to the difficulty ranking) suggests that this is not an adequate differentiation. On the qualitative side, there were difficulties in identifying the correct drivers and the weights for the PSFs in the scenarios.

In the Enhanced Bayesian THERP analysis, in some cases it may be sufficient to have the general effect of the PSFs be correct, while individual PSF weights might not correspond to the empirical data. The reason for this is that the HEP is driven by the time available for the task and that each of the PSFs is treated the same on the mathematical side, so the expert judgments that are obtained with the method are not necessarily required to identify each PSF correctly to arrive at the "correct" HEP number; rather, it may be enough that the task analysis with respect to the time available is accurate, and that the combined effect of the PSFs reflects the overall difficulty of the HFE. Obviously, however, this effect may not always produce an appropriate HEP if the PSFs are misunderstood, and the reliability of the method may not be good if sometimes accurate and sometimes inaccurate HEPs are obtained without really understanding what is going on in the scenario (i.e., they didn't identify the correct PSFs).

5.8.1 Strengths

Uncertainty analysis

The Enhanced Bayesian THERP method provides a systematic method for assessing the uncertainties in the analysis. Expert judgments of PSF weights are handled as observations of random variables. Prior distributions for PSFs are updated in a Bayesian fashion, which leads to posterior distributions for the factors. This approach allows the calculation of probability distributions and confidence limits for resulting HEP values.

Traceability of the quantification

The mathematical side of the method is easy to trace. The quantitative effect of the PSF weights is explicitly stated, but the basis for it may not always be clear. As with most of the PSF-based methods, the mechanistic steps of the quantification that is performed after the PSF weights are chosen are easily traceable, while the basis for choosing the PSF weights is not easily traceable, especially if the qualitative analysis is poorly documented. In this method, the reasoning behind the PSF weights is dependent on several different experts, and their reasoning might vary, so there is not necessarily a consensus for the qualitative analysis of the scenarios. In the SGTR analysis, the qualitative analysis was limited to reasons given by the experts for the values they had chosen. A descriptive qualitative analysis was supplied in the LOFW analysis.

Ease of use

The Enhanced Bayesian THERP method uses a simplified approach to task analysis, combined with a time correlation curve and five PSFs. The guidance for assessing weights is quite freeform, which leads into a method that is easy to use.

5.8.2 Weaknesses

Qualitative performance

The Enhanced Bayesian THERP method uses expert judgment to assess values for the PSFs. The other input from the expert panel is freeform qualitative statements about the basis for each PSF weight. The quantitative strength of the method is to form a Bayesian consensus of the PSF weights. The weights and their justifications can be contradictory. This reflects the different experts' differing views on how the operators would handle the scenarios, and the resulting composite HEP is the average of all possible predicted responses to the scenarios. No systematic method of forming a similar consensus of the qualitative statements exists. The qualitative SGTR analysis reflected the disagreements of the analysts, while in the LOFW this was developed into a more consistent qualitative assessment. There is, however, no guidance in the method for forming a consensus of the qualitative statements.

Guidance for PSF weights

The Enhanced Bayesian THERP method uses a set of five PSFs, each of which is given one of five possible weights (0.2, 0.5, 1, 2, 5) by each member of the expert panel. The method has guidance on how to assign the different possible values to the PSFs, but the guidance is limited, with only about one sentence for each possible weight/PSF combination. The guidance is also generic, such as “Mental load is considerable, situation is serious, a serious decision needs to be made,” without additional information.

PSF set

The method uses five PSFs: quality and importance of procedures, quality and importance of training, feedback from process/HMI, mental load, communication, and coordination. The definitions and guidance relating to the PSFs is also generic in nature, which means that each of these five PSFs can correspond to several of the drivers used in the Empirical HRA study. For example, the “Mental load” PSF used in the Enhanced Bayesian THERP method can be interpreted to include characteristics of both the “Stress” and the “Scenario complexity” factors identified in the empirical data. In both the SGTR and LOFW scenarios, this limited the ability to identify main drivers in the scenarios by lowering the resolution with which the difficulties of the scenarios could be characterized within the Enhanced Bayesian THERP method. Additionally, the qualitative reasoning behind the PSF weights was not enough to explain the in-depth reason for the chosen value.

Insights for Error Reduction

The Enhanced Bayesian THERP method does not specifically consider error reduction, but the insights from the most important negative PSFs can be used to identify areas for improvement, similar to any other HRA method.

Dependency

The method could benefit from increased guidance on how to handle dependencies (this is the case for many methods, and for HRA in general). This is apparent in the LOFW scenario analysis, where the analysis would have been more accurate if tasks 2A and 2B had not been assessed in isolation from previous tasks (i.e., if the analysis had considered actions 1A1 and 1B1 instead of 2A and 2B). The effect on the results is apparent, since the analysis considered little time to be available for 2A and 2B, even though these are dependent on 1A and 1B. If they were assessed as a continuation of the previous tasks, the results would be closer to the empirical data.

5.9 Overall Assessment of HEART (Ringhals)

This assessment is based on an overall evaluation of the Ringhals team’s analyses of both the SGTR and the LOFW scenarios, in which the Human Error Assessment and Reduction Technique (HEART) method [16] was applied.

In HEART, an HFE is quantified by matching a generic task type (GTTs, of which there are six) and adjusting the nominal HEP for this task type to account for the effect of error producing conditions (EPCs). The core of the method is the list of EPCs (over 30), each with a maximum multiplier corresponding to that EPC’s impact on the HEP at its most severe. To quantify, the GTT’s nominal HEP is adjusted by applying a proportion of the maximum effect for each EPC.

As applied in this study, the qualitative HEART analysis consists of identifying the EPCs relevant to the HFE and justifying each proportion of effect in terms of specific issues.

The qualitative predictions of the HEART analyses for the SGTR and LOFW scenarios were moderately poor to fair, while the quantitative predictions, on the whole, can only be rated as fair. The good quantitative performance seen in the HEART analyses of the SGTR HFEs, where the obtained ranking and consistency relative to the empirical bounds were among the best, was not repeated in the LOFW analyses.

5.9.1 Strengths

Focus on identification of error-producing conditions (negative factors driving performance)

While HEART does not define or specify a qualitative analysis approach, one of HEART's strengths is its focus on identifying whether these EPCs are present for a given HFE. By definition, an EPC is a driving factor; consequently, a HEART analysis focuses on identifying factors or conditions that drive performance.

5.9.2 Weaknesses

Coverage of generic task types and lack of guidance for identifying GTTs applicable to HFEs

The identification of the GTT applicable to an HFE, which anchors its quantification, may be difficult in HEART. For the HFEs in both the SGTR and LOFW scenarios, the same GTT, "Shift system state following procedure," was repeatedly used. Most GTTs were not applicable, and the assigned GTT was not a clear match. A mismatch in the GTT assignment can in some cases be compensated by identifying the difficult elements of the task as an EPC.

Lack of guidance for assessing proportion of maximum effect of the EPCs applicable to HFE

In a HEART analysis, the identification of an EPC as applicable is part of the qualitative analysis. Quantification of the HFE requires an assessment of the proportion of the maximum effect. There are no scales or anchor points for assessing these proportions. This lack of guidance for deriving the quantification inputs from a qualitative analysis leads to traceability issues (e.g., why is the proportion 0.2 vs. 0.6 for a given HFE?), and would be expected to lead to issues with inter-analyst consistency (method reliability).

Difficulties with modeling complex HFEs

Complex HFEs, which have many subtasks for situation assessment, response selection or planning, or execution, as well as multiple opportunities for errors, can be difficult to model with HEART. The method does not provide guidance for decomposing an HFE into subtasks, either for qualitative analysis or for quantification. A related issue is that the definitions and descriptions of the GTTs do not clearly address the decision making aspect of the HFEs. An open question for analysts is whether to model the HFE with one or two GTTs, although this would generally result in very different estimates of the HEP.

No means to model interactions between negative performance factors or EPCs

The HEART method treats the effects of the various EPCs as independent multipliers. There is no explicit mechanism for addressing potential interactions between the EPCs. For this study, the HEART analyses therefore did not address the overall operational expressions or HFE failure scenarios.

No means to model mitigating effects of positive factors

The method's design does not consider the reduction of the failure probabilities or mitigation of the effects of EPCs by positive factors. The definitions of some GTTs include some positive factors, such as "routine, highly-practiced task" or "following a procedure." In practice, however, these GTTs may not be applicable to a given HFE where the analyst may wish to take credit for this factor.

It should be noted that some of these weaknesses are well known, having been identified in previous studies, and that there are proprietary versions of HEART with additional guidance, as well as an effort to develop a new version of HEART, called NARA [17]. Neither this guidance nor the NARA method was available for the Empirical Study.

5.10 Overall Assessment of K-HRA (KAERI)

The Korean Human Reliability Analysis (K-HRA) method [17] is a thorough and sound extension of THERP and ASEP. It offers a clear decision tree approach that allows the ready extraction of drivers that can contribute to errors and provides a separate consideration of diagnosis and execution factors, which facilitates consideration of a wide range of error contributors.

The K-HRA analyses offered reasonable predictions predicated on logical assumptions. These predictions did not, however, always match the actual crew performance. It is possible that some factors, like operational culture differences between the Korean and the Halden crews, may have shaped the K-HRA analyses. Nonetheless, the assumptions and predictions in K-HRA were not unreasonable. Thus, it is not clear whether the K-HRA method is asking the right questions for the analysis. The sometimes poor matching between predicted and actual drivers suggests that additional guidance on the assignment of specific drivers and how to perform the qualitative analysis would be appropriate. There seemed to be a particularly large co-occurrence of drivers. The method does not control for double-counting of similar effects, and the available documentation does not articulate special considerations for the orthogonality of the drivers. Reviewing the interplay of drivers may further enhance the method's predictive reliability. K-HRA is ultimately a highly usable and efficient method, but its predictive ability may be hampered by the process of accounting for somewhat ambiguous performance drivers.

5.10.1 Strengths

Traceability

K-HRA offers high traceability of the quantification. The decision tree approach, with specific negative and positive assignments, translates directly into the calculation of the HEP. As with most HRA methods, there is significant room for analyst interpretation of these assignments. The decision tree approach makes decisions clear, but it may not always clearly document the rationale for a decision. As is the case with decision tree approaches, the general reason for a particular assignment is automatically recorded by selection of a specific pathway in K-HRA. However, if the analyst does not also document the rationale for selecting a particular pathway, the exact assumptions for level assignments may not be clear or replicable. In this case, the analysts have done a very good job of providing additional documentation of decisions made in the analysis. However, as with other decision tree approaches, it appears possible to complete a K-HRA analysis without the thoroughness demonstrated in the present analysis.

Easy to use method

Quantifying the HEP in K-HRA is straightforward: a simple set of level assignments (in most cases, encompassing both negative and positive influences) is made along potential driving factors for execution and diagnosis to compute the basic HEP. Separate basic HEPs are generated for execution and diagnosis and summed. THERP-style dependency is then considered in order to adjust the HEP and produce the final, conditional HEP. The entire quantitative analysis is based on a decision tree, but is accomplished in a straightforward spreadsheet, in which input states beget clear HEP outputs. The analysis successfully predicted the complex case to be more difficult than the base across the study scenarios. The K-HRA analysis did accurately predict all difficult HFEs, compared to the actual crew runs. While the method accurately predicted truly difficult tasks, it may not be sufficiently conservative in all predictions.

Insights for error reduction

The method, with its strong use of performance drivers, does a good job of accounting for different opportunities for error. Its use of separate diagnosis and execution inputs further leads it to consideration of a wide range of error contributors. While the available English level documentation does not discuss error reduction, the assessor believes the method is well suited to this application. However, as noted below, its predictive ability sometimes varied from actual observed performance. K-HRA seems to do a good job of predicting the most error-likely tasks, but it may not always predict the correct drivers that lead to those errors. This result clearly suggests that additional research is needed to verify the validity of the method.

5.10.2 Weaknesses

Conservative identification of drivers

The drivers available in the K-HRA method align closely with those used in the empirical assessment. However, the K-HRA analysis was overall moderately poor in terms of its predictive power for drivers. The factor that K-HRA most closely captured was HMI. Across most HFEs, K-HRA overestimated the negative influence of drivers. Its predicted positive drivers did not consistently match the observed positive drivers on crew performance.

It seems that many of the drivers tended to cluster together (e.g., Time Pressure, Stress, and Execution Complexity tended to co-occur). Although these factors were not typically observed together in the crews, it is not unreasonable or unusual in HRA to group these drivers. The true orthogonality of driver combinations, as well as each HRA method's definitional orthogonality, are not clearly understood. While K-HRA may seem quick to attribute multiple negative (or sometimes positive drivers), the co-occurrence of these drivers helps to ensure that the method covers performances that may actually occur. The downside of this is that there may be some double-counting of effects; the positive side is that the method is more likely to offer a conservative account of performance, as is appropriate for certain applications of HRA.

Lack of explicit treatment of operational expressions

The K-HRA method is simplified in the spirit of ASEP, and does not provide explicit guidance or treatment of the qualitative approach beyond the PSFs covered in the decision trees. In practice, the method predicted crew performance based largely on the crew's familiarity with the scenario, and as a function of how much time is available to complete the task. For the base case scenarios, it was assumed that the crew should be quite familiar with such scenarios and should perform well, with the possible exception of the tight time constraints posed on each task. Those complex scenarios that deviated from the familiar or expected course of activities were predicted to present considerably more difficulty to the crews.

Limited guidance on the method and application of the method

Guidance for the K-HRA method is quite limited in English. K-HRA serves as a tool for the Korean nuclear industry, and method documentation is accordingly available primarily in Korean. However, K-HRA is based on popular methods, which are extensively documented, and for which good guidance is widely available.

The method is a simplification of THERP and ASEP. Part of this simplification is manifested in limited guidance provided on performing a qualitative analysis, beyond reviewing dominant drivers used in the methods. There is room for interpretation in many of the driver-level assignments, leading to potential differences between analysts. Additional guidance may be needed to complete assignments correctly.

5.11 Overall assessment of MERMOS (EDF)

This assessment is based on an overall evaluation of the Electricité de France (EDF) team's analyses of both the SGTR and the LOFW scenarios with the Méthode d'Evaluation de la Réalisacion des Missions Operateur la Sûreté (MERMOS [19]and [20]) method.

5.11.1 Strengths

Method's focus on identifying HFE failure narratives (how the failure occurs)

For each HFE, the HRA analyst applying the MERMOS method identifies failure narratives or failure scenarios that describe how scenario, crew, and task characteristics may interact to result in the failure of the HFE. This process means that the predicted failure probabilities are based on and directly connected to specific qualitative findings concerning the HFE and its context.

Systematic process for identification of potential failure narratives

Although the qualitative analysis to identify specific failure narratives relevant to a given HFE relies on the expertise of the HRA analysis team, the process for identifying these narratives is systematic and provides a basis for failure narrative categories. This contributes to the traceability of the analysis and supports the analysis team (and, to some extent, external reviewers) in reviewing the comprehensiveness and plausibility of the scenarios.

Multiple failure narratives can be considered for an HFE

MERMOS analyses frequently identify multiple failure narratives for a given HFE. This provides the chance for the HRA analysis team to consider potential variability in the characteristics of the crew, in the evolution of the scenario, or in the way a crew may respond to a situation in the plant or manage a task.

Generation of insights for error reduction

Failure narratives describe how elements of the scenario, task, HMI, and operator aids may contribute to the HFE. The failure scenarios can be directly understood by plant experts, and the specificity and level of detail of these narratives makes them directly usable for error reduction. The quantification of the HFE relies on the presence of the assessed probability of these elements in a given situation. This highly traceable relationship between the failure scenarios and the resulting HEP makes the HEP very responsive to changes in the system, interface, training, or aids.

Qualitative predictions

In terms of qualitative predictive performance, the MERMOS analysis was rated moderately good to good. Particularly for the more challenging HFEs, the elements of the failure narratives predicted in the MERMOS analysis were supported by the empirical evidence. The specific narratives, which describe how the identified elements may interact to lead to failure, were also supported by the observations of how the failures occurred in the simulator.

5.11.2 Weaknesses

Quantitative differentiation

The qualitative performance of the analysis was not matched by its quantitative performance. While most HEPs were consistent with the empirical bounds, the ranking of the HFEs was not so consistent with the empirical evidence. In both the SGTR and LOFW phases, the quantitative differentiation among the HFEs was less than expected, given the qualitative predictions. In addition, the empirical evidence tended to support a greater differentiation of the HFEs. Whether this was only the case with the present analysis, or whether it points to a general weakness in the method, could not be determined in this study.

Extensive reliance on expert judgment and impact on external reviews

Both in the identification of scenarios and in the quantification of the elements of the failure narratives, the MERMOS method relies extensively on the HRA team's expertise and familiarity with the plant and operations, without reference to external evidence. For an external reviewer, evaluating the plausibility of the HEPs obtained with the method may be difficult.

Inter-analyst consistency (method reliability) is an open issue

Inter-analyst consistency for MERMOS is an open issue, given that MERMOS analysis relies strongly, as all methods do, on expertise for the qualitative analysis (identification or construction of failure narratives) and quantification (quantifying the probability of the contributing elements of the failure narratives). A database of past MERMOS analyses, which covers the HFEs in EDF's probabilistic safety analyses (PSAs), supports analysts by providing a catalog of failure narratives. This database was used in the LOFW phase. Its potential for supporting the qualitative analysis is clear, though its impact on the quantitative analysis is less clear.

5.12 Overall Assessment of PANAME (IRSN)

This assessment is based on an overall evaluation of the Institut de Radioprotection et de Sûreté Nucléaire (IRSN) team's analyses of both the SGTR and the LOFW scenarios, applying the New Action Plan for the Improvement of the Human Reliability Analysis Model (PANAME) method [1].

5.12.1 Strengths

Guidance for Users

The PANAME method includes extensive guidance on its use, especially for determining the weights of the PSFs. The guidance for determining PSFs includes specific indications that the analyst should look for: for example, in the case of the procedure quality PSF, the analyst is directed to determine whether the EOPs include a fold-out page that is required for the scenario at hand. This can be considered a good thing in the sense that it increases consistency between different applications of the methodology. Similarly, this makes the method approachable for persons with different levels of HRA knowledge.

Like several other HRA methods, the time correlation curve for diagnosis success is based on Swain's work on THERP [9]. The guidance for choosing between curves corresponding to different levels of difficulty is easy to understand in the PANAME method.

For determining the context factor (a composite of six PSFs) and choosing the time correlation curve, PANAME utilizes a decision tree-like approach, which increases the usability.

Traceability of the analysis

Quantification is easily traceable in PANAME. The PSFs' effect on the quantitative HEP is explicitly stated in the documentation. The mathematics of the method are simple to understand. Depending on six PSFs, a single context factor is chosen. The context factor can have a value of 1/3, 1, or 3. This should allow the analyst to discern between easy and difficult scenarios; however, the resolution of the method is not that high.

Traceability of the chosen context factor is clear from the use of the decision tree-like approach, even though in the LOFW analysis the reasons are also stated explicitly in the accompanying analysis worksheets. The guidance for using the method also plays a part in the traceability of the analysis.

Presentation of results

The documentation of the analysis was improved from the SGTR to LOFW scenarios. The worksheet included with the LOFW scenario analysis is useful for understanding the analysis, and for giving insight into the component factors of the analysis. The composition of the total human error probability is easily seen as the quantitative side of the analysis, and is broken down into an easily understood form.

5.12.2 Weaknesses

Rigid Guidance

The guidance for the PANAME method is listed above as strength, but there is another side to this. While the guidance is complete and specific in terms of PSFs treated in the method, this can also have a limiting effect on the analysis, if followed strictly. That is, the range of PSFs may not be broad enough, and the specific instances of the PSFs covered may not cover other related aspects that might be addressed under a given PSF. The specific requirements listed in the guidance for assessing a certain value for a PSF might not be applicable in each case, and the PSF might not be factored into the analysis. The results may lead to an inaccurate analysis if the correct reasoning for a PSF weight is not in the guidance. For example, for HFE-1B in the SGTR scenario, the PANAME analysis assessed the context factor as nominal, despite missing indications. The decision tree rigidly prevented the analysts from assessing other values for the context factor.

Quantitative resolution and calibration of the method

The PSFs of the PANAME method are combined into a single context factor that can have values generally ranging from 1/3 to 3. This means that the representation of the difficulty range resulting from the context of the task is limited compared to most other methods, and that most of the differentiation between easy and difficult scenarios has to come from the diagnosis part. This diagnosis part is sensitive to time, and, to some degree, to the assessed difficulty of the task. The PANAME method's context factor should be compared to empirical data and other methods. The quantitative constants of the method are different from those of many other methods.

Additionally, the high diagnosis failure probabilities caused some problems in the complex LOFW scenario because the method assessed the total failure probabilities to be equal to 1, meaning that there was no possibility of success. This was mostly because the assumptions used for available diagnosis time turned out to be wrong when compared to the empirical data. The method generally yielded pessimistic results for the LOFW scenario.

Sensitivity analyses are not covered explicitly in the method description. In effect, the upper and lower values were achieved by varying the time available for the task. The method is very time-sensitive, due to the variability of diagnosis failure as a function of time, but this approach does not cover the method's sensitivity to other factors.

Improved qualitative scenario analysis is needed

The PANAME analysis did not capture the negative drivers adequately in either the SGTR or the LOFW scenarios. The method itself is not geared towards producing qualitative information. The decision tree approach to analyzing scenarios necessarily limits the degree to which the qualitative differences of the scenarios can be assessed.

Insights for Error Reduction was not achieved

The PANAME method does not offer specific guidance for error reduction. The output of the method could be used to identify the most important negative drivers affecting the HFEs, but the low resolution of the context factors (PSFs) limits the usefulness of this approach.

5.13 Overall Assessment of SPAR-H (INL)

The Standardized Plant Analysis Risk-Human Reliability Analysis (SPAR-H) method (NUREG/CR-6883) [21] was developed as a simplified HRA quantification technique based around eight predefined PSFs and separate nominal HEPs for diagnosis and action tasks. The SPAR-H method is easy to apply in order to arrive at the HEP, which may be seen as the method's greatest strength. This ease of use may be misleading, because in practice there can be great complexity in performing the underlying qualitative analysis and mapping the findings of that analysis to the SPAR-H PSFs. These shortcomings are mostly seen as the byproduct of inadequate guidance on performing a successful and complete SPAR-H analysis. The current documentation does a good job of explaining the method, but it falls short in providing examples and guidance on deciding between competing levels of assignment for aspects such as PSFs. Moreover, the underlying qualitative analysis process is not clearly documented in SPAR-H. The SPAR-H method is easy to apply, but it is also potentially easy to misapply. Additional guidance would help to prevent the unintentional misapplication of the method and potentially inappropriate results.

5.13.1 Strengths

Ease of use

The SPAR-H method is a simplified technique that uses easy-to-follow worksheets. Compared to other simplified techniques like ASEP, SPAR-H features few exceptions to the process flow, and the worksheets can be completed with minimal experience or training. With supporting documentation, the SPAR-H method provides a simple and traceable approach to quantification. While the ease of use is a benefit to experienced analysts, it also presents pitfalls for less experienced analysts, as it belies the need for a thorough qualitative analysis, which is not elucidated in the SPAR-H guidance. Moreover, because the method uses a checkbox approach, it is actually possible to complete an analysis without a thorough understanding of the HFE. The possibility for misapplication or overconfidence by inexperienced analysts must be considered as a limitation of the method.

Insights for error reduction

The PSFs included in SPAR-H should allow insights into improving safety and reducing errors, although the current method does not specify the process for doing so. In order to assist in error reduction, additional guidance is needed for performing the qualitative analysis and assessing the influencing factors at a more scenario-specific level. The findings produced by the PSFs are at a fairly coarse level, and would likely not allow detailed insights into error reduction. Individual analyses using SPAR-H might, of course, provide the right level of detail, but this would be attributable to the analyst's skill rather than to the guidance in the method. Still, the PSF approach allows a streamlined pinpointing of error sources that holds potential for error reduction applications.

5.13.2 Weaknesses

Lack of qualitative analysis guidance

The SPAR-H method lacks guidance for performing the qualitative analysis to systematically determine which aspects of a scenario affect crew performance. In fact, the method documentation explicitly labels the method as a quantitative method and defers to other methods like ATHEANA in order to complete a detailed qualitative analysis. It is, in the assessor's opinion, not reasonable to expect an adequate quantitative output without a reasonable qualitative analysis that is tied to the quantitative model first, and this deficiency hinders the utility of the method. The Idaho National Laboratory (INL) SPAR-H team seemed to perform a more thorough qualitative analysis than is required by the SPAR-H method. The analysis provided thoroughly documented operational expressions, which are neither required nor explained by the method.

The level of assignment of the PSFs in SPAR-H is subjective. The identification of the relevant PSFs seemed to be guided by the analysts' knowledge and understanding. It is not clear how the SPAR-H method guidance was used to identify important drivers. The mapping from the operational expressions to the PSFs did not seem to be entirely successful, suggesting the need for better guidance in SPAR-H on performing a qualitative analysis, particularly to support the selection and weighting of the PSFs. Across the scenario HFEs, the INL SPAR-H analysis team seemed to map the qualitative analysis to the PSFs consistently, but the reviewers wondered whether another underlying qualitative analysis or a different analysis team would have arrived at the same mapping. Furthermore, it was not always clear how decisions were made as to which multiplier should be chosen to determine the strength of the driver. The analysis team did a good job of documenting the assumptions behind their PSF assignments, but the ultimate mapping to the assignment level and multiplier was not always transparent.

An explicit, detailed operational story/description was not provided in the INL SPAR-H analysis for comparison with the empirical data because the SPAR-H method is PSF-driven. Some assumptions that were made about what would occur, especially for the easier HFEs, did not appear to be based on an examination and understanding of potential issues the crews might have (e.g., choices the operators could consider to accomplish an action). The SPAR-H method did not guide the analysts to consider such factors when assigning PSFs. For example, the assignment of the nominal level to the "Procedures" PSF in SPAR-H on the grounds that it entails a "procedure-driven action" does not adequately gauge the completeness or the suitability of the procedures.

PSF ratings and the use of diagnosis and their impact on quantification

The HEPs calculated across the SPAR-H analyses generally reflected differences between base and complex scenarios. However, the method produced alternating conservative and optimistic HEP values for several relatively easy HFEs and one extremely conservative HEP, compared to actual crew performance. SPAR-H includes multipliers to account for the PSFs, which allows the analyst a lot of flexibility in deriving HEPs. However, appropriate PSF level assignment is not clearly documented or constrained in the method, making it more likely to produce different values for different analysts.

The SPAR-H guidance (NUREG/CR-6883 [21]) is not clear as to whether diagnosis should be modeled in applying procedures. On the one hand, it states that "guidance for action has to do with carrying out one or more activities (e.g., steps or tasks) indicated by diagnosis, operating rules, or written procedures." On the other hand, it also states that "when answering the question 'Does this task contain a significant amount of diagnosis activity?' one should consider whether the operator or crew has to expend mental energy to observe and interpret what information is present (or not present), determine what that means, think of possible causes and decide what to do about it." Thus, if the analyst judges the procedure to be simple procedure-following, one may model it as pure action. However, if the analyst judges the procedure to contain interpretation, as would be the case with judgments of increasing levels, one may include the diagnosis portion when analyzing procedure-following. Since all HFEs in the scenario addressed cognitive tasks, it is troubling that the SPAR-H analysis would readily discount some tasks as being solely “Action”-oriented in the SGTR scenarios. Because the classification of a task as “Diagnosis” or “Action” directly affects the nominal HEP in SPAR-H (resulting in a nominal HEP of 1E-2 and 1E-3, respectively), this practice in SPAR-H has tremendous implications for the quantitative result of an analysis. For the LOFW scenarios, the INL SPAR-H analysis team adopted the position that every HFE should include both diagnosis and action components. Review and revision of the nominal HEP values is recommended for “Diagnostic” and “Action” tasks, as well as for some of the PSF multipliers (like the multiplier of 50 for missing parts of the HMI), because there is a strong potential to produce overly optimistic or overly pessimistic results.

Conservative treatment of dependency

The use of dependence drove up the HEPs considerably, but it was not clear that this met the intent of the analysts. For example, in the LOFW scenario, HFE-2A was considered unlikely by the analysis team, but the overall HEP rose from 5.77E-2 without dependency to 1.92E-1 with dependency. This represents a highly likely event probabilistically, which is a poor reflection of the qualitative insights by the analysts. This finding suggests that additional guidance on dependency assignment would be useful to help analysts better reflect their intent in the analyses. The SPAR-H method features a simplified version of THERP dependency, applied between HFEs. In this simplification, dependency becomes more likely for scenarios involving more HFEs. The inherent conservatism in this assumption is not clearly justified in the method’s documentation.

5.14 Overall assessment of SPAR-H (NRC)

This assessment is based on an overall evaluation of the U.S. Nuclear Regulatory Commission (NRC) team’s analyses of both the SGTR and the LOFW scenarios.

5.14.1 Strengths

Captured the complexity

In general, the LOFW analysis was better than the SGTR analysis; the qualitative analysis in particular was improved over the SGTR analysis. The analyses were more detailed and had good descriptions of the use of time and the impact of the failing SG level measurements on complexity. The fact that the analysis team did a better and more thorough job on the LOFW analysis than on the SGTR analysis is not really a strength of the method. However, it does show that the method is able to incorporate detailed qualitative descriptions and use these descriptions and knowledge in the quantification. On the other hand, one could say that the method is in many cases dependent on a detailed knowledge of the scenarios and the situations, and the method does not guide the analysts to perform a detailed qualitative scenario analysis (e.g., a walkthrough of the procedures with an operator) to provide the knowledge. See below on weaknesses, in the discussion on the need for a detailed qualitative scenario analysis.

The quantification itself is easily traceable in SPAR-H

The simplicity of the base probabilities and the adjustable PSF multipliers makes it very easy to trace where the numbers come from in SPAR-H. Thus, the traceability of the quantification itself for SPAR-H, meaning the link between the PSF weights and the HEP value, is good. However, the basis for the judgments on the PSFs still relies on the documentation provided by the analysts. The method does encourage appropriate documentation.

Simple descriptions of tasks, link to PRA

The analyses included short descriptions of the main tasks for each HFE, and also descriptions of the actions as represented in the SPAR models. This simplifies the review of the HRA as part of the PRA.

Easy to use

The simplicity of SPAR-H may be seen as an advantage. In the present analyses, the explanations of how the PSFs would impact the simple, "standard" HFEs were good; thus, for quantification of "vanilla," or nominal, scenarios, SPAR-H may be appropriate. The challenge and warning is that SPAR-H may be too weak in separating a simple HFE from a difficult one (i.e., identifying the difficult HFEs). In these analyses the team often did not manage to pinpoint the difficult HFEs, so the method may lack adequate sensitivity. Thus, simplicity is not always an advantage.

5.14.2 Weaknesses

Optimistic predictions of the most difficult HFEs

In both the SGTR and the LOFW, the NRC SPAR-H analyses were optimistic in their predictions of the difficult HFEs. It seems that this was mainly because (1) they used only the task type action instead of addressing both the action and the diagnosis (especially in the LOFW scenarios), and (2) when they did address diagnosis, they chose which PSF and what weight to apply for the complexity issues in the scenarios that did not match the empirical observations. Since the HRA team appeared to follow the guidance in the SPAR-H method in making these choices, it seems that the guidance would need to be modified (see the paragraph below on the Complexity and the Ergonomics PSFs). Note that the INL SPAR-H team (the other SPAR-H application) was not optimistic in these HFEs, so the guidance did not lead the different teams to the same results, meaning that this is an inter-rater issue.

Task types Diagnosis and Action should be included in all analyses of all HFEs

The empirical results of this study have established that procedure-driven actions should not be assumed to deal with execution-type tasks only. The crews continually perform cognitive activities while executing procedures (see the conclusions and the empirical results chapters.) The NRC SPAR-H team chose only the task type "action" for most of the HFEs, giving 1E-3 as the base probability before adjustments by PSFs ("diagnosis" task type has a base probability of 1E-2). Thus, this choice has a considerable impact on the HEPs. The NRC SPAR-H team justifies their choices thoroughly, based on direct citations from the SPAR-H documentation. One interpretation of the guidance is that if the crews are following procedures, it is not necessary to analyze the HFE as including a diagnosis activity. Another interpretation is that if the procedure-handling includes any diagnostic activity, one should include the diagnosis activity in the analysis. The documentation needs to be made clearer, and the data from this study suggests that the best strategy would be to always address diagnosis and execution aspects when quantifying a post-initiator HFE.

SPAR-H should clarify when to use Complexity and when to use Ergonomics/HMI

The guidance for SPAR-H needs to be improved regarding the choice of PSFs. The choice between the “Complexity” PSF or the “Ergonomics/HMI” PSF in particular should be clarified, possibly with more examples. In both the SGTR and the LOFW scenarios, the NRC SPAR-H team analyzed the misleading indicators as part of the “Complexity” PSF. They cited the SPAR-H guidance directly for this choice, and they did follow the guidance correctly as far as this assessor can judge. The INL team instead used the “Ergonomics/HMI” PSF, where the SPAR-H guidance for the PSF rating seems to apply directly. The NRC team evaluated this PSF but interpreted the overall scope of the Ergonomics/HMI PSF as referring to limitations in the overall HMI design (and excluding failures of the HMI). This should be sorted out in the guidance documents of the method. The Ergonomics/HMI PSF also has a much larger weight and, thus, a higher maximum impact on the probability than the Complexity PSF. If the analyst judges the impact of complexity issues to be severe, he/she might feel inclined to choose the Ergonomics PSF in order to credit the phenomenon correctly regarding the impact on the HEP, and in that way bend the method. A critical review of these two PSFs should be done in SPAR-H.

Rating of PSF weights (choosing multipliers) can cause inter-rater variability

The decision as to which PSFs to rate positively and which to rate negatively is clearly based on the analysts’ judgment in using SPAR-H, and it is not always obvious why the choices are made. Deciding which and how many PSFs to include as negative or positive influences and how to assign the PSF levels seems like it can be a complicated process in SPAR-H, at least for these types of scenarios. Some additional guidance in SPAR-H as to how to consider the PSFs together and make such judgments would be very useful. A better qualitative analysis would improve the basis for these judgments.

Decisions about multipliers can be based on a number of factors, and SPAR-H probably intends to be relatively flexible in this regard; ultimately, it is up to the analyst. However, this can lead to poor reproducibility. If analysts are expected to consider the relative weights across PSFs, which appears necessary, additional guidance and documentation would be helpful.

The justification for each choice regarding the levels of the PSFs is encouraged in the comments section of the SPAR-H data sheets, but this is very much up to each analyst. Some extra requirements or structure for documenting the choices in operational terms would be beneficial. This would improve the traceability of the basis for the quantification.

More detailed qualitative scenario analysis is needed

In spite of the improved qualitative analysis in the LOFW scenarios as compared to the SGTR, none of the analyses of the scenarios were detailed enough to capture the drivers and the operational situations that created difficulties for the operating crews. The analyses lack the level of detail in describing operations to explain issues about procedures in difficult scenarios. For example, in the LOFW HFE-1B, it was not enough to strictly follow the procedures, since these did not cover the failing level measurements or the parallel problems with the condensate pumps. It seems that stronger requirements for a more detailed task or scenario analysis should be guided by the method.

In the SGTR scenarios, the NRC SPAR-H team used pretty much the same PSFs, with similar weights, for all of the HFEs in the whole sequence of each scenario. For the complex scenario, they judged the complexity and stress to be similar for almost all of the HFEs throughout the scenario. The assumption that only a few PSFs can be used to identify the correct drivers and their impact for all HFEs in each scenario turned out to be too simplistic. For each HFE, various drivers were identified in the empirical data, based on operational issues on a more detailed level. A more detailed analysis for each HFE seems to be necessary to identify the right drivers. A weakness of the method is that it does not state in enough detail how far the analyst needs to go in analyzing the scenario and procedures in order to get a good understanding of the operational difficulties the crews will face. One solution to this might be to use an initial screening, and, if one suspects that a scenario is very complex, one may consider using another method, such as ATHEANA, to identify the difficulties involved for the crew. SPAR-H itself proposes this in the foreword of NUREG/CR-6883 [21]: “Nonetheless, as a simplified method, SPAR-H has inherent modeling and analysis limitations that should be clearly understood. The SPAR-H Method should not necessarily be preferred over more sophisticated and detailed approaches, such as A Technique for Human Event Analysis (ATHEANA), in situations that require detailed analysis of the human performance aspects of an event.” SPAR-H also proposes the ATHEANA search process as a means of gaining knowledge of the tasks and contexts to be rated (ibid., p. 56).

Insights for Error Reduction were not achieved

This SPAR-H analysis does not provide many operational details for specific parts of each HFE (e.g., specific parts of predicted difficulties in procedure use are not discussed). The team does, however, note which parts of the procedures are relevant and which conditions and goals the crews are aiming for. Overall, the analysis gives few insights into error reduction.

Dependency

Treatment of dependency should be improved in HRA. SPAR-H utilizes a simplified version of THERP dependency, applied between HFEs. The NRC team discussed two ways of treating dependency for the two relevant actions in the LOFW scenarios. First they discussed what could be considered normal for SPAR-H, modeling two HFEs and applying the dependency calculation. However, they ended up modeling the two HFEs as one composite action. This is an interesting choice, but is not dependent on which method they were using. The treatment of dependency should be improved for most HRA methods, not only for SPAR-H.

6. CONCLUSIONS ABOUT HRA AND THE HRA METHODS

A review of the strengths and weaknesses of the various methods described above in Section 5 identified a number of common features among groups of methods that are important to understanding limitations in current human reliability analysis (HRA) methodology and how to improve it. These are discussed in this section. While important limitations from all methods were identified, it should also be noted that all of the HRA methods had good features that often allowed them to obtain relevant information about the conditions influencing performance, identify important drivers of performance, and produce reasonable human error probabilities (HEPs). The methods are all based to some extent on psychological/cognitive models, and, with a few exceptions, they have generally not had the benefit of empirical studies for either their development or their testing. This study at least provided evidence that the frameworks employed by the different methods are not unreasonable, and can be used to predict performance and estimate the likelihood of success/failure for many cases. Useful information was often obtained from applying the methods, and the limitations can be seen as reflecting a lack of robustness in HRA methods, rather than poor overall validity. Thus, a particularly important benefit of this empirical study is that it identified a clear means to improve HRA methodology and increase its accuracy and reliability.

6.1 Explaining variability in results

This study identified significant variability in both the qualitative and quantitative HRA results. It is important to understand the sources of variability between the methods, and some of the factors or dimensions that can affect the variability in predictions are discussed in the sections below. It should be noted that variability should not be unexpected, since the methods have very different theoretical bases and approaches for understanding and quantifying human failure events (HFEs). Examples of these differences include:

- Identification of failure mechanisms at a fairly detailed level (e.g., ATHEANA (error-forcing context and unsafe acts), MERMOS (stories), CBDT (failure mechanisms))
- Identification of generic failure types (e.g., CREAM, HEART)
- Detailed task analysis and (among other aspects) use of time-reliability correlations to treat crew cognitive activities (e.g., THERP, ASEP)
- Reliance on assessment of a selected set of performance-shaping factors (PSFs) (e.g., SPAR-H)

Given the differences in the methods, some of the factors that can affect the variability in predictions can be seen as method-driven, including:

- The ability of the method to capture significant influences on behavior
- The depth of qualitative analysis acceptable to the method, and the degree to which it leads to an understanding of the underlying dynamics of the scenario and driving factors
- The ability of the method to accommodate the analysts' knowledge and understanding of the HFE and scenario context in a way that allows a characterization of the relative difficulty of the actions associated with the HFEs

- Guidance provided by the methods

Other issues may be analyst-driven, including:

- Whether the method has been applied as intended
- The depth of the qualitative analysis undertaken to understand the underlying dynamics of the scenario and factor it into the estimation; this can go beyond what was required by the method, and, to some extent, is a function of the two factors listed immediately below
- The team experience in HRA and with the method applied
- The degree of expertise in human performance and plant operations needed to apply the method

This project had a limited capacity to cast light on some of these factors. Certainly the last two items were not easily evaluated, although some relevant discussion is provided below. Furthermore, in some cases, it is difficult to distinguish between the effect of the method and the effect of the analysts. As discussed earlier (in the introduction to Section 5), a different study would be required to empirically separate many of these aspects. Nevertheless, in this study, we have investigated the analyses both qualitatively and quantitatively, to the extent possible, and have obtained reasonable insights into the strengths and weaknesses of the methods that may enable or hinder analysts in performing good analyses, and into the important characteristics that need to be addressed to reduce the variability in HRA results.

6.2 Summary of Lessons Learned about the HRA Methods

Address cognitive activities more comprehensively

The data and observations on the crew responses in the simulated scenarios confirmed the general observation that crews perform cognitive activities together with the collection of plant information throughout the scenario's evolution. These cognitive activities and information gathering are intended to help the crews (1) understand the plant situation and decide on the appropriate response, and (2) make decisions during the execution of the selected response plan. In current nuclear power plant (NPP) operations, these cognitive activities are supported by emergency operating procedures (EOPs) and other procedures, and occur while the operators are following EOPs.

HRA methods commonly use the term "diagnosis" to refer generally to cognitive activity; in addition, diagnosis is often broken down into Detection-Diagnosis-Decision phases, and is viewed as preceding an "execution" or "implementation" phase. This terminology and its interpretation place an understandable emphasis on the decision to take a particular action, because this must be successful before the execution of the action is considered. The decision to take a particular action requires that the key plant cues be perceived/detected, and that the crew develop an appropriate understanding of the plant state (diagnosis). Thus, this terminology tends to be oriented to the cognitive activities aimed at "(1) understand[ing] the plant situation and [deciding] on the appropriate response," which may also be referred to as "initial diagnosis," rather than those intended to help "(2) make decisions during the execution of the selected response plan."

Within this study, the HFEs with an important component of "initial diagnosis" (cognitive activities with aim "(1)") include, for instance, HFE-1A in the steam generator tube rupture (SGTR) scenarios, where the crew must diagnose that there is an SGTR and identify the affected steam generator, and HFE-5B1 (again in SGTR), where the crew must first diagnose that there is a leaking power-operated relief valve (PORV). In contrast, HFEs with limited or no "initial diagnosis" part but with cognitive activity during execution include HFEs 2A, 2B, 3A, and 3B (in SGTR), which deal with reactor coolant system (RCS) cooldown and depressurization. For these, the cognitive activity deals with the selection of the appropriate cooldown rate[5] and the decision to stop cooldown or depressurization when the desired plant state is reached; an incorrect cooldown rate or a premature decision to stop cooldown or depressurization leads to a failure to implement the action.

These HFE examples show how cognitive activities with aim "(2)" may lead to the failure of a given HFE during the execution of a selected response. The assessments of the HRA predictions show that some analyses focused on "initial diagnosis" (and the decision on an appropriate response for this plant state), and did not fully consider the challenges with cognitive activities during the execution of a selected response. In summary, a comprehensive analysis of cognitive challenges and possible failures needs to consider both "initial diagnosis" and decisions related to executing a response plan. In analyzing an HFE modeled as following the success of a preceding HFE, it is also important to examine the scope of the cognitive activities (whether initial diagnosis or decisions related to execution) and the associated failure modes modeled for the preceding HFE in order to ensure that the full scope of cognitive activities related to the crew's overall response to the scenario (which may in some cases be divided among HFEs) is addressed.

In addition to the ambiguity or bias introduced by referring to all cognitive activity as "diagnosis," the guidance for several of the methods, including SPAR-H, ASEP, and CBDT allows analysts to make a modeling decision not to explicitly address the cognitive demands associated with following emergency procedures during the execution of the response. While each of these methods includes its own approach to address and quantify the cognitive aspects of a task, analysts have the option to model HFEs subsequent to the initial HFE or the identification of the event (e.g., entering the correct procedure) as purely task-oriented (i.e., purely execution or implementation). For example, for some HFEs, the SPAR-H and ASEP analyses did not include a cognitive contribution to the HEP, and, in the CBDT+THERP analysis, the analysts decided not to use the CBDT to estimate the HEP for some cognition-based HFEs addressed in an earlier event (e.g., an event in the model) and instead included only the execution contribution using THERP.

In addition, the results of the study clearly showed that failure to adequately consider the crews' cognitive activities and related potential failure mechanisms while they are working through the procedures can in many cases lead to a failure to identify important influencing factors and result in underestimations of HEPs. Thus, it was concluded that cognitive activities involved in following EOPs and related procedures should always be addressed in modeling and predicting crew behavior. Since for most methods these activities are addressed by applying their diagnosis model, this implies that in general a diagnosis portion of all HFEs (along with the execution portion) should be examined and quantified to ensure that potential problems are not missed. However, it should be noted that not all methods include an approach for quantifying

[5] In contrast to the selection of the cooldown rate, the decision to cool down is given as a part of the selection of the response appropriate for the diagnosis of SGTR.

the cognition-oriented aspects of response implementation when the actions are more complex than simple skill-of-the-craft.

Identification of failure mechanism and contextual factors

There was substantial evidence in the study that methods that focus on identifying failure mechanisms (ways the crews could fail a particular task) and the contextual factors that enable them (e.g., CBDT, ATHEANA, CESA, MERMOS) tended to produce richer content in the qualitative analysis than the PSF-focused methods (e.g., SPAR-H, ASEP, Enhanced Bayesian THERP, PANAME, and similar methods, such as CREAM and HEART). Moreover, the resulting operational stories reflected a more detailed prediction of what could or would occur in responding to the scenario. However, richer operational stories did not necessarily lead to HEPs that were more consistent with the empirical data, so other factors are also involved (e.g., reliable processes and associated guidance for translating the richer information into HEPs). Nevertheless, it seemed clear that, given the variety of conditions that can occur in an accident scenario, a thorough assessment of failure mechanisms and context will be needed for consistent and reasonable results. That is, the simpler methods do not appear to have the capacity to cover a broad enough range of conditions to consistently produce reliable results (but see the discussion of PSFs below). Moreover, for those methods that rely on the assessment of PSFs to estimate HEPs, considering the possible failure mechanisms or causes could provide a rationale for identifying the more important PSFs and their effects.

Judging the influence of PSFs and choosing the right PSFs

Not surprisingly, in the HRA analyses using PSFs (or similar, such as the common performance conditions used in CREAM and the error-producing conditions included in HEART), the evaluation of the degree of influence of the different PSFs considered by the method was an important factor. In both the loss of feedwater (LOFW) and SGTR scenarios, the assessment group identified inconsistencies in the ratings of the PSFs in those HRA methods highly based on PSFs. For example, although the present study only had one case where a single method was used by two different teams (SPAR-H), in a couple of cases the methods were similar (e.g., DT+ASEP and CBDT+THERP, along with ASEP and ASEP/THERP). Observable variations in the HEPs for the same HFEs both in the SGTR and the LOFW scenarios were seen across these methods, and differences were seen in both the selection and weighting of the PSFs thought to be important. Clearly, in many cases, these judgments can be difficult, and the results of some methods were very sensitive to these sometimes subtle judgments.

Two aspects of the analysis contributed to these inconsistencies. First, the HRA teams did not develop the same degree of qualitative understanding of the details of the scenario. Second, there were differences in the interpretation of the scope of the PSFs and in the ratings assigned to the PSF for a given issue or performance condition. In most of the HRA methods using PSFs (e.g., SPAR-H, ASEP, ASEP/THERP (THERP itself uses PSFs only to a limited extent), Enhanced Bayesian THERP, K-HRA, and PANAME), and other methods, such as HEART and CREAM, that require similar types of judgments regarding the task types and performance conditions, and CBDT and CESA, which require judgments on the levels of various conditions, the guidance provided to support these judgments is limited. Consequently, support for consistent transformation of qualitative insights into consistent inputs for quantification is needed. Of course, the failure cause-/context-based methods (e.g., ATHEANA, MERMOS) are not immune to this issue, but there is an emphasis in those methods on obtaining additional information to support the judgments. While there will always be some subjectivity involved, the study indicated that all of the methods need methodological improvements or improvements in the guidance related to judging which factors should be considered and how to evaluate and weight them (e.g., the level or strength of a factor or set of factors relative to an HFE).

Range of PSFs covered

Another PSF-related issue concerns whether an adequate range of PSFs is addressed by a given method. There was evidence in the study that in some cases the PSF-based methods (including CESA, CREAM, HEART, and the CBDT approach) did not capture some of the relevant influencing factors identified in the data simply because they were not addressed by the method. Some examples include: (1) for workload, the effect of concurrent tasks vs. the load from the tasks associated with the HFE; (2) for procedural guidance, the steps that guide crew assessment and response selection vs. focusing on the steps detailing the manipulations to be executed; or (3) for human-machine interface (HMI), the situation-specific salience of the cues given static or dynamic HMI interactions vs. the ergonomic quality of the interface. This finding suggests that to be able to reliably predict performance, HRA methods need to cover an appropriate range of PSFs.

While this seems to be a solid conclusion from the study, some methods (e.g., Enhanced Bayesian THERP) take the position that not all possible PSFs need to be included or evaluated exactly right to produce reasonable HEPs (in part because the time available is a key measure for this approach). CREAM seems to take a similar perspective by narrowing down to specific task types and using corresponding PSFs, but there were many misses in identifying important PSFs, as seen in the crew data, and misses in terms of the difficulty reflected in some of the HEPs. Nevertheless, the notion is simply that, with a few key factors, an adequate and reliable assessment of the likelihood of failure can be obtained in most cases. There was in fact some evidence that the PSF-based methods sometimes produced reasonable HEPs without identifying all relevant PSFs, particularly for the easy HFEs. However, whether this reflects an inherent characteristic of the method, or whether it was just a coincidental effect, could not be clearly determined. Similarly, it was true that other methods that attempt to address a wide range of contextual factors, such as ATHEANA and MERMOS, did not always obtain reasonable HEPs, even when identifying the correct set of factors; yet these methods often seemed to do better in the qualitative analysis, when the HFEs were relatively difficult. While the present study was not able to resolve this issue, it does seem that it would be a good question to address in future HRA empirical studies: that is, can methods using a key subset of factors, a corresponding qualitative analysis, and a dovetailing quantification process produce reliable and reasonable HEPs for most scenarios? Of course, a single method that adequately provides guidance for covering the full range of conditions in a relatively straightforward manner and consistently produces reasonable HEPs would be ideal.

Detailed guidance and analyst expertise

The study provides strong evidence that all HRA methods continue to involve significant expert judgment, and that the quality of the results can depend to a great extent on decisions about what to include in the analysis and how to include it, and decisions about the level or expected impact of PSFs (discussed further above). In some cases, analyst expertise was used to extend the methods and improve their overall performance. In general, this implies that improved methodology and better guidance for collecting and systematically using the right information is needed for many HRA methods to help reduce variability and support the expert judgment required in applying HRA methods. Furthermore, the study found discrepancies between method descriptions and their actual applications by the analysts. Some analysts interpreted the methods based on informal “consensus” practices not included in the method description. These and related findings point to the need for developing more structured guidance and tools to ensure more coherent and consistent method application and to lead less experienced analysts to search for and address appropriate information in obtaining HEPs, thus reducing variability. This need for improved guidance does not eliminate the need for an interdisciplinary HRA analysis team, with an understanding of the plant behavior and accident evolutions, plant operations, and the probabilistic safety analysis (PSA) model of the accident sequences.

Crew characteristics and representing crew variability

It appeared that crew factors, such as team dynamics, work processes, communication strategies, sense of urgency, and willingness to take knowledge-based actions can have significant effects on crew performance. The effects from these factors can be positive for some crews and negative for others within the same accident scenario, as their effects are moderated or reinforced by other crew characteristics and/or situational features. While such factors can certainly be worth investigating, it is often difficult in the context of the probabilistic risk analysis (PRA) to observe enough crews in the simulator and collect appropriate information to be able to identify systematic crew characteristics and evaluate their potential influence on the scenarios.

Moreover, crew-to-crew variability is not explicitly considered by many methods. Several methods (e.g., SPAR-H, ASEP, HEART, CBDT) consider a “representative” crew (a crew with characteristics judged to be average or typical) in a “base case” quantification. In contrast, methods that explicitly consider scenario variations (e.g., ATHEANA, MERMOS) can address crew-to-crew variability in estimating the HEP if they choose to, and ATHEANA does provide some guidance for doing so. In fact, this option can in principle be used with any other method, by developing different HEPs for different PSFs that reflect the impact of different crew characteristics and performing a weighted sum of the HEPs. Of course, a process for how to include the impact of the crew characteristics on HEPs would be needed. The human cognitive reliability/operator reliability experiments (HCR/ORE) method also accounts for crew-to-crew variability in performance through its variance (sigma) parameter; this variability is measured if plant-specific simulator data is used to obtain the parameters of the time/reliability correlation (TRC), or it can be selected from generic data obtained from similar scenarios in the ORE experiments. The appropriate scenario category for the sigma to use is selected by the analysts.

The question for HRA is to what degree these issues need to be taken into account, and how feasible it is to try and do so. As noted above, current HRA methods take them into account to varying degrees, and it can be difficult to obtain the needed information. Furthermore, the effects need to be systematic (e.g., half the crews do something one way while the other do it another) and have a significant impact on performance in order to warrant inclusion. Given the current state-of-the-art in HRA for treating these potential effects, they may often have to be evaluated using sensitivity analyses on the HRA results in order to see if the effects are important enough to investigate and explicitly incorporate into the analysis.

Ambiguity of HRA dependence guidance

In the empirical data from the LOFW scenarios, all of the crews that failed to implement bleed and feed (B&F) before dryout subsequently initiated B&F before core damage. This suggests that the actions at least are not completely dependent. If that was the case, all of the crews who failed to initiate B&F before dryout would have a negligible chance of success after dryout and before core damage. On the contrary, however, it seems that three things contributed to the success of the operating crews: (1) the crews had more time to analyze the situation; (2) the crews had additional cues, especially the flat steam generator level trend indications pointing to the indicators' unreliability; and (3) the situation after dryout was less complex to the crews because the concurrent goals and tasks of dealing with condensate pumps or feedwater pumps to feed the steam generator (SG) were no longer applicable. One should note that these HFEs comprise one example of dependency, while other HFEs in other scenarios might constitute different types of dependency.

Most of the HRA teams analyzed the conditional HFE and addressed potential dependence with a THERP-based dependence model, and obtained HEPs for the conditional HFEs that were pessimistic compared to the empirical data. In considering potential dependence, they were consistent with the common practice of accounting only for positive dependence, which refers in this case to the failure of a preceding task increasing the HEP of the subsequent task (relative to the case where the preceding task is successful). An analysis of the empirical data does not make it clear whether a "negative" dependence relationship between the preceding and conditional HFEs is applicable in this case. The factors leading to the failure of the preceding HEP did not necessarily reduce the failure probability for the conditional HEP. Instead, the plant context after failure apparently evolved to the point where the decision became simpler. These findings point out that it may be important to balance potential positive dependence effects from an initial failure with the impact of new information and changes in conditions. However, they also support the idea that, at least in some cases, it may be important to consider the potential for negative dependence, even when previous failures occur. Significant improvement in the treatment of dependence is needed for all methods. In particular, it would appear that analysts need to understand the dynamic nature of the plant status evolution and the information flow and procedural guidance that the evolution entails, rather than the current emphasis on factors like same crew, same procedure, or same location, which focus on more static aspects.

Traceability

Traceability is an important aspect of HRA. In this study, two different aspects of traceability were evident in the HRA analyses: (1) traceability of the quantification itself, given the choices made in the analysis; and (2) traceability of how the judgments from any qualitative analysis are reflected in the method's representation (e.g., basis for choices of PSFs and their weights). For some PSF-based methods, the first aspect of traceability may be good. For example, in SPAR-H, the simplicity of the base probabilities and the adjustments of the multipliers make the quantification itself very easily traceable, once the weights are decided. This aspect of traceability is strongly related to reproducibility in the sense that if the analysts make the same assumptions, they will get the same HEP. For the very same method, the second aspect of traceability may be not as good, since the way in which weights are decided from the qualitative analysis is not as easily traceable, and relies heavily on the documentation provided by the analysts.

For the context or variable scenario-based methods, such as, MERMOS or ATHEANA, this picture may be almost the opposite. These methods have established good approaches for identifying and translating qualitative analysis and judgments into understanding of the conditions facing the operators, and they develop strong operational stories as a basis for quantification. However, these methods lack an easily traceable way of translating these scenario stories into HEPs during the quantification process, and there is no guarantee of reproducibility, even when the analysts agree on the assumptions and aspects of the scenario descriptions.[6]

6.3 Specific recommendations for improving guidance, practice, and methods

It is worth emphasizing that the Empirical Study is based on a specific, limited set of HFEs. While this set of HFEs presents human performance and analysis issues that are expected to be broadly representative of the spectrum of issues that may be encountered in a PSA, the key observations do depend on generalizing on the basis of a small set of HFEs.

As also noted elsewhere in this report, a second caution is that the study design (essentially one analysis team per method, one method per analysis team, with rare exceptions) does not allow the findings to separate the effect of the analyst, sometimes also referred to as the “user effect,” from that of the HRA method. Nevertheless, the examination of the submitted HRA analyses did identify aspects of the methods that would be expected to be problematic for consistency and repeatability.

6.3.1 Address both qualitative and quantitative aspects of HRA

Overall, the assessments of the HRA analyses performed with the methods were generally able to identify (1) the more challenging HFEs and (2) a number of the main issues and factors contributing to unreliability. For the simpler HFEs that presented little or no challenge to the operator crews, the empirical evidence was usually not sufficient to either support or refute qualitative predictions because the identified human performance issues (if any) were at a low probability level. On the other hand, the difficult HFEs were empirically identified from the observation data, using observed crew performance difficulties or the crew performance measures based on plant parameters. When one or more crews failed to meet the HFE success criteria, the resulting empirical HEPs had narrow confidence bounds. Using the empirical data for these difficult HFEs, the qualitative as well as the quantitative predictions of the methods could be assessed. The assessments of qualitative predictions provide insights on the ability of the different methods to identify and represent situations and issues actually observed to contribute to crew difficulties, and to the potential or actual failure to meet the success criteria. The narrow confidence bounds of the HEPs for these HFEs provided information on whether the HRA method then yielded appropriately large HEPs.

Looking at the method assessments shows that, for some methods, it was difficult to address some of the empirically observed performance issues at all. For other methods, the performance issues could be identified and treated, but the HEPs obtained for the difficult HFEs were too small (below the lower confidence bound of the corresponding empirical HEP). These findings suggest that, in order to improve the predictive performance of HRA analyses, both the qualitative and quantitative aspects of HRA analyses need to be addressed.

Although the method assessments focused on the extent to which the empirical findings supported the predictions from each HRA team applying a given method, the process of summarizing their qualitative predictions and comparing these to the empirical evidence pointed to significant differences in the qualitative predictions among the analysis teams. Even before quantification, there were differences in the scope of the issues sought or factors

[6] In MERMOS, all scenarios have failure probability 1, so if "the analysts agree on the assumptions and aspects of the scenario descriptions" they will agree on the HEPs of the HFEs. However, different analysts would likely make different assumptions, that is, they would find different failure stories and would give different probabilities to the situational features that "cause" the stories. In this sense, it is similar to older methods where analysts weight and rate the PSFs differently.

treated, as well as in the level of detail at which they were examined. These differences in qualitative analysis findings and the qualitative predictive performance suggest that improvements to HRA methods and guidance should not focus solely on improved consistency in quantification, or on the quantitative calibration of the HRA methods. Improvements related to quantification alone would not lead to more consistent HEPs unless the quantification inputs were also consistent.

Instead, improvements to the qualitative analysis process and guidance are needed to ensure that this stage of the HRA properly identifies the critical tasks (the key tasks needed to succeed), characterizes the performance contexts in terms of the performance factors and operational challenges faced by the operators, and identifies those features of the task and scenario most likely to contribute to the failure of the HFE. Some of these improvements are discussed below, in 6.3.2.

Attention to the qualitative-quantitative interface

As noted, improvements in guidance and practice are needed in both the qualitative analysis steps and the quantification steps of an HRA analysis. Special attention is needed at the qualitative-quantitative interface; in other words, the transformation of the qualitative analysis findings into the quantitative analysis inputs. The findings of a qualitative analysis can be, or frequently are, expressed in terms of performance issues and challenges, such as the applicability of procedural guidance, difficulty in interpreting decision criteria (as written or as a function of plant parameter behavior), cues that are masked by a failure, and the coordination and timing requirements for manipulations. For many PSF-based methods, these findings then need to be expressed as PSF ratings.

Another aspect of the qualitative-quantitative interface is how to represent these qualitative findings in the model used to quantify the HFE. As discussed further below, this has to do with the extent to which HFEs are decomposed into subtasks and the level of modeling appropriate for the basic failure probabilities provided by a method.

To summarize, to improve the predictive power of HRA, additional guidance and enhancements may be needed for qualitative analysis as well as for quantification, with attention paid to ensuring a close coupling between these analysis stages. Some suggestions related to qualitative analysis in general are presented in the following section (6.3.2), while Section 6.3.3 discusses potential areas for improvement in method guidance.

6.3.2 Enhancing qualitative analysis

Significant differences were observed in the depth and level of detail within the qualitative analyses, when the qualitative findings are compared with each other, as well as in the degree to which the qualitative predictions were supported by the empirical data. With respect to the predictive performance, it was observed that the differences could not be explained (anecdotally) solely by the level of detail or the effort used in the qualitative analysis. The more detailed approaches certainly demonstrated a greater potential for identifying some issues, which in a simpler qualitative analysis may not be addressed at all; however, there were no indications of a more systematic relationship between detailed qualitative analysis and improved quantitative predictive performance.

Before discussing the areas of the qualitative analysis process in which enhancements could be useful, it is worth noting that a sound qualitative analysis is a premise of many HRA methods, starting with the THERP method. When HRA methods are compared over time, it can be seen that many of the recent method developments have indeed added guidance for the qualitative analysis process, whereas the documentation of older methods only refers to the qualitative analysis as a prerequisite step. The question of whether a method and its guidance actually supported the qualitative analysis as performed by an HRA team was discussed in the Empirical Study workshops (in which the assessment group interacted with the HRA teams). Some analysis teams pointed out that they applied what they considered known good practices for HRA qualitative analysis. It was also pointed out that some of the guidance added in more recent methods frequently represented and summarized such practices.

Qualitative analysis needs to be performed beyond the minimum requirements of most individual methods

Examining some of the HRA teams' qualitative analyses and their results shows that, without explicit guidance provided by an HRA method, some teams performed a qualitative analysis that consisted essentially of characterizing and rating the performance factors used in the method. For the various methods where the HFE is basically decomposed into an assessment/decision component and an execution/implementation component, this approach made it difficult to address the more complex HFEs, which had multiple subtasks and took place over a significant evolution of the scenario. In addition, such qualitative analyses focused on, and were limited to, the human performance issues addressed by a given method. For the set of HFEs in the study, the empirical observations highlighted issues for which it was difficult for the HRA teams to select an appropriate task type or PSF effect to represent the issue and its impact on performance. Given the limitations of individual HRA methods, it is important for HRA teams to strive for a qualitative HFE analysis that is broader than the scope of the factors explicitly treated by a given method. However, as is discussed below, this extended analysis may not be easily incorporated into the quantification model of a given method. In itself, an extended qualitative analysis may not provide improved quantitative predictions, unless the qualitative and quantitative elements of the HRA method are closely coupled.

Scales and anchored ratings for PSFs

For a number of the performance factors commonly used in the various HRA methods, the qualitative analyses showed that the set of issues or task/scenario features that the HRA analysts examined for a given PSF varied significantly. For instance, for the factor "Procedural Guidance," some analyses focused on the number and wording of key procedural steps, whereas others examined the applicability of the procedural guidance for the given scenario and the specificity of the instructions and decision criteria. Analogous differences may be found for such factors as "Human-Machine Interface," where some analyses focused on the ergonomics of the interface while others considered the availability of the relevant plant parameters and the saliency of the indications.

Such differences in the scopes of the PSFs addressed by the analysis teams combine with shortcomings in the PSF rating scales (when the qualitative findings are expressed as quantification model inputs) to yield the differences in the scope, depth, and findings of the qualitative analysis in HRA. These suggest that enhancing the rating scales for PSFs in terms of providing anchors for the ratings could promote consistency in the scope of factors addressed by different analysts. In addition, such scales should also support more consistency in the quantification model inputs (among analysts using the same method), leading to more reliable (repeatable) quantitative results.

Addressing the dynamics of operator actions and an operational perspective on task performance

The simplest, most basic HFEs are generally characterized by a single assessment/decision subtask and a single manipulation, or a short series of manipulations (e.g., to initiate a system). As expected, there were fewer differences in the qualitative analyses of such HFEs. Even if there were differences in the scope and depth of the analyses of the PSFs, most analysts predicted few performance issues (typically with a low likelihood) and yielded HFE probabilities at a level where the empirical data was inadequate as evidence either for or against these probabilities (failure probabilities on the order of 1E-2 and below).

In contrast, there tended to be significant variability in the HRA analyses of HFEs with multiple assessment/decision subtasks and execution subtasks, which took place over time and required observation and processing of plant parameters. These operator actions and the corresponding HFEs may be characterized as more complex. The successful completion of these actions could require the operator crew to perform a series of plant or equipment state assessments in order to reach the procedural steps guiding the execution of the task. In addition, the relevant procedural guidance and steps may be in different procedure steps, or in multiple procedures. Finally, in execution/implementation, the success of the action could require waiting for, collecting, and assessing plant feedback before continuing or completing the action.

For such HFEs, the narrative-based HRA methods, such as MERMOS and ATHEANA, appeared to present some advantages. Narrative-based methods usually do not try to address the HFE based on decomposition, instead giving attention to the unfolding of the scenario, to the evolving perspective of the operating crew, and to how the task is performed in operational terms, since these elements provide the basis for the context-based failure narratives on which their quantification is based. Other methods, such as CESA and CBDT, which examine how each HFE subtask may fail as a function of crew performance and the scenario evolution, also have some capacity to capture and model context-related (in contrast to task-based) failure mechanisms and their impacts on the HFE failure probability.

For HRA methods for which decomposition is fundamental to the quantification model, the qualitative analysis could be strengthened by giving more attention to identifying how the operator action (the set of tasks related to a given HFE) is embedded into the scenario operationally, including interactions with other ongoing tasks, and to identifying the specific assessment/decision subtasks and execution/implementation subtasks over time, before attempting to decompose the operator action into its assessment/decision component(s) and its execution components(s).

6.3.3 Improving guidance, practices, and methods

Some of the directions in which the qualitative analysis stage of an HRA may be improved have been discussed immediately above. The qualitative-quantitative interface has also been mentioned. This section focuses on potential improvements in the quantification aspects of HRA methods.

Addressing the overlap among PSFs in quantification

The role of the qualitative analysis is to identify key (sub)tasks required for success of the HFE, to characterize the performance contexts for these subtasks and operational challenges faced by the operators, and to identify those features of the task and scenario most likely to contribute to the failure of the HFE. The set of HRA analyses performed for the Empirical Study highlighted the fact that, in many methods, analysts may treat a given performance issue or challenging aspect of the context with different PSFs. This may occur because (1) the scope of the individual PSFs is ambiguously defined, or (2) because analysts interpret PSFs more broadly in order to represent issues not clearly addressed by the HRA method. Note that there may also be unavoidable interactions among the PSFs (e.g., “Scenario complexity” and “Procedural guidance”), although this tended to affect the empirical data analysis in terms of identifying the relevant driving factors of performance. The selection of a PSF to represent a given issue will usually have a quantitative impact, that is, lead to different HEPs based on the selection (e.g., because the multipliers associated with different PSFs are different).

One example of an overlap among PSFs in quantification relates to the lack of plant information readily available to help the crew to reach a plant state assessment. In some analyses, such issues were modeled in terms of “Human-Machine Interface” by considering the availability and presentation of plant information within the control room interface, whereas other analysts represented the same issue in terms of “Scenario Complexity” because the lack of information made the scenario more difficult to understand.

Comparisons among the HRA analyses using different methods were not performed in the Empirical Study to address this type of issue specifically. However, such observations suggest that guidance can be developed by having multiple teams of analysts modeling complex HFEs with a given method.

Differences in modeling for quantification

Modeling for quantification refers to the selection and composition of building blocks (decomposition into basic model elements) in order to quantify an HFE. For a given HRA method, the design of the Empirical Study precluded identifying how different analysts may decompose differently, since each method was only applied by one analysis team, except for in one case. However, the quantification of HFEs was examined, for instance, when the failure probability estimates were a poor match to the empirical reference probability bounds. Such examinations, and the workshop discussions with the analysis team, showed that the analysis team in some cases considered different decomposition models to quantify the HFE (or different PSFs to represent a given performance issue), and that these different models yield, not unexpectedly, different HFE failure probabilities.

Consequently, one of the observations is that some methods need to provide additional information to specify the level of detail at which their basic elements (quantification model building blocks) are intended to be used. It would also be helpful if the method documentation would provide sample quantifications of more complex HFEs.

Reasonableness checking of HFE probabilities

In the Empirical Study, the HFE probabilities resulting from the application of each HRA method were assessed with respect to their relative values (rank order of HFEs by failure probabilities) and the overall differentiation among these probabilities. For some of the HRA analyses, the sets of failure probabilities obtained show limited differentiation. In some cases, the failure probabilities obtained by the HRA team fall into a narrow range (e.g., the HEPs are within a factor of 2 or 3 of each other), in contrast to the qualitative findings of the team, which range from an expectation of no significant performance challenges to the identification of multiple potential error mechanisms.

This suggests that the HRA team did not check the reasonableness of the obtained probabilities, or performed an inadequate check. Several analysis teams confirmed that their reasonableness check was limited, and possibly more limited than what would be performed in the context of an actual PSA study. The scenario, procedural, and task information provided to the analysis teams in the Empirical Study was extensive, and included procedural guidance, the expected scenario and parameter evolutions over time, and other details concerning the hardware failures underlying the complex scenarios. This may have caused some analysis teams to focus on understanding and integrating the implications and impacts of these details, and on documenting their qualitative assessments, perhaps at the expense of reviewing the quantitative results for overall coherence.

Although HEPs are often checked for reasonableness in external reviews of the PSAs/HRAs, there appears to be little documented guidance on how to perform reasonableness checks when each individual HEP cannot be reviewed in detail and emphasis is placed on the relative values of the HEPs. A reasonableness check examines whether HEPs of comparable magnitudes are obtained for “similar” HFEs, and whether the HEPs estimated for HFEs with more challenging performance conditions are indeed larger. Although they inherently involve multiple dimensions, some of the factors to be considered in terms of similarity and levels of challenge include:

- time available (time window)
- decision complexity, basic vs. complex scenarios (number of issues, need to prioritize)
- task complexity, number of tasks, need for manual control, fine-tuning, adjustment
- number of issues, adverse PSFs, and failure modes identified for the HFE

Comparing related HFEs (for the same tasks in different scenarios) or HFEs with similar performance conditions, as represented by these factors, typically leads analysts to review the contributions to the HEPs to determine whether they correspond to their expectations, based on their qualitative analysis. For the HRA teams that did perform such checks, the identified discrepancies between HEP results and qualitative expectations would lead them to review the quantification, and, in some cases, to adjust the quantification of the HFEs.

In summary, the development of guidance for reasonableness checks would help to promote a structured review of HRA results that emphasizes the consistency between qualitative findings and quantification results.

6.3.4 Towards hybrid HRA methods

The recommendations of the previous sections have focused on enhancing analysis practices and potential improvements for each of the HRA methods. Potential improvements to the individual methods can be identified, and, in many cases, are being addressed by the developers of the individual HRA methods; however, the Empirical Study results, taken as a whole, also provide some indications that support combining the effective elements and features of the different HRA methods as a way forward, that is, integrating these elements into a hybrid method.

Some of the findings that suggest hybrid methods include:

- Within the set of HFEs and the various human performance issues relevant to these HFEs, no method consistently outperformed the other methods.
- The methods did not perform equally well on the SGTR and LOFW HFEs. Overall, some analysis teams performed better in the later LOFW phase, which could be an effect of the analysts learning to analyze the second set better in terms of the measures used within the Empirical Study. However, the predictive performance for the LOFW HFEs was poorer for other teams, suggesting that some methods may be better at treating some kinds of HFEs than others.
- There were fairly clear differences in the methods' ability to model assessment/decision and implementation/execution, respectively. The structured methods tended to handle the latter better, while narrative-based methods had advantages for assessment/decision issues.
- Most methods could be used to model simple HFEs, but the simpler methods, with fewer degrees of freedom for structuring the quantification model, were difficult to apply to the complex, multiple-subtask HFEs in the sense that they couldn't adequately cover the situation.
- For each HFE, no method showed consistently conservative or optimistic tendencies relative to all methods' predictions for that HFE. This suggests that none of the methods are particularly suited for "scoping" vs. "detailed" analyses.
- Similarly, simple methods were not obviously conservative or more conservative. This suggests that the more detailed methods are not trading off simplicity to obtain more realistic failure probabilities, and cannot be viewed as requiring more effort to obtain more realistic, less conservative HEPs.

The aim of a hybrid method would be to guide a richer qualitative analysis, providing a broad scope of performance factors and failure modes and mechanisms, while maintaining the repeatability of the methods with more structured quantification. The narrative-based methods, such as MERMOS and ATHEANA, are broader and more flexible with respect to the scope of the factors and failure mechanisms. While such methods circumvent the need to use a specific quantification model (assessment/decision + execution or application of model "building blocks"), the quantification at this stage relies strongly on expert judgment. With no basic HEPs and few guideline values, their quantitative performance was modest, reinforcing the need for a more structured or at least a better anchored quantification process. (Separately, given the constraints of the Empirical Study, these methods' use of or reliance on detailed plant-specific information on crew behaviors and situation-specific response strategies may also have played a role.)

In summary, the Empirical Study has highlighted some of the overall requirements for HRA methods, especially those related to guidance and supporting consistent analysis practices, and its method assessments identify features of the various HRA methods that a hybrid method could incorporate.

7. CONCLUSIONS ON THE USE OF EMPIRICAL HRA DATA AND BENCHMARKING

7.1 Feasibility of benchmarking against simulator data

The human reliability analysis (HRA) Empirical Study is designed around a simulator study with up to 14 licensed operator crews from two units of the participating nuclear power plant. In the steam generator tube rupture (SGTR) phase, 14 observations of both scenarios were made, while 10 observations were obtained for both of the loss of feedwater (LOFW) scenarios. While this makes the simulator study remarkably large, the sample size for deriving reference human error probabilities (HEPs) from the evidence remains small. Quantitatively, the data represent a mixture of strong and weak quantitative evidence. When a failure is observed in a small sample, the evidence is strong; on the other hand, when no failures are observed, the evidence for the HEP is weak. Specifically, this means that the simulator data does not provide strong quantitative evidence for HEPs much lower than 0.05 to 0.1; in other words, for human failure events (HFEs) where the performance of the crews easily meets the HFE success criteria, the quantitative observations concerning the number of crews not meeting the success criteria would be “consistent” with a range of several orders of magnitude for predicted HEPs.

In the study, these limitations on the purely quantitative aspects of the reference data were addressed with (1) a combination of benchmarking methodology features that accounted for qualitative evidence as well as a qualitative ranking of the HFEs, and (2) a selection of scenarios and HFEs that included HFEs representing a range of difficulties, and included very adverse, challenging scenario contexts.

The results of the study show that the predictive performance of the various HRA methods could be evaluated.[7] Thus, reference data for the benchmarking of the HRA methods can be obtained from a simulator study based on a relatively small number of observations of each HFE.

The benchmarking methodology features and scenario/HFE selection are further discussed below, in Section 7.3.

7.2 Usefulness and acceptance of results

The study addressed the predictive performance of HRA methods both quantitatively and qualitatively, rather than the consistency or convergence of HEP estimates from different HRA methods. The HRA Empirical Study was able to identify or cast further light on the strengths and weaknesses of the HRA methods with respect to their predictive performance.

For the quantitative aspects, these strengths and weaknesses relate to, for instance, the underestimation of challenging HEPs (in these cases, the empirical evidence is strong), to the ranking of HFEs, and to the differentiation among HFEs. With respect to the qualitative aspects, the strengths and weaknesses identified relate to the identification of the performance factors that increase the likelihood of task failure and the operational manifestation of these factors (how these factors specifically affect the response of the crews).

[7] As noted, the method assessments have focused on the performance of each method vs. the empirical data, and on the related identification of each method’s strengths and weaknesses. At the same time, there are some indications for the relative performance of the methods. However, the differences in method performance between the SGTR and LOFW HFEs suggest that conclusions should not be drawn on the basis of method-to-method comparisons; predictive performance for HFEs in other PSA-relevant scenarios could be expected to differ.

Overall, the results consist of findings and insights on the adequacy of the set of performance factors addressed by a method, on the appropriateness of the scope of the performance factors, on the quantitative relationship between the identified performance factors and HEP estimates, and on how analysts interpret and apply a given method. These results may serve as an impetus for the developers of HRA methods and associated guidance to extend a given method or its guidance. In this regard, the focus on assessing the prediction of actual task performance in simulated accident scenarios is especially useful because it deals with representing and modelling concrete, observed crew behaviours and responses.

For HRA practitioners, the results of the Empirical Study and the documented analyses that have been performed provide insights concerning whether, how well, and the ease with which specific HRA methods treat some scenario challenges and crew behaviours that are generally relevant to a range of HFEs in probabilistic safety analysis (PSA) scenarios. While specific to the type of plant and procedures, the documented crew responses to these scenarios highlight for practitioners the variability in performance, including successful performance, and the operational challenges that crews face in emergency scenarios.

On the other hand, some limitations and cautions on the generalizability of the findings are warranted. One limitation of the study design is that each HRA method (with one exception) was only applied by a single analysis team. As such, it is difficult to generalize the comparison findings to the method in general. In the worst case, the comparison findings may reflect the peculiarities of one analysis team and not prove representative of other applications of the same method. Since there is no way to gauge inter-analyst and intra-method variability, given the makeup of the analysis teams in the present study, future HRA benchmarking efforts should attempt to provide more than one analysis team per method. It is not felt, however, that this limitation hindered successful insights into the methods in this study. There were considerable lessons learned about the methods in terms of their utility for qualitative and quantitative predictions. Additional insights into the process were documented formally and anecdotally by the analysis teams, allowing the study's assessment team to provide informed discussions on the strengths and weaknesses of the individual methods with respect to their use in the International HRA Empirical Study. Further generalization was not attempted or warranted.

Just as one must be cautious not to generalize the results of one team's analysis to an entire HRA method, one must take care to consider the specific scenarios that are being analyzed. The present study provided two scenarios—an SGTR and an LOFW scenario, along with base and complex case variants—to allow a fairly diverse sample of the types of analyses for which the HRA methods might be used. HRA methods were designed for different purposes, and no single scenario is sufficient to comprehensively gauge the merits or limitations of a particular method. There were differences in the predictive performance of the methods between the SGTR and LOFW scenarios; with other scenarios and HFEs, some further differences in the performance of the methods could be expected. Additional scenarios to broaden the gamut of HRA activities are a logical extension of the current study.

7.3 Assessment of benchmarking methodology features

This section discusses some of the key features of the benchmarking methodology used to assess HRA methods in the Empirical Study.

The Empirical Study is based on the assessment of the HRA analysis predictions against reference data from a simulator study, rather than on assessing convergence between the methods. (Feature 1)

The reference data derived from a simulator study provided a rich set of qualitative and quantitative data for assessing the predictive performance of the methods. The availability of the observations on which the reference data are based (e.g., descriptions and data concerning the response of each crew) added to the transparency of the reference data and helped to resolve differences in the terms used in different methods (e.g., for performance factors), and in the interpretation of these terms by different analysts.

Two challenges that need to be addressed when performing such a study arose. These are to some extent inherent to empirical HRA data from simulations, and were not deemed problematic for the results. First, the HFE success criteria may need to differ from those used in a PSA setting. For instance, the available time (time window) for the completion of a task needs to be addressed with care because crews can take actions that affect the plant response, and, therefore, the effective time window. Secondly, a simulated scenario represents a specific realization of a PSA scenario, corresponding to one way in which a functional failure (system unavailability) may arise. The specificity of the simulated realization (e.g., the cause of unavailability of a system or its manifestation) must be taken into account, since the observed crew performance underlying the empirically-based reference data is only directly applicable to this realization. In contrast, the HRA of an HFE in a PSA typically does not distinguish among the specific causes of a functional failure, which may affect crew performance.

Qualitative predictions were assessed in addition to the quantitative predictions. (Feature 2)

Analysis predictions and empirical simulator data regarding the most important factors driving crew performance and the associated impact on the crew response in operational terms were used in the benchmark. In addition, the empirical qualitative findings from the simulators were used to supplement the observed failure counts in order to derive a relative ranking of the HFEs as a reference.

To establish the empirical reference data for the assessments, several of these qualitative elements required subjective judgments from the study team; the qualitative empirical data were then obtained in a consensus process. With the simulator data and observations available, there were no significant differences within the study team, and the HRA teams did not raise issues concerning this data. The subjective judgments underlying some of the qualitative empirical data were not problematic for the assessments of the methods.

It should be noted that some of the performance factors used in various predictive HRA methods were problematic in the simulator study. Whereas the crew observations could be used straightforwardly to determine both the presence of procedural guidance issues (applicability in the scenario, clarity of the wording) and their impact on crew performance, the relationship between an adverse performance factor and its impact on the crew responses could not be established for all performance factors. The impact of such factors as "time pressure" and "experience" are frequently not directly observable, even though there is evidence that the factor is adverse. The validity of evaluating such factors in estimating HEPs is not questioned. Instead, the issue is that it can be difficult in this type of simulator study, where the possibility of controlling for these factors is limited, to determine from the empirical evidence whether and how the presence of the adverse factors affected crew performance.

The selection of scenarios/HFEs for the benchmark aimed to include similar tasks in more adverse (complex) scenarios, as well as in more basic, straightforward scenarios. (Feature 3)

The inclusion of HFEs that represent a range of difficulties and similar/related tasks (HFEs) in scenarios ranging from basic to more adverse scenarios was an essential element of the study. Challenging HFEs are needed to determine whether the treatment of performance factors by an HRA method may lead to underestimation of HFEs in adverse scenarios. Many of the tasks in the straightforward scenarios were performed successfully by the crews, with no notable factors or issues. For these HFEs, the simulator study provided little qualitative or quantitative data to assess the predictions of the HRA methods (unless these methods had unexpectedly high HEPs for these HFEs). On the other hand, including HFEs in both adverse and straightforward scenarios was useful to establish a baseline performance.

This study used two variants each of the SGTR and LOFW scenarios, for a total of four scenarios and 13 HFEs. These scenarios and HFEs were selected to include HFEs that were expected to significantly challenge the crews, especially in terms of cognitive, decision making, and plant state assessment. The scenarios and HFEs selected for the benchmark fulfilled the objective of having HFEs with a range of difficulties and including comparable tasks differentiated by the scenario context (adverse vs. straightforward).

The benchmark addressed the performance factors, as well as the associated operational issues and challenges. (Feature 4)

Addressing the operational issues and challenges related to the performance factors (i.e., underlying the rating of the performance factors) was very effective in the study. First, it provided a transparent basis for the performance factors, in light of the fact that the definitions of the performance factors (and their scope) varied among the methods. Secondly, it provided more specific information on how the HRA teams understood the HFEs and the expected performance of the crews. For instance, with respect to procedural guidance, the assessments could take into account whether the analysis teams had identified the specific procedure steps and the aspects of these steps that would lead to potential problems for the crews. Addressing the operational issues underlying performance factor ratings ensured an evidence-based assessment of the qualitative predictions.

7.4 Future HRA studies utilizing simulator data

As noted in the sections above, utilizing simulator data for these kinds of studies was shown to be highly useful. In this study, the simulator data was used as the empirical basis against which the results from the HRA predictions were compared. A benefit of this type of study was that all HRA teams were on a neutral playground, for example, all had access to same information; as a result the study could focus more on why the results were what they were, rather than on other aspects, for example on validating the underlying theoretical models.

In general, the promising results from this study encourage and promote the use of simulator data in the future for HRA in many different ways. It illustrated the potential of using and aggregating empirical simulator results from multiple studies to strengthen the empirical basis for both method assessment and extending the scope of methods to address some of the identified shortcomings. A follow-up study performed in a U.S. nuclear power plant (NPP) training simulator ([23] and [24]), utilized all methodological facets developed in this study including the types of scenarios used.

As a next step, simulator results could be utilized as part of a database. In this regard, there are several ongoing initiatives dealing with the use of simulator data to support HRA (e.g., see [26]). One may think of different usages, but the main idea is to collect simulator data in databases[8] in a way that can improve HRA. Several issues would have to be addressed, such as the definitions and frames for the data, how to collect the data, how to use the database and how to inform the HRA analysts in practical terms. While there could be other sources of HRA data, this study reinforced the relevance of simulator data for HRA in general.

[8] The word «database» is used in a rather broad manner in this context, applying to both pure narrative information and more structured numbers in a strictly defined framework.

8. ACHIEVEMENTS AND OVERALL CONCLUSIONS

The International human reliability analysis (HRA) Empirical Study is the first major study to directly compare HRA predictions with actual operating crew performance in probabilistic risk analysis (PRA)-related accident scenarios conducted in a full-scope nuclear power plant (NPP) simulator, the Halden Machine-Human Laboratory (HAMMLAB). Related studies within the nuclear field have only compared HRA methods to each other, while studies comparing HRA methods to empirical data have been limited, and were performed with data from domains other than NPP control rooms [4] and [22] for more detail). Using a manageable number of operating crews and simulated accident scenarios, the present study produced a large set of diverse findings on the different HRA methods and their application. The findings allowed both qualitative and quantitative HRA issues to be explored, and, although additional work will bring more evidence, they provided a strong empirical basis with which to evaluate the strengths and weaknesses of the various methods and HRA in general, and to identify methodological enhancements needed to improve the reliability and accuracy of HRA methods.

One of the main conclusions from the study is the importance of the qualitative analysis when performing HRA. It was shown that without a qualitative analysis that covers a thorough set of scenario conditions and influencing factors, one that is concerned with potential failure mechanisms, failure modes, and associated causes, and one that carefully examines the crews' interactions with the procedures, given the scenario conditions and available information, HRA methods will have an inadequate basis to identify important performance drivers and obtain realistic human error probability (HEP) estimates. Of course, it was also shown that without a reliable, structured process and associated guidance for translating the results of such a good qualitative analysis into HEPs, the benefits may not be realized. The various methods vary significantly in the nature and degree of the qualitative analysis required. While a good qualitative analysis (including a task analysis) is a relative strength of some methods, it is clear that all of the methods need improvement in this area. This conclusion is apparent from the findings discussed above in Sections 5 and 6 and summarized below.

Key insights regarding HRA and HRA methods included the following:

- Due to the dynamic nature of even the more straightforward accident scenarios, operating crews' cognitive activities while working through the procedures should always be considered in modeling and predicting crew behavior. That is, procedure-following during accident scenarios should not necessarily be assumed to be a simple response execution process with limited (or no) need for cognitive activities, such as plant/equipment state assessment, prioritization, and option selection, as is allowed by some methods.

- Methods that focus on identifying failure mechanisms (ways the crews could fail a particular task) and the contextual factors that enable them will generally produce richer content in the qualitative analysis than the performance-shaping factor (PSF)-focused methods, and should provide a better basis for estimating HEPs, as well as promoting the reduction of human errors. However, to benefit from the improved basis, a systematic and repeatable means for transforming the information into HEPs is needed.

- Selecting the most important PSFs and judging their degree of influence is sometimes difficult. The study indicated that all of the methods need methodological improvements or improvements in the guidance related to judging which factors should be considered and how to evaluate and weight them (e.g., the level or strength of a factor or set of factors relative to an HFE).

- The range or scope of PSFs covered by many existing methods may not always be adequate for reliable and accurate HRA. Some examples include (1) for workload, the effect of concurrent tasks vs. the load from the tasks associated with the HFE; (2) for procedural guidance, the steps that guide crew assessment and response selection vs. focusing on the steps detailing the manipulations to be executed; or (3) for human-machine interface (HMI), the situation-specific salience of the cues, given static or dynamic HMI interactions, vs. the ergonomic quality of the interface.

- In several instances (particularly in the SGTR scenarios), the HEPs produced by the HRA methods were not sensitive to the difficulties of the HFEs identified in the HRA team's qualitative analysis. Lack of sensitivity indicates that these methods may not provide an adequate range of influencing factors, allow an appropriate assessment of the PSFs, or appropriately evaluate diagnosis activities. The lack of appropriate differentiation points out that reasonableness checks on obtained HEPs should always be a key practice in HRA. Reasonableness checks (e.g., checking the rank order of the human failure events (HFEs) by their HEPs) were not always performed by the analysts in this study. While such checks will be subjective to some degree, the qualitative analysis should provide a basis to differentiate between HFEs so that examination of "their level of difficulty" produces a reasonable rank of HEPs. Such a reasonableness check helps analysts to develop insights about their results and identify cases where HFEs and their HEPs may need additional examination (e.g., see the NRC good practices document, NUREG-1792 [5]).

- In spite of the apparent limitations of the various methods, there was no evidence that any of the methods showed a systematic bias towards producing either conservative or optimistic HEPs, given an appropriate analysis.

- However, if the methods do not adequately cover the important qualitative information and do not have a reliable means of translating that information into HEPs, one would generally expect HRA to underestimate potential problems and HEPs, which obviously would not be desirable.

As is reflected in the bullets above, it should be noted that the quantitative empirical data and comparisons obtained in the study greatly supplemented the qualitative comparisons and insights identified. This was the case, even though there was a relatively small set of observations (at least from a statistical perspective), in terms of numbers of crews and scenarios. Generally, the quantitative empirical data and comparisons gave a very good starting point for assessing the qualitative predictions of the methods by prioritizing these qualitative findings and providing a measure of the significance of the predicted or observed performance issues. Thus, the present study demonstrated that important information on HRAs and HRA methods can be obtained without using impractically large numbers of operating crews and scenarios, which is an important achievement.

While the findings from this study have provided significant information on how to enhance HRA and improve results, additional studies will help to substantiate the results and address the generalizability of the findings. One limitation of the present study was that the experimental design made it difficult to always separate method from analyst effects. At a minimum, multiple HRA teams using the same method will be needed to assess the reliability of the results from the different methods and allow inferences about those aspects of the different methods' qualitative analyses that lead to shortcomings in their predictive validity; indeed, a new study ([23] and [24]) is being conducted by many of the participants from the International study. This study, referred to as the U.S. HRA Empirical Study, uses operating crew data collected in a U.S. NPP simulator, and is testing the consistency of the qualitative and quantitative HRA predictions between HRA teams using a given method. In other words, one of the aims of the U.S. HRA Empirical Study is to examine the "user effect," that is, the impact of the analysis team on identifying the performance issues in the scenarios and the use of this information to produce HEPs that reflect the relative difficulty of the HFEs, as derived from the empirical data. The U.S. study also addresses other concerns with methodological aspects of the present study.

As discussed in Section 7, another major benefit of the International study was the lessons learned about conducting HRA studies using nuclear power plant simulators which are being used in follow-up studies, such as the U.S. HRA Empirical Study ([22] and [24]). In particular, methodological tools such as (1) the development of the experimental design, focusing on evaluating HRA methods, (2) the methodology for collecting crew data, and (3) the methodology for data-to-method comparisons are proving to be very useful achievements from the study.

The experimental design of the study was tailored to HRA needs, which established the use and usefulness of simulators in HRA. This included designing scenarios similar to those modelled in probabilistic risk analysis (PRA) and analyzing the crew performance data in terms of tasks corresponding to PRA-type HFEs and in terms of the corresponding HFE boundary conditions, objective performance measures, and success criteria.

Furthermore, in its treatment of PSFs for the analysis of the simulator data and the assessment of predictions, the empirical study addressed the operational aspects of the PSFs. The PSF ratings were tied to the specific features or issues that would impact the crew performance and related to the affected HFE subtasks, both decision-related subtasks and manipulations. The data analysis also examined the connections between the PSFs and associated context features and issues and the observed performances and crew outcomes. The study also provided valuable empirical evidence on how crews will perform and why, and documented the variability in actual crew performance. These findings illustrated that crew performance variability, which is not typically directly considered in HRA, can be important under some circumstances. Whether it can be reasonably addressed in HRA modelling has yet to be determined (see Section 6 for additional discussion).

Finally, in addition to the findings on HRA methods, which are being used to enhance the reliability and accuracy of HRA methods (e.g., the "hybrid" HRA method being developed by the U.S. Nuclear Regulatory Commission (NRC) and the Electric Power Research Institute [25]) and the important lessons learned about conducting HRA studies using simulators, it should be noted that such studies also provide significant value to the participating plants. Based on the feedback given to Halden by the utility that supported the simulator runs, it appears that such studies and interactions with utilities may result in a number of performance benefits, including improvement to plant procedures and training programs. The U.S. NPP supporting the U.S. HRA Empirical Study [23] and [24] has provided similar feedback, particularly in terms of identified improvements for their training program.

9. REFERENCES

[1] Lois, E., Dang, V.N., Forester, J., Broberg, H., Massaiu, S., Hildebrandt, M., Braarud, P.Ø., Parry, G., Julius, J., Boring, R., Männistö, I., & Bye, A. (2009). International HRA Empirical Study - Phase 1 Report: Description of Overall Approach and Pilot Phase Results from Comparing HRA Methods to Simulator Performance Data. NUREG/IA-0216, Vol. 1, U.S. NRC. HWR-844, OECD Halden Reactor Project.

[2] Bye, A., Lois, E., Dang, V.N., Parry, G., Forester, J., Massaiu, S., Boring, R., Braarud, P.Ø., Broberg, H., Julius, J., Männistö, I., & Nelson, P. (2010). The International HRA Empirical Study - Phase 2 Report: Results from Comparing HRA Methods Predictions to HAMMLAB Simulator Data on SGTR Scenarios. NUREG/IA-0216, Vol. 2, U.S. NRC. HWR-915, OECD Halden Reactor Project.

[3] Dang, V.N., Forester, J., Boring, R., Broberg, H., Massaiu, S., Julius, J., Männistö, I., Liao, H., Nelson, P., Lois, E., & Bye, A. (2011). The International HRA Empirical Study - Phase 3 Report: Results from Comparing HRA Methods Predictions to HAMMLAB Simulator Data on LOFW Scenarios. NUREG/IA-0216, Vol. 3, U.S. NRC. HWR-951, OECD Halden Reactor Project.

[4] Boring, R. L., Hendrickson, S. M. L., Forester, J., Tran, T. Q., & Lois, E. (2010). Issues in benchmarking human reliability analysis methods: A literature review. Reliability Engineering & System Safety 95: 591-605.

[5] Kolaczkowski, A., Forester, J., Lois, E., Cooper, S. (2005). Good Practices for Implementing Human Reliability Analysis (HRA), NUREG-1792. US Nuclear Regulatory Commission, Washington, DC.

[6] Brown, L.D., Cai, T.T., & DasGupta, A. (2001). Interval Estimation for a Binomial Proportion. Statistical Science 16(2): 101-133.

[7] Siu, N.O., & Kelly D.L. (1998). Bayesian parameter estimation in probabilistic risk assessment. Reliability Engineering & System Safety 62: 89-116.

[8] Swain, A.D. (1987). Accident Sequence Evaluation Program Human Reliability Analysis Procedure. NUREG/CR-4772/SAND86-1996, Sandia National Laboratories for the U.S. Nuclear Regulatory Commission, Washington, DC, February 1987.

[9] Swain, A. D & Guttman, H. E. (1983). Handbook of human reliability analysis with emphasis on nuclear power plant applications. NUREG/CR-1278-F, U.S. Nuclear Regulatory Commission.

[10] Technical Basis and Implementation Guidelines for A Technique for Human Event Analysis (ATHEANA), NUREG-1624, Rev. 1, USNRC, Washington, D.C., May 2000.

[11] Forester, J., Kolaczkowski, A., Cooper, S., Bley, D., and Lois, E. (2007). ATHEANA User's Guide, NUREG-1880. US Nuclear Regulatory Commission, Washington, DC.

[12] Julius, J., Grobbelaar, J., Spiegel, D., & Rahn, F. (2005). The EPRI HRA Calculator® User’s Manual, Version 3.0, Product ID #1008238, Electric Power Research Institute, Palo Alto, CA.

[13] Parry, G. et al. (1992). An Approach to the Analysis of Operator Actions in PRA, EPRI TR-100259, Electric Power Research Institute, Palo Alto, CA.

[14] Reer, B. (2009). Outline of a Method for Quantifying Errors of Commission. LEA 09-302, Laboratory for Energy Systems Analysis, Paul Scherrer Institute, Villigen PSI, Switzerland.

[15] Hollnagel, E. (1998). Cognitive reliability and error analysis method CREAM. Elsevier Science Ltd.

[16] Williams, J. C. (1986) HEART–A proposed method for assessing and reducing human error. Proceedings of the 9th Advances in Reliability Technology Symposium, University of Bradford, UK, 2-4 April, 1986, pp B3/R/1-B/3/R/13.

[17] Kirwan B, Gibson H, Kennedy R, Edmunds J, Cooksley G, & Umbers, I. (2004), Nuclear Action Reliability Assessment (NARA): A Data-Based HRA Tool. Proc. 7th Int. Conf. on Probabilistic Safety Assessment and Management (PSAM 7 – ESREL '04), Berlin, Springer-Verlag.

[18] Wondea Jung, et. al., A Standard HRA Method foe PSA in Nuclear Power Plant; K-HRA Method, KAERI/TR-2961/2005,2005

[19] Bieder, C., Le Bot, P., Desmares, E., Cara, F., & Bonnet, J. L. (1998). MERMOS: EDF's New Advanced HRA Method. Probabilistic Safety Assessment and Management, PSAM4, New York, USA, September 13-18, 1998.

[20] Le Bot, P., Bieder, C., & Cara F. (1999). MERMOS, a second generation HRA method: what it does and doesn't do. International Topical Meeting on Probabilistic Safety Assessment, PSA'99, Washington, D.C., USA, August 22–26, 1999.

[21] Gertman, D., Blackman, H., Marble, J., Byers, J., Haney, L., & Smith, C. (2005): "The SPAR-H Human Reliability Analysis Method," NUREG/CR-6883. Washington, D.C., U.S. Nuclear Regulatory Commission.

[22] Boring, R. L., Forester, J., Bye, A., Dang, V. N., & Lois, E. (2010). Lessons Learned on Benchmarking from the International Human Reliability Analysis Empirical Study. Proceedings of the 10th International Probabilistic Safety Assessment and Management Conference, Seattle, Washington, USA.Bye, A., Dang, V. N., Forester, J., Hildebrandt, M., Marble, J., & Liao, H., & Lois, E. (2012). Overview and First Results of the US Empirical HRA Study. Proceedings of the 11th International Probabilistic Safety Assessment and Management Conference, Helsinki, Finland.

[23] Bye, A., Dang, V. N., Forester, J., Hildebrandt, M., Marble, J., & Liao, H., & Lois, E. (2012). Overview and First Results of the US Empirical HRA Study. Proceedings of the 11th International Probabilistic Safety Assessment and Management Conference, Helsinki, Finland

[24] Marble, J., Liao, H., Forester, J., Bye, A., Dang, V. N., Presley, M., & Lois, E. (2012). Results and Insights Derived from the Intra-Method Comparisons of the US HRA Empirical Study. Proceedings of the 11th International Probabilistic Safety Assessment and Management Conference, Helsinki, Finland.

[25] Hendrickson, S. M. L., Parry, G., Forester, J., Dang, V. N., Whaley, A., Lewis, S., Lois, E., & Xing, J. (2012). Towards an Improved HRA Method. Proceedings of the 11th International Probabilistic Safety Assessment and Management Conference, June 25-29, 2012, Helsinki, Finland.

[26] Chang, Y. J., Mosleh, A., Roth, E., Richards, R., Shen, S-H, Bley, D., & Kirwan, B. (2012). Methodology for collection and analysis of simulator data for HRA applications. PSAM11/ERREL2012 Conference, Helsinki, Finland, June 25-29, 2012

APPENDIX A: DETAILED SCENARIO DESCRIPTIONS AND HFE DEFINITIONS

A.1 SGTR base scenario

In this scenario, a steam generator tube rupture (SGTR) is initiated in steam generator (SG) #1 that is sufficient to cause nearly immediate secondary radiation alarms and other abnormal indications/alarms, such as SG #1 abnormal level and lowering pressurizer. Conditions, while continually degrading, are not sufficient to cause an immediate automatic scram. About three minutes after the tube rupture initiation, the large screen display indicates lowering pressurizer pressure and level, increased charging flow (as it attempts to make up for the loss of reactor coolant from the tube break), increasing SG #1 level, and a slight imbalance in feedwater flow to the SGs. If the crew also calls up the radiation monitoring display screen, they will see higher radiation indications associated with SG #1. It is expected that at this point, or as conditions continue to deteriorate over the next few minutes, the crew is likely to manually scram the plant. Even if they do not, an automatic scram will eventually occur, due to low pressurizer pressure or some other trip setting. Either way (manual or auto scram), the crew is expected to then enter the E-0 procedure.

Typically at about 10 minutes after entering E-0 (if the crew has not been delayed in their responses to the steps in the E-0 procedure), the crew should be reaching step 19, the first step in E-0 at which the crew should transfer to procedure E-3 (the SGTR procedure) in response to the radiation indications of an SGTR. At this point, secondary radiation is high (has been virtually from the beginning) and SG #1 level becomes elevated as compared to the other SGs once the level indications are restored following the scram, but it takes a while longer before SG pressures divert. Post-trip, auxiliary feedwater system (AFWS) input feed imbalances may also exist among the SGs. While expectations are that the crew may enter E-3 at this point, it is noted that a couple of steps later in E-0, there is another step calling for transition to E-3 based on an SG-level-checking step (if any SG level is rising uncontrollably, go to E-3).

If/when the crew enters E-3, the scenario proceeds in response to the crew’s actions, with no failures or other complicating factors induced by the simulation design: that is, the plant response will be based on what the crew does in carrying out procedure E-3. In general, the expectation is that the crew will perform four primary tasks corresponding to the human failure events (HFEs) defined for the base SGTR scenario. These tasks include (1) identifying which SG is ruptured and isolating it, (2) cooling down the reactor coolant system (RCS) expeditiously by dumping steam, (3) depressurizing the RCS expeditiously using the pressurizer sprays but also likely by using a pressurizer power-operated relief valve (PORV) to expedite the depressurization, and (4) stopping safety injection (SI) upon indication that the SI termination criteria are met. Note that [2] concentrates mainly on the HFEs following the SG isolation, as the qualitative analysis of the HFEs for identification and isolation was the topic of the pilot phase, described in [1].

A.2 SGTR complex scenario

This scenario is similar to the SGTR base scenario, except for five very significant differences. Of these, two are relevant for the present analysis:

- The event starts off with a major steamline break with a nearly coincident SGTR in SG #1 that will cause an immediate automatic scram and expectations that the crew will enter the E-0 procedure.

- Auto closure (as expected) of the main steamline isolation valves (MSIVs) in response to the steamline break, but along with failure of any remaining secondary radiation indications (not immediately known or expected by the crew) as part of the simulation design.

The steamline break serves to “drive” the plant response early in the scenario with the initial plant behavior like that expected in response to a significant steamline break, with quick closure of the MSIVs. This fact, along with the failure of all secondary radiation indications/alarms, is expected to “mask,” at least initially, the nearly coincident occurrence of the SGTR in SG #1. This should make it considerably more difficult for the crew to diagnose the existence of the SGTR, especially in response to step 19 in the E-0 procedure, which concerns elevated radiation indications.

If/when the crew does enter E-3 with the same tasks as described for the SGTR base scenario expected (see the base case scenario description above for the various opportunities to transfer to E-3), one bus failure will also be initiated in the simulation design in order to ensure that the crews are forced to use a pressurizer PORV to perform the desired RCS depressurization (the bus failure will cause the failure of a reactor coolant pump (RCP) that reduces the pressurizer spray efficiency). Once the desired RCS depressurization is completed (this is expected to take ~5-10 minutes), the crew, in following the steps in E-3, is directed to close the PORV. At this point, unbeknownst to the crew, the PORV will remain partially open, allowing about 6% flow. For one half of the crews, the PORV position indication will show “closed”; for the other half it will show “open.” At the PORV closure step in E-3, it is expected that if the desired closed indication is not immediately evident (which it won’t be for the crews for which the valve shows “open”), the crew is supposed to give a closing order to the PORV block valve associated with the PORV of interest. The next step in E-3 calls upon an indication that is readily viewable (i.e., RCS pressure and whether it is rising). RCS pressure will essentially be stable or only rising extremely slowly (because of the leaking PORV), rather than much more quickly, as expected. RCS pressure will tend to lower quickly, as the leaking PORV provides sufficient pressure relief to make it difficult to maintain pressure. All of this could be a sign to check additional supporting indications that will show increasing adverse indications for the pressure relief tank (PRT), including temperature and level continuing to rise and subsequent loss of pressure when the PRT rupture disk fails, all of which are signs of a continuing leak that needs to be isolated. If this additional evidence is viewed and acted upon by the operators, the operators need to conclude that there is strong evidence of a leaking PORV, attempt to close the associated block valve (i.e., give it a closing order), and transfer to procedure ECA-3.1.

A.3 SGTR HFE definitions and event tree

Figure A-1 below represents a typical probabilistic risk analysis (PRA) event tree for an SGTR event. It is presented here to provide an overall PRA context for the HFEs to be evaluated. Its sequence end states (outcomes) refer to whether the reactor core is safe in the long term, or whether there is core damage (CD). Those paths through the event tree and the relevant HFEs of interest for the current study are set in bold. All other sequences on the event tree, and those system successes or failures or operator actions associated with refueling water storage tank (RWST) refill, were not simulated.

As a model of an accident sequence, the event tree represents, in a general manner, the way the operators are trained to respond to an SGTR event with the E-3 procedure. However, when performing a PRA, the success criteria for the events are typically determined by successfully avoiding irreversible changes to the plant state that affect the likelihood of core damage. For this exercise, the training staff expectations of the operator responses were considered in determining the success criteria. These expectations are reflected in the crews' training. In applying the procedures, the operators are also trained to be concerned about more intermediate and detailed goals that are particularly relevant to an SGTR event. The operators are taught that "success" means "timely operator intervention in order to limit the radiological releases and prevent steam generator (SG) overfill" (a quote from a basis document for the procedures), and to terminate primary-to-secondary leakage expeditiously. They want to limit the radiological releases that are, in part, a function of how long it takes before the rupture is mitigated, and they do not want to overfill the ruptured SG, since this could cause an SG pressure relief valve to open (thereby allowing more release), or worse yet, cause a main steamline break or leak (also allowing more release, as well as further complicating the shutdown).

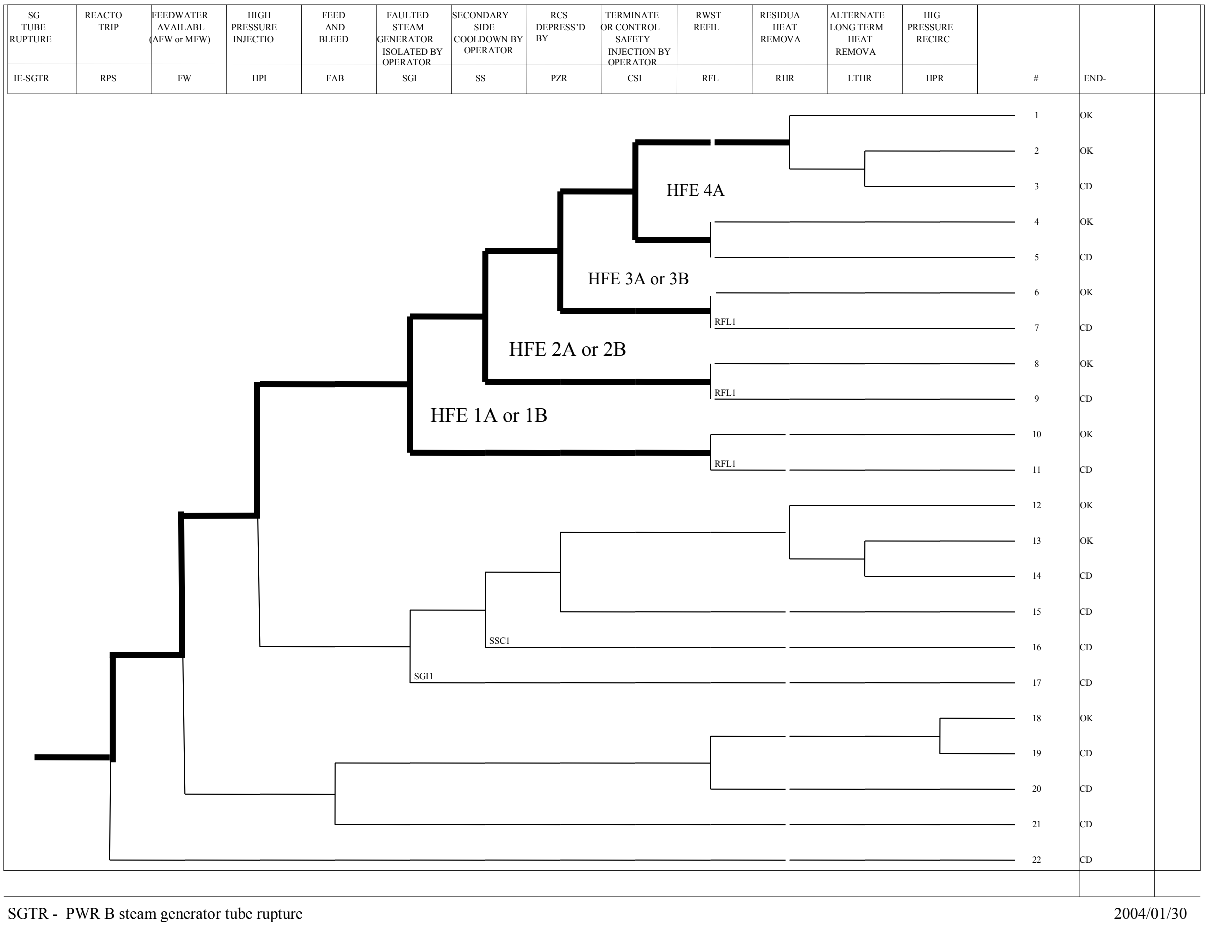

Figure A-1 Event tree for SGTR scenario

This overall more pertinent goal (in the operators' minds) of limiting the radiological release is achieved by performing the tasks in the E-3 procedure. For the HFEs analyzed in this report, the relevant tasks are identifying and isolating the ruptured SG, cooling down and depressurizing the RCS system, stopping the SI, and achieving primary-secondary pressure balance. Because of the overall goal of limiting radiological release, the operators are trained to perform these actions expeditiously, and the procedures are designed accordingly.

The operators are also taught that failure in any of these tasks has undesirable consequences: for instance, if the affected SG is not identified and isolated, releases will remain high, which is an outcome to be avoided.

The operators, in their training, are taught about these undesirable consequences, and that they need to perform the tasks expeditiously and correctly, as specified in the procedures. They are also taught that, in order to limit the release, all tasks should be completed before the ruptured SG overfills; thus, while operators do not think of the task in terms of clock time, they are aware of the need to get through the tasks with some urgency in order to meet the overall goal. Based on this awareness, when they simulate an SGTR event in their training, some level of expectation exists regarding typical response times to perform the various tasks. The HFE definitions of success-failure are based on these temporal expectations, along with what is to be accomplished for each task. While the threshold times to perform each task, as provided in the HFE definitions, are not exact, they do represent times by which the operators could be viewed as being slower than expected, since the overall goal could then be jeopardized.

Based on these considerations, the HFEs were defined as follows:

HFE-1 (A & B): Failure of the crew to identify and isolate the ruptured SG:

Success requires that the crew:

- Enters procedure E-3 (preferably from E-0 Step 19).
- Has closed/isolated all steam outlet paths from the ruptured SG (SG #1)
- Has stopped all feed to the ruptured SG, as long as the ruptured SG level is at least 10%, as indicated on the narrow range SG level indications (to ensure that the SG U-tubes will remain covered).
- It is expected to take the crew about 8-10 minutes after entering E-0 to reach the vicinity of step 19 in E-0. Allowing at least a few minutes before plant trip for the crew to observe and evaluate the initial indications of the tube rupture, about 8-10 minutes to enter and get through E-0 to step 19, five minutes to actually enter E-3 and perform the initial isolations/stoppages, and an additional few minutes for reasonably acceptable variability among crew responses, we assume that failure to successfully perform the above within 20 minutes (base case, HFE-1A) or 25 minutes (complex case, HFE-1B) *once the tube rupture occurs* (which is the start of the event) constitutes "failure," as this would be a slower response than expected/desired.

Note: the isolation manipulations involve the following, and would typically take less than three minutes to do:

Control room actions. These are all expected to be done by the crew, and are part of the HFE:

- Verify steam dump to atmosphere valve set point at 70.5 bar.
- Verify blow down isolation.
- Verify main feedwater isolation.
- Close steam valve to turbine-driven auxiliary feedwater (AFW) pump.
- Close main steamline isolation valve and its bypass valve.
- Stop AFW when level is greater than 10%.

Local actions. The crew should at least call for these actions, which are part of this HFE.

- Verify steam dump to atmosphere valve closed.
- Lock steam valve to turbine-driven AFW pump.
- Verify steam traps closed.

HFE-2 (A & B): Failure of the crew to cooldown the reactor coolant system expeditiously:

The crew is supposed to cool down much faster than 100 F/hr for the SGTR base scenario. This is anticipated to be performed by dumping steam from one or more intact SGs. Success requires that the crew:

- Performs the cooldown using either or both the steam dump valves to the atmosphere or to the main condenser, such that an RCS temperature corresponding to the pressure in the faulted SG is reached, along with corresponding adequate RCS subcooling (see the enclosed subcooling margin figure at the end of this document), and then subsequently terminates the cooldown.

- Maintains the RCS temperature below the limit value.

- It is expected that this initial cooling should take about 10 minutes, if performed in the desired expeditious manner, once the cooldown step (step 7 in E-3) is reached. We will assume that failure to successfully perform the expeditious cooldown and then terminate the cooldown while meeting the above criteria within 15 minutes of reaching the cooldown step in E-3 (step 7) constitutes “failure,” as this would be a slower response than expected/desired, even allowing for some variability in the speed of the crews.

HFE-3 (A & B): Failure of the crew to depressurize the RCS expeditiously:

(To minimize the break flow and refill the pressurizer for the SGTR base scenario.) While the goal is to perform the depressurization and then subsequently terminate depressurization once the crew achieves an RCS pressure lower than the pressure in the ruptured SG, ultimate success (so as to be able to move on in the procedure) requires that the crew:

- Achieves and maintains a pressurizer level greater than 10%.

- Avoids exceeding a pressurizer level greater than 75% (the crew should stop depressurization even if the RCS pressure is higher than the pressure in the ruptured SG).

- Avoids going too low in subcooling by virtue of not maintaining the RCS pressure and temperature within the allowed range, using the prescribed subcooling margin.

- Since the desire is to perform this expeditiously, the depressurization should be completed in less than 10 minutes once the depressurization step in E-3 (step 16) is reached. Allowing for reasonably acceptable variability among the crews, we will assume that failure to perform an expeditious depressurization while meeting the above success criteria within 15 minutes of reaching the depressurization step in E-3 (step 16) constitutes "failure," as this would be a slower response than expected/desired.

HFE-4A: Failure of the crew to stop the safety injection (SI):

(Such that only a single charging/SI pump is running/injecting, and the SI flowpath is isolated).

Success requires that the crew:

- Stops all hi head SI pumps except, for a single pump, isolates SI flowpath, and establishes charging with the single remaining pump when the shutoff criteria (see the E-3 procedure) are met so that the crew can maintain RCS coolant level and pressure control.

- Performs the stoppage before the RCS repressurizes to the point of exceeding the ruptured SG pressure (assuming it was lower after the cooldown and depressurization).

- It is preferable for the stoppage to occur before the ruptured SG is overfilled (sustained 100% level on indicating wide range), but this is not a requirement.

- Note that the manipulations involved with the first bullet require that the following be performed in order to be "successful":

    - Stop all but one charging pump with its suction remaining aligned to the RWST (it should already be that way) and verify that the charging pumps' minflow valves are open.

    - Isolate the boron injection tank (BIT) by closing the two BIT inlet isolation valves as well as the two BIT outlet isolation valves, and verifying that the BIT bypass valve is closed.

HFE-5B1: failure of the crew to give a closing order to the PORV block valve associated with the partially open PORV within 5 minutes of closing the PORV (but it remains partially open, allowing ~6% flow) used to depressurize the RCS:

This action would recognize that the PORV path may be open or leaking, based on all supporting indications (e.g., pressurized relief tank (PRT) indications), even though the PORV position indication shows "closed." (Half of the crews will be given the "closed" indication for the PORV position).

HFE-5B2: failure of the crew to give a closing order to the PORV block valve associated with the partially open PORV within 5 minutes of closing the PORV (but it remains partially open, allowing ~6% flow) used to depressurize the RCS:

This action would recognize that the PORV path may be open or leaking, given that the PORV position indication shows "open," along with the other supporting indications of the leak path. (Half of the crews will be given the "open" indication for the PORV position.)

HFE-4A applies to the base scenario only. HFEs 5B1 and 5B2 apply to the SGTR complex scenario only, and are two different versions of HFE-5B. Half of the crews were in a group analyzed per HFE-5B1, and the other half were in a group analyzed per HFE-5B2.

A.4 LOFW base scenario

The LOFW HFEs analyzed in this study occur in two versions of an LOFW scenario, a base case (total LOFW without further complications) and a complex case (LOFW with further complications). In both versions, the main tasks for the crews are to (1) detect loss of feedwater (FW), (2) try to reestablish FW, and (3) start bleed and feed (B&F). In the complex scenario, the action "depressurize SG" is also part of the reestablishment of FW.

In a situation following a total LOFW, the reactor core is cooled by vaporization of the remaining water in the SGs. The first goal for the operating crews is to try to reestablish feedwater. If feedwater cannot be reestablished, the SGs will eventually become empty and unable to cool the core. To establish another means of core cooling before the SGs are empty, bleed and feed (B&F) of the reactor coolant system should be started. Primary B&F consists of manually starting SI pumps and opening the pressurizer relief valves. The criteria for starting B&F according to the plant's procedure for sustained LOFW (functional restoration procedure FR-H.1) are that the SG wide range (WR) level should be less than 12% in two out of three SGs, or that the reactor pressure is high due to loss of secondary heat sink. These criteria are the cues for detecting the sustained LOFW. According to the emergency procedures, the crews shall try to restore FW to the SGs until these criteria are met, in which case B&F is required.

A.5 LOFW complex scenario

The complex scenario contained multiple issues. The first issue was that one condensate pump was successfully running, leading the crew to depressurize the SGs to establish condensate flow; however, the running condensate pump was degraded, and gave a pressure so low that the SGs became empty before the pressure could be reduced enough to successfully inject water. The procedure step to depressurize is complicated, and this action both kept the crew busy and gave them a concrete chance to reestablish feed water to the SGs. The crews were directed by procedure FR-H.1 to depressurize the SGs to inject condensate flow.

In addition to this, in the complex scenario, two of the three SGs had WR level indicators that would incorrectly show a steady (flat) value somewhat above 12% when the actual level would be 0%, as shown in Figure A-1. The two failing SG levels both indicated a level above the 12% criterion to start B&F. To be able to follow the criteria, the crews had to identify and diagnose the indicator failures, since the criteria, interpreted literally, would never be met.

A.6 Main tasks in the LOFW scenarios

- Detect LOFW. Following procedures to start monitoring the critical safety functions is important in quickly transferring to the correct procedure, FR-H.1. When the crew transfers to procedure FR-H.1, they stop the RCPs in step 3. The time from the start of the scenario to the time that the crew stops the RCPs can be used to measure how fast they detect LOFW (Tables 4-1 and 4-2).

- Re-establish FW. The FR-H.1 procedure will guide the crew in trying to re-establish feed flow to the SGs. The crew needs to organize actions to check the status of all possible ways to feed the steam generators, and try to re-establish the different sources of FW.

- Depressurize SG (complex scenario only). If at least one condensate pump is running, flow to the SGs can be established by depressurizing the SGs to a pressure lower than the discharge pressure of the condensate pump(s). In the complex scenario, one condensate pump is running, and the crews will be guided in procedure FR-H.1 to depressurize the SGs. The guiding procedure step 7 is complicated to follow (e.g., using auxiliary spray and blocking SI signals). One difficulty is that the procedure step instructs the crews to depressurize to less than 35 bars, because the condensate pumps normally give around 40 bars. In the complex scenario, the one running condensate pump only gave about 26 bars, and depressurizing to less than 35 bars according to the procedure was not enough to establish feed flow. Because the running condensate pump gave a lower pressure than normal, condensate flow could not be established before the SGs were empty. If the crews managed to establish flow to the SGs from condensate, the condensate pump was tripped.

- Start B&F. In a situation of total LOFW, the reactor core is cooled by vaporization of the remaining water in the SGs. If feedwater cannot be re-established, the SGs will eventually become empty and unable to cool the core. It is important to establish another means of core cooling before the SGs are empty. This is done by initiating B&F (i.e., starting SI and opening the PORVs). The criteria for starting B&F in the FR-H.1 procedure are that the WR level should be less than 12% in two out of three SGs, or that the reactor pressure should be high due to loss of secondary heat sink. To be able to start B&F in time, the crews need to monitor the SG levels. In the complex case, the WR level was failed in two of the three SGs. To be able to follow the criterion to start B&F at 12% WR level, the crews had to identify the indication failures. While the WR level measurement of SG #2 was correct for the entire scenario, the measurement for SG #3 was failed 14% high from the start, and consequently indicated 14% when SG #3 was, in fact, empty. The WR level measurement for SG #1 initially worked correctly, but became stuck at 16% and remained at that level for the rest of the scenario, showing a flat line in the trend displays. The WR level measurements for SGs #2 and #3 showed flat lines (at 0 and 14%) when the SG became empty (see Figure A-1).

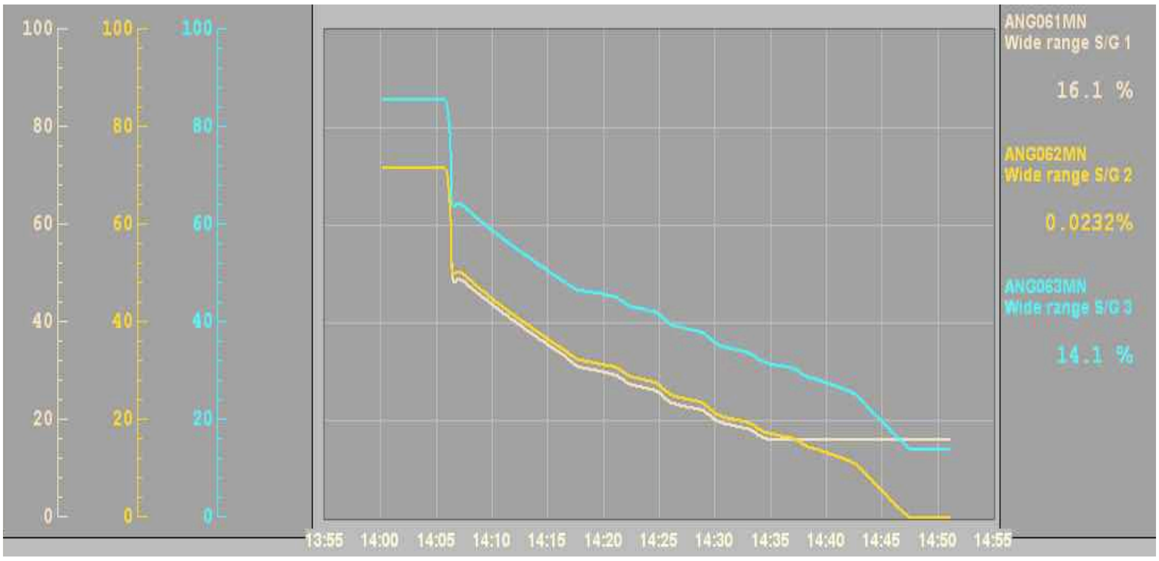

Figure A-2 Example of wide range steam generator measurements in the complex LOFW scenario

A.7 LOFW HFE definitions and event tree

Figure A-2 below represents the PRA event tree for an LOFW event. It is presented here to provide an overall PRA context for the HFEs to be evaluated. Its sequence end states (outcomes) refer to whether the reactor core is safe in the long term, or whether there is CD. Those paths through the event tree and the relevant HFEs of interest for the current study are described below. All other sequences on the event tree, and those system successes or failures or operator actions following late recovery of B&F (X4L), were not simulated.

The HFEs of interest for the study were defined as follows:

- X4 = Initiation of Primary B&F = Establish/Initiate B&F before SG dryout. SG dryout occurs when there is no water left in the SGs, indicated by 0% WR SG level.
- X4L = Late Recovery Before Core Damage = Establish/Initiate B&F within 25 minutes of SG dryout. This HFE is conditional on X4 (failure of B&F before dryout).

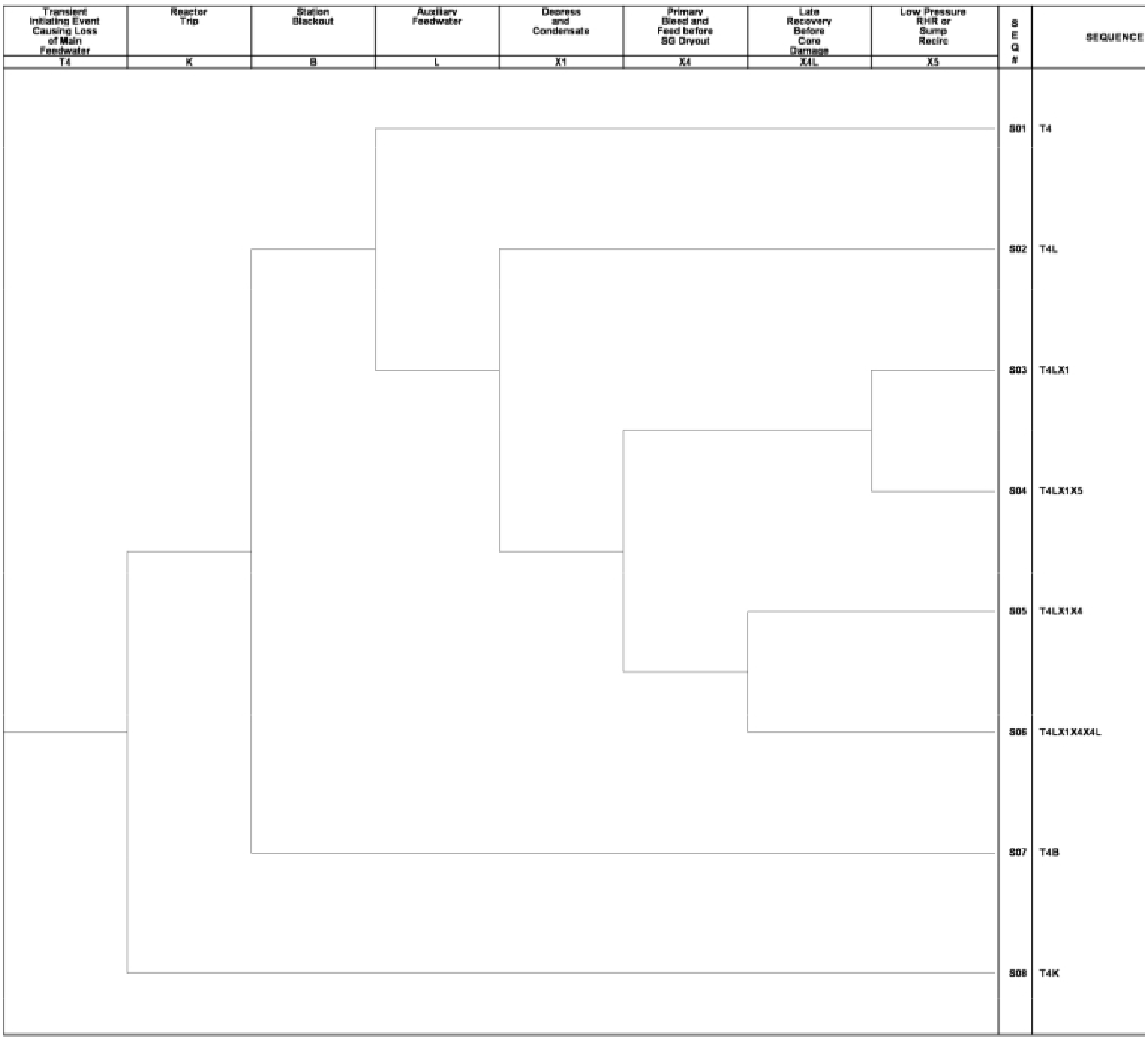

Figure A-3 Loss of feedwater event tree

The HFEs to be estimated by the HRA teams are coded as follows:

- HFE-1A: X4 in the base case
- HFE-2A: X4L in the base case
- HFE-1A1: X4*X4L in the base case
- HFE-1B: X4 in the complex case
- HFE-2B: X4L in the complex case
- HFE-1B1: X4*X4L in the complex case

NRC FORM 335
(12-2010)
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-2168,
Volumes 1, 2, and 3
DRAFT

2. TITLE AND SUBTITLE
Environmental Impact Statement for an Early Site Permit (ESP) at the PSEG Site, Draft Report for Comment

3. DATE REPORT PUBLISHED

MONTH	YEAR
August	2014

4. FIN OR GRANT NUMBER

5. AUTHOR(S)
See Appendix A.

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.)
Division of New Reactor Licensing
Office of New Reactors
U.S. Nuclear Regulatory Commission
Washington, DC 20555-0001

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
Docket No. 52-043

11. ABSTRACT (200 words or less)

This environmental impact statement (EIS) has been prepared in response to an application submitted to the U.S. Nuclear Regulatory Commission (NRC) by PSEG Power, LLC, and PSEG Nuclear, LLC (PSEG), for an early site permit (ESP). The proposed action requested in the PSEG application is the NRC issuance of an ESP for the PSEG Site located adjacent to the existing Hope Creek and Salem Generating Stations.

This draft supplemental environmental impact statement includes the preliminary analysis that evaluates the environmental impacts of the proposed action and alternatives to the proposed action.

After considering the environmental aspects of the proposed NRC action, the NRC staff's preliminary recommendation to the Commission is that the ESP be issued as requested. This recommendation is based on (1) the application submitted by PSEG, including Revision 3 of the Environmental Report (ER), and the PSEG responses to requests for additional information from the NRC and USACE staffs; (2) consultation with Federal, State, Tribal, and local agencies; (3) the staff's independent review; (4) the staff's consideration of comments related to the environmental review that were received during the public scoping process; and (5) the assessments summarized in this EIS, including the potential mitigation measures identified in the ER and this EIS.

12. KEY WORDS/DESCRIPTORS (List words or phrases that will assist researchers in locating the report.)
PSEG ESP
PSEG Site
Draft Environmental Impact Statement, DEIS
National Environmental Policy Act, NEP A
NUREG-2168

13. AVAILABILITY STATEMENT
unlimited

14. SECURITY CLASSIFICATION
(This Page)
unclassified
(This Report)
unclassified

15. NUMBER OF PAGES

16. PRICE

UNITED STATES
NUCLEAR REGULATORY COMMISSION
WASHINGTON, DC 20555-0001

OFFICIAL BUSINESS

BLOG CHAT You Tube

NUREG-2127

The International HRA Empirical Study – Lessons Learned from Comparing HRA Methods Predictions to HAMMLAB Simulator Data

August 2014