SAND2007-4037C

An Empirical Study of HRA Methods - Overall design and issues

Vinh N. Dang[1], Andreas Bye[2], Erasmia Lois[3], John Forester[4], Alan M. Kolaczkowski[5], Per Øivind Braarud[2]
[1] Paul Scherrer Institute, Switzerland, vinh.dang@psi.ch
[2] OECD Halden Reactor Project, Norway
[3] U.S. Nuclear Regulatory Commission, Rockville, MD, USA
[4] Sandia National Laboratories, Albuquerque, NM, USA
[5] SAIC, Eugene, OR, USA

Abstract—**A diversity of Human Reliability Analysis (HRA) methods are currently available to treat human performance in Probabilistic Risk Assessments (PRAs). This range of methods reflects traditional concerns with human-machine interfaces and with the basic feasibility of actions in PRA scenarios as well as the more recent attention paid to Errors of Commission and decision-making performance. Given the differences in the scope of the methods and their underlying models, there is a substantial interest in assessing HRA methods and ultimately in validating the approaches and models underlying these methods. A significant step in this direction is an international evaluation study of HRA methods, based on comparing the observed performance in simulator experiments with the outcomes predicted in HRA analyses. Its aim is to develop an empirically-based understanding of the performance, strengths, and weaknesses of the methods. This paper presents the overall methodology for this initial assessment study.**

I. Introduction

A diversity of Human Reliability Analysis (HRA) methods are currently available to treat human performance in Probabilistic Risk Assessments (PRAs). This range of methods reflects traditional concerns with human-machine interfaces and with the basic feasibility of actions in PRA scenarios as well as the more recent attention paid to Errors of Commission and decision-making performance. Given the differences in the scope of the methods and their underlying models, there is a substantial interest in assessing HRA methods and ultimately in validating the approaches and models underlying these methods. A significant step in this direction is an international evaluation study of HRA methods, based on comparing the observed performance in simulator experiments with the outcomes predicted in HRA analyses. Its aim is to develop an empirically-based understanding of the performance, strengths, and weaknesses of the methods. It is expected that the results of this work will provide the technical basis for the development of improved HRA guidance and, if necessary, improved HRA methods.

As a first step in the overall HRA method evaluation study, a pilot study is currently under way to obtain initial data and help establish a methodology for assessing HRA methods using simulator data. The operating crews from a nuclear power plant participated in a series of scenarios in the Halden Reactor Project's HAMMLAB (HAlden huMan-Machine LABoratory) simulator facility in late 2006. Without knowledge of the crews' performances, HRA analysis teams are performing predictive analyses of the scenarios. This paper presents the methodology for this pilot assessment study, highlighting the major aspects of its three elements: the design of the scenarios and experimental data collection and analysis, the predictive HRA analyses, and the approaches for the comparison of the predicted and observed outcomes. Companion papers [1][2] address the study design issues associated with each of these areas in more detail.

A number of organizations from ten countries are participating in the study; these include industry, regulators, and the research community. In particular, the U.S. NRC has played a major role in supporting the preparation and execution of the study. A series of workshops and meetings is organized in order to facilitate the study. A steering group has been established.

II. Overview of the Study Design

In the period from October to December 2006, an experiment was performed in Hammlab called Performance Shaping Factors (PSF) and Masking, utilizing PRA relevant scenarios to study the effects of masking and other PSFs for HRA [3]. 14 crews of three licensed PWR operators per crew participated in this study. All crews responded to two versions (a base case and a more challenging case) of two scenarios, a steam generator tube rupture (SGTR) and a total loss of feed water (LOFW). This paper does not include the specifics of the scenarios that are used. In view of the HRA and data analysis workload needed for the evaluation study as well as the desire to assess the evaluation study methodology itself, it was decided to use one half of the data, i.e. both variants of one scenario, in the pilot study in 2007, and the other scenario runs for a follow-up study in which improvements to the methodology could be incorporated as needed.

A. Participants and Their Roles

The study is designed around four sets of participants:

- HRA teams: each team applies one or more HRA methods to obtain predictions for the Human Failure Events (HFEs) in the scenarios defined for the study.
- Halden experimental staff: the simulator sessions are conducted in the Hammlab facility by Halden staff. The

staff has the main responsibility for the analysis of the experimental data.

- Operator crews: a set of licensed operator crews respond to the scenarios simulated in Hammlab. Each crew responds to 4 scenarios, consisting of a base and "complex" variant of two scenario types.
- Study assessment group: An information package (analysis inputs) for the HRA teams is prepared. Together with Halden staff the group answers requests from the HRA teams for additional information and questions concerning ambiguities in the instructions and assumptions. During this stage, the assessment group does not receive any information about the actual crew performances, in order to avoid coloring the information provided in ways that could bias the HRA analyses. Finally, the study assessment group reviews the HRA team responses and performs the assessment and comparison of the predicted outcomes vs. experimental outcomes, together with experimental staff from Halden.

The overall design of this study is shown in Figure 1. The activities of the teams applying the HRA methods ("HRA Teams") are shown on the left. The collection of the simulator data is performed by the Halden staff, using measures that have been used in previous work in Hammlab. The data analysis is performed primarily by Halden staff as well, with the assessment group contributing in the HRA-related aspects. The comparison of predicted outcomes with experimental outcomes is mainly the responsibility of the assessment group assisted by Halden staff.

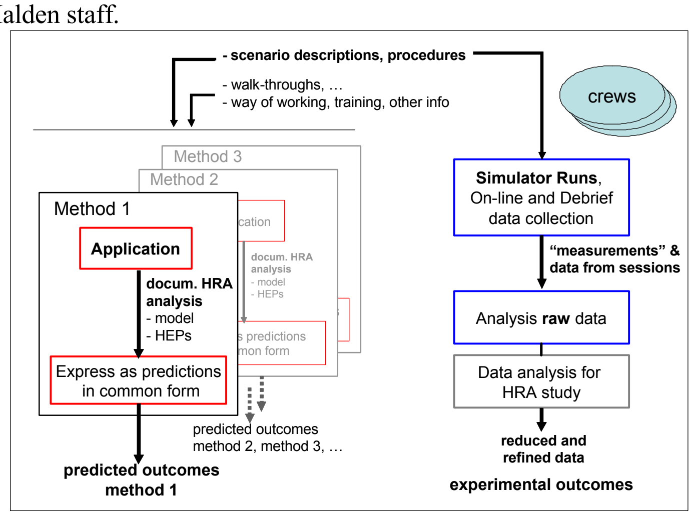

Figure 1. Overview of the study.

B. *Focus on the HRA of post-initiator actions*

The study focuses on the HRA of the personnel actions required in the response to PRA initiating events. This focus is motivated by the significant research and development efforts on HRA methods addressing the issue of errors of commission and decision-making performance. These include for instance [4-7]. In addition, the Hammlab simulator facility is oriented to HMI and human performance issues in the context of normal and emergency operation.

Although scenarios related to full-power operation as well as low power and shutdown operation could be treated in the study, full-power PRA scenarios have been selected because of their risk significance and because the configurations of the plant are fewer and better defined. Moreover, the typically larger time windows for operator actions in shutdown PRA scenarios lead to additional problems for the Hammlab data collection (for instance, the length of the simulator sessions would have to be longer to account for the slower dynamics of shutdown scenarios).

C. *The effect of highly reliable performance*

Human performance levels in nuclear power plant operations are generally high because of the comprehensive analysis of potential accident scenarios, the support provided to the operators by abnormal and emergency procedures, and the extensive, on-going training programs. As a result, operator crews are expected to make few errors in most scenarios; moreover, unrecovered errors are expected to be even less likely (the probability of HFEs in PRAs in general accounts for recovery factors, for instance, that a fellow crew member detects an omission).

The scenarios that are selected for HRA-related studies at Halden are designed to present a significant challenge to the operators, relative to, for instance, design basis accidents. While the crews can be expected to experience some difficulties, the overall performance is still expected to be generally successful. Naturally, it would in principle be possible to design scenarios that are so difficult that a significant number of the operating crews could be expected to fail on the same task (HFE). In this context, a significant number of the crews means 20 to 50% or more of the crews. Such scenarios, however, are not especially interesting in terms of HRA quantification. The consensus is that the main issue with scenarios that practically "force" errors is to identify plausible candidates, scenarios that do not require extremely unlikely combinations of conditions.

Therefore, a major challenge for an empirical study of HRA is that if performance is reliable, few observations of crew responses will include failures for PRA-level HFEs. Consequently, the crew responses in a set of simulator experiments cannot usually be used to obtain an empirical, "observed" HEP.

However, for some types of human failure events even within the scenarios designed in this study, it is possible to create required actions and action sequences that are so difficult that a large number of the operating crews would be expected to fail. These HFEs do not necessarily represent the typical HFEs in PRA, for which the HRA methods are designed to model, but are defined on a lower task level. The predicted and observed HEPs thus cannot be compared in a straightforward way. Nevertheless, the qualitative information related to the challenging parts of the scenario will be informing for the study.

To summarize, the reliable performance expected in HRA-related scenarios has the following implications for the study:

- a comparison of observed HEP for an HFE with predicted HEPs is generally not possible

- the study focuses on comparing the qualitative predictions of the HRA analyses with the tendencies observed in the Hammlab sessions

A comparison of the HEPs obtained by the different teams using different methods on a given HFE is of course possible. On the other hand, such a comparison would only be minimally informed by the simulator observations. As a result, the comparison of HEPs is not emphasized in the study.

D. Focus on the qualitative outcomes predicted in HRA analyses

At a high level, HRA methods have the same purpose (or aims) due to the role of the HRA within the PRA. These common aims are:

- identification of the HFEs to be included in the PRA accident sequence model.
- qualitative analysis of the HFEs, i.e. the tasks and the performance conditions, in order to identify the main failure modes and the plant- and scenario-specific influences (often called performance shaping factors) that may affect these failure modes
- quantification of the probability of these HFEs, the HEPs, based on the qualitative analysis.

As discussed in the preceding section, there are difficulties in setting up the study such that the HEPs predicted by applying HRA methods can be compared to the outcomes of the experiments observed in Hammlab.

With regard to the identification of the HFEs for a given scenario, this is typically done in PRAs as an interaction between HRA analysts and the accident sequence analysis. With the exception of some of the more recent methods, the task analysis approaches used for HFE identification are not part of the HRA methods themselves. For the study, the study organizers have performed this step and identified the HFEs of interest, for which the HRA analyses are to be performed.

- Since the approaches for HFE identification are not prescribed and not part of most methods, asking the teams to perform HFE identification would not test the individual HRA methods and would introduce differences due to the selected approach
- The HFEs represent not only the key required operator actions but also the modeling of the sequence (sequence delineation), leaving this step to the HRA teams would also risk that each team would define the scenario models and HFEs differently. Several scenario models and the associated HFE definitions could be valid but the simulator data analysis would be complicated by the need to match these variations in the HFE definitions.

It should be noted that defining the HFEs for the HRA teams does not eliminate the qualitative analyses to be performed. As noted by Kirwan [8, p. 318], "targeted task analyses" should be performed in support of the HRA. This process identifies the main failure modes and the plant- and scenario-specific influences on human performance. Requirement HLR-HR-G of [9] lists a number of these influences. The most important influences or factors are sometimes referred to as the factors "driving" performance or the "driving factors" of performance. Comparing the specific factors identified as driving factors by the HRA teams for the defined HFEs with those observed in Hammlab is the main focus of the comparison.

E. Inter-rater and inter-method reliability

In many ways, this empirical evaluation study of HRA methods includes many elements of a benchmark. These include

- the involvement of multiple methods and analysis teams
- predictive analysis using these methods, performed without knowledge of the experimental results
- the collection of experimental data, and
- the comparison of the experimental results with analysis predictions

In spite of these elements, this initial pilot study is not in itself a benchmark study. As noted, its aims are to better understand the strengths and weakness of the individual HRA methods in light of the empirical data, to obtain initial data relevant to benchmarking, and to help establish a methodology for assessing HRA methods using simulator data.

Therefore, some issues are left unresolved in this first pilot study that would be important for a benchmarking study. Although there are HRA "good practices", e.g. the recent [10], there can be significant variability in the way HRA methods are applied and it can be difficult to put in provisions to ensure consistency in the application of the methods. Aspects of inter-rater reliability (consistency of results obtained from two or more sets of analysts) are left out in the initial pilot study, there are not multiple teams applying each method. Finally, given the explorative goal of the pilot study, we do not apply any measure to ascertain that the HRA teams are similarly qualified. Also, each HRA team only applies one method (in general). Nevertheless, each HRA team includes expert users of the methods that it applies, and in some cases, the teams include the developers of the methods.

In sum, this first pilot study does not intend to address the issues of inter-rater reliability and inter-method reliability. The focus is on the quality of the information provided by each of the methods, when applied by knowledgeable users.

III. Applications of HRA methods and reporting of predicted outcomes

All of the HRA teams received an information package, to be used as the basis for the application of an HRA method (or several). This information package included the following items:

1. Overview (of the information package) and instructions to the HRA teams
2. Administrative information and agreement forms
3. Study outline
4. Hammlab information
5. Scenario description and HFEs
6. Characterization of the crews, their work practices and training
7. Procedures used in Hammlab
8. Forms for the responses of the HRA teams

More generally, the package provides information about the organization of the study, the general performance conditions (e.g. information about the interface, the work practices of the crews, the procedures), information about the specific scenarios simulated in Hammlab, and forms for the HRA team responses. It can be seen that only item 5 is scenario-specific.

In item 7, it has not been possible to provide a complete set of Hammlab procedures in English due to availability and organizational issues. Consequently, the procedures included in the package are limited to those expected to be used in the scenario variants for this phase although the study organizers recognizes that information in other procedures may have an influence on the crews' performance in the scenario.

The HRA teams have the opportunity to request clarifications or additional information during the course of analyzing the scenarios. To ensure a common understanding of the scenarios and predictions and consistent assumptions among the HRA teams, all questions and all answers are provided to all teams.

The HRA teams provide their analyses and the outcomes they predict in three ways. Form A is a free-form response and consists of open-ended questions. Form B is a so-called "closed-form" response, in which the responses are structured. Finally, the HRA teams are asked to provide their analysis documented according to PRA (good) practice.

For example, Form A question 2 begins: "Provide a summary of the most influencing factors on the crews' behavior with respect to this HFE and why they are important… Do use the terminology of the HRA method [that you are applying]". In contrast, on Form B, the teams are asked to provide their predictions using a set of terms based on the HERA classification [11]. They are asked to select the most important factors from among a pre-defined set of influencing factors and to rank these. The Form B response has two purposes

- to add some measure of consistency in the terminology used in the predictions of the outcomes
- to allow the analysis of the simulator experimental data to be oriented to a single terminology, thereby reducing the effort required.

On the other hand, having both forms of responses will support other comparisons at a later stage.

IV. Experimental method and measures

The study decided to utilize the data from the PSF/Masking experiment that had an extensive data collection in the fall 2006. Thus, the design of the scenarios and all details of the data collection were decided in this project. A description of the design, the experimental measures, etc. is given in [3]. An extract of this is given below.

14 crews with licensed PWR operators participated in the study. Each crew consisted of a Shift Supervisor, a Reactor Operator and an Assisting Reactor Operator. The Hammlab PWR simulator is a full scope simulator of a French plant (CP0 series). Hammlab uses a computerised human machine interface for the PWR simulator. The Hammlab PWR procedures are based on the procedures used at the participating operators' home plant. The procedures are adapted to the simulated PWR and the Hammlab interface. The participating operators' home plant uses the Emergency Response Guidelines (ERGs) developed by the Westinghouse Owners Group.

The crews' home plant has conventional control rooms with panels, and alarm tiles. The Hammlab PWR simulator is based on digital instrumentation and control. In addition, there are a few differences in systems/equipment in the actual plant and those simulated in the Halden PWR simulator, so the simulator is not precisely simulating the actual plant (e.g., the PORVs are different). Therefore, prior to participating in the experimental scenarios, the crews were trained on how to use the screen based interface and on the differences between their actual plant and the simulator.

The data collection included:

- Crew interview: After each scenario the crew participates in an interview focusing sequentially on phases of the scenario.
- Operators PSF ratings: After the interview for each scenario is completed, the operators individually perform the rating of several PSFs for all scenario phases.
- Operator Background Questionnaire.
- Observer PSF ratings and comments: An observer sitting in the control room rates four PSF items for each scenario phase and provide free text comments for the same phases.
- OPAS and performance rating: Under each scenario run, a process expert fills in the Operator Performance Rating System from the gallery, by checking the completion of a set of predefined crews' actions and detections. He/she also rates the crews' overall performance of scenarios phases.
- Observer comments: Under each scenario run, a process expert verbally comments on interesting aspects of crews' activity and process development.
- Logs: All crews' activities on the simulator are logged.
- Audio/videos: Two fixed cameras behind the operators and two head mounted cameras on shift supervisor and reactor operator are employed. All operators are equipped with wireless microphones.

The detailed performance measures comprise extensive information about the various phases of the scenario. These phases correspond to the defined HFEs. All the various experimental measures, including extensive data collection on influencing PSFs and narratives about crew behavior, enable detailed descriptions of what the crews did, when they did it, and why. This constitutes a good basis for qualitative comparisons with the HRA method predictions for each HFE.

V. Comparison of predicted and experimental outcomes

The outcomes predicted in the HRA analyses performed by the teams are compared with the outcomes obtained from the Hammlab experiments on several levels. Analytical predictions are compared with experimental outcomes for each of the following:

- the factors that most influence the performance of the crews in these scenarios ("driving factors")
- the level of difficulty associated with the operator actions of interest (the HFEs), the reason for these

difficulties (or ease), and how the difficulties are expressed in operational and scenario-specific terms

The predicted and experimental outcomes in the structured, i.e. closed, form are intended to contribute considerably in the comparison, since this form is common to all of the prediction results (from all teams and methods). This requires though, that the methods can find the proper mapping of its results onto the form. It also depends on that the experimental results can be aggregated in a meaningful way into the form, that originally are designed for coding of one event at the time. The predicted outcomes as expressed in the open-ended responses and the documentation of each HRA analysis are used to provide a complementary view of the predictions. The extent to which the open-ended responses can be used directly for comparison to the experimental outcomes, will in the first phase depend on the clearness of the terms used by the teams in their responses. Nevertheless, the extensive experimental performance measures about driving factors and the data on why crews were doing what when, should lay the ground for a very interesting and effective comparison based on the open form as well.

The study and experimental plans do not anticipate that the Hammlab experiments can support the derivation of Human Error Probabilities (HEPs) from the experimental data. The number of sessions and crews (sample size) is too small and the expected levels of performance of the crews are too high, even in the more challenging scenario variants. As a result, the HEPs obtained for the HFEs may be compared only across the HRA methods and teams. The bulk of the comparative analyses focus on the ability of the methods to predict the tendencies of behavior and performance in the scenarios, i.e., on qualitative rather than the quantitative insights obtained by the methods.

Acknowledgment

The authors gratefully acknowledge the contributions of Helena Broberg, Salvatore Massaiu and Michael Hildebrandt, the Halden Project, and Ron Boring, Idaho National Lab, for major parts of the experimental work done so far in the project. Also, the additional members of the project assessment group, Gareth Parry, NRC, Bruce Hallbert, INL, Jeff Julius, Scientech, has provided invaluable inputs in the review work. The work of the HRA teams, 14 teams providing 15 HRA analyses, has of course been of invaluable importance. In addition, the interest and views expressed by the numerous participants to the preparatory workshops provided essential inputs to the design of the study.

The contributions of the OECD Halden Reactor Project to the Empirical Study of HRA Methods are supported by the Joint Programme of the OECD Halden Reactor Project. In addition, parts of this work are funded by the U.S. Nuclear Regulatory Commission (USNRC) and performed at Sandia National Laboratories. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000. The work was also supported by the Swiss Federal Nuclear Inspectorate (DIS-Vertag Nr. 82610). The opinions expressed in this paper are those of the authors and not those of the USNRC or of the authors' organizations.

References

[1] H. Broberg, A.M. Kolaczkowski, "Constraints in Designing Simulator Scenarios and Identifying Human Failure Events for Testing HRA Methods," 2007 IEEE 8th Human Factors and Power Plants (HFPP) and 13th Annual HPRCT Meeting, August 26-31, 2007, Monterey, CA, USA (*this conference*).

[2] J. Forester, A.M. Kolaczkowski, V.N. Dang, E. Lois, "Human Reliability Analysis (HRA) in the Context of HRA Testing with Empirical Data," 2007 IEEE 8th Human Factors and Power Plants (HFPP) and 13th Annual HPRCT Meeting, August 26-31, 2007, Monterey, CA, USA (*this conference*).

[3] P.Ø. Braarud, H. Broberg, S. Massaiu, "Performance shaping factors and masking experiment 2006: Project status and planned analysis", Proc. of the Enlarged Halden Programme Group meeting, Storefjell, Norway, 11-16 March 2007.

[4] P. Le Bot, F. Cara and C. Bieder, "MERMOS, A second generation HRA method: What it does and doesn't do", In: Proc. of the International Topical Meeting on Probabilistic Safety Assessment (PSA '99), 22–26 August 1999, Washington DC vol. II, American Nuclear Society, La Grange Park, IL (1999) p. 852–860 .

[5] J. Forester, D. Bley, S. Cooper, E. Lois, N. Siu, A. Kolaczkowski, J. Wreathall, "Expert elicitation approach for performing ATHEANA quantification," Reliability Engineering and System Safety, Elsevier, 2004, 83(2): pp. 207-220.

[6] J.W. Kim, W. Jung, J. Park, "A systematic approach to analysing errors of commission from diagnosis failure in accident progression," Reliability Engineering and System Safety, Elsevier, 2005, 89(2), pp. 137-150.

[7] B. Reer, V.N. Dang, S. Hirschberg, "The CESA method and its application in a plant-specific pilot study on errors of commission," Reliability Engineering and System Safety, Elsevier, 2004, 83(2), pp. 187-205.

[8] B. Kirwan. A Guide to Practical Human Reliability Assessment. New York: CRC Press, 1994.

[9] American Society of Mechanical Engineers, ASME Standard for Probabilistic Risk Assessment for Nuclear Power Plant Applications, ASME RA-S-2002 (and addendum ASME-RA-S-2003, 2003)., New York: ASME, 2002.

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

[11] B. Hallbert, R. Boring, D. Gertman, D. Dudenhoeffer, A. Whaley, J. Marble, J. Joe, E. Lois, Human Event Repository and Analysis (HERA) System – Overview, NUREG/CR-6903 Vol. 1, INL/EXT-06-11528, Idaho National Laboratory for the U.S. Nuclear Regulatory Commission, Washington DC, USA, July 2006.