Lessons Learned on Human Reliability Analysis (HRA) Methods from the International HRA Empirical Study[1]

J.A. Forester[a*], E. Lois[b], V.N. Dang[c], A. Bye[d], G. Parry[b+], and J. Julius[e]

[a]Sandia National Laboratories, Albuquerque, USA
[b]U.S. Nuclear Regulatory Commission, Washington DC, USA
[c]Paul Scherrer Institute, Villigen PSI, Switzerland
[d]OECD Halden Reactor Project, Halden, Norway
[e]Scientech, Seattle, USA

Abstract: In the International HRA Empirical Study, human reliability analysis (HRA) method predictions for human failure events (HFEs) in steam generator tube rupture and loss of feedwater scenarios were compared against the performance of real crews in a nuclear power plant control room simulator. The comparisons examined both the qualitative and quantitative HRA method predictions. This paper discusses some of the lessons learned about HRA methods that have been identified to date. General strengths and weaknesses of HRA methods are addressed, along with the reasons for any limitations in the predictive results produced by the methods. However, the discussions of the lessons learned in this paper must be considered a "snapshot." While most of the data has been analyzed, more detailed analysis of the results from specific HRA methods are ongoing and additional information may emerge.

Keywords: Human reliability analysis, HRA, benchmarking, simulator studies.

1. INTRODUCTION

Since human reliability analysis (HRA) can be an important contributor to the results of probabilistic risk assessment (PRA), which is an important tool used by the nuclear power industry and others in safety evaluations, the Office of Nuclear Regulatory Research (RES) of the Nuclear Regulatory Commission (NRC) is sponsoring work in an effort to improve the robustness of HRA methods and practices. Among other efforts, RES participates and supports an international collaborative effort (International HRA Empirical Study) was initiated to empirically assess, on the basis of data, the general strengths and weaknesses of a variety of HRA methods and to examine the reasons for any limitations in the predictive results produced by the methods.

The study involved the use of Halden Reactor Project's HAMMLAB (HAlden huMan-Machine LABoratory) nuclear power plant simulator facility. HRA analysis teams performed predictive analyses of operating crew performance in several accident scenarios and the results of these analyses were compared with reference data derived from the actual performance of real crews in the scenarios. The comparisons examined both the qualitative and quantitative method predictions. This paper discusses the lessons learned about HRA methods that have been identified to date. It addresses aspects of HRA methods identified as needing improvement and discusses potential follow-on studies needed to further understand reasons for the identified differences in HRA results. It should be noted that the discussions of the lessons learned in this paper must be considered a "snapshot." While most of the data has been analyzed, more detailed analysis of the results from specific HRA methods are ongoing and additional information may emerge.

The experimental methodology for the study, including the scenarios examined and human failure events (HFEs) quantified, the data collection and analysis process, and the process used to compare HRA method predictions with crew data from the simulator are presented in detail in several reports

[1] The opinions expressed in this paper are those of the authors and not those of the USNRC or of the authors' organizations

[*] jafores@sandia.gov
[+] Currently with ERIN Engineering and Research Inc., Walnut Creek, CA

[1-4]. A brief summary is provided below. Additional papers addressing different aspects of the study are included in this conference [5-7].

A number of organizations from ten countries participated in the study; these include industry, regulators, and the research community. The U.S. NRC in particular played a major role in supporting the preparation and execution of the study.

2. SUMMARY OF THE STUDY METHODOLOGY

The study utilized a set of data generated during a large-scale HAMMLAB experiment in 2006. Fourteen crews of licensed pressurized water reactor (PWR) operators performed four experimental trials each, namely base and complex conditions for both a Steam Generator Tube Rupture (SGTR) and Loss of Feedwater (LOFW) scenario. Each crew consisted of a Shift Supervisor, Reactor Operator and Assisting Reactor Operator. Although a turbine operator is also normally present on crews at the plant, they were not included on the crews in this study. The HAMMLAB PWR simulator is a full-scope simulator of a French plant (CP0 series) using a computerized human-machine interface. The HAMMLAB PWR procedures are based on the procedures at the participating operators' home plant. A total of 9 HFEs were defined for the two SGTR scenarios and 4 HFEs were defined for the two LOFW scenarios. In all but one case, the base and complex scenario included matching pairs of HFEs corresponding to the same tasks in the two scenario variants. In the SGTR scenario, the conditions for the complex case precluded the need for one action modelled in the base case. In addition to data collected during the simulation, DVDs were made of the crews performing in the simulator and these were reviewed to determine the final characterization of the crews' performance.

Thirteen HRA teams using thirteen HRA methods participated in the study. Two teams used the same method (SPAR-H), and one team used two different methods. The HRA teams were provided an information package that included the scenario and HFE descriptions, relevant procedures, information about the simulator, information on the operating crews, and other HRA related information. Further, the HRA teams requested and received additional information in a question-and-answer process, with all HRA teams receiving all questions and all answers. Thus, all HRA teams had access to the same information as a basis for their predictions about crew performance and human error probabilities. The HRA teams were asked to deliver:

- their 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 driving factors (PSFs), and 3) "operational expressions" or stories.
- the "normal" documentation of their HRA analysis and quantification, as in a PRA.

In one form or another, all HRA methods evaluate factors that can influence crews' performance in determining HEPs. The most important influences or factors affecting crew performance are sometimes referred to as the factors "driving" performance, the "driving factors" of performance, or the main "performance shaping factors" (PSFs). Comparing the specific factors or PSFs identified as driving factors for the defined HFEs by the HRA teams based on their method, with those observed in HAMMLAB, is a main focus of the comparisons performed for this study. In addition, the HRA teams were asked to provide a description of what they thought would occur operationally during the scenario runs (i.e., how the crews would respond in operational terms, what problems they might encounter, and what would be influencing their behaviour). These descriptions are referred to as operational stories or expressions.

The empirical simulator data, which are compared to the outcomes predicted by the HRA teams, describe the performance of the participating crews on the required actions (as defined for each HFE) in the study's scenarios. In the Halden data analysis, the individual crew performances were first analyzed to arrive at an integral understanding of each crew's performance. In a second stage, the integrated, summary data at the individual crew level were analyzed and combined to describe the performance at the aggregated (all crews) level. The aggregated performance of the HFE related

actions by the crews is described in three ways, which correspond to the ways in which the HRA teams were asked to report their predictions. These are namely:

- Performance on the HFE related actions expressed in operational terms ("operational descriptions")
- Assessment of the PSFs (main drivers) for each action.
- Number of crews failing to meet the success criteria for each action and an assessment of the difficulty of the action

The PSFs evaluated included adequacy of time, time pressure, stress, scenario complexity, indications of conditions, execution complexity, training, experience, procedural guidance, human-machine interface, work processes, team dynamics and communication. The selection and definitions of PSFs were based on the HRA Good Practices document (NUREG-1792) [8], but also included factors that the HAMMLAB analysts considered necessary to explain the behaviour of the crews in the simulator scenarios.

In addition, the HFE related actions were ranked relative to their difficulty. This evaluation was made by considering all available information on the performance of the tasks making up the actions. This implies that the ranking is not based on mere counting of 'failing crews'. Rather, the ranking took into account:

- The number of 'failing' crews and 'near misses'. Failures and near misses are the 'crews with operational problems' in performing the actions.

- Difficulty in operational terms. That is, which actions and the associated scenarios appeared to give at least some crews problems, even if they eventually met the response criteria defined for the HFE.

The final ranking was agreed upon by group consensus, where both experimentalists and the assessment group participated.

The outcomes predicted in the HRA analyses performed by the teams were compared with the outcomes obtained from the HAMMLAB experiments on several levels. Analytical predictions were compared with experimental outcomes for each of the following (the elements of response Form A):

- The level of difficulty associated with the operator actions of interest (with the HFEs). For the HRA predictions, the level of difficulty is represented by the HEP.

- The factors that most influence the performance of the crews in these scenarios (PSFs), called driving factors.

- The reason for the difficulties (or ease) with which the crews perform the tasks associated with each HFE, and how these difficulties are expressed in operational and scenario-specific terms ("operational expressions").

3. LESSONS LEARNED ON HRA METHODS

Based on both the qualitative and quantitative comparisons discussed above, a number of lessons learned about HRA methods were identified. These lessons learned take into account both the results of the SGTR scenarios and the LOFW scenarios, which turned out to be important since the HRA data patterns varied somewhat across the two sets of scenarios. However, as noted in the introduction, the discussions of the lessons learned in this paper must be considered a "snapshot," since detailed analysis of the results from specific HRA methods are ongoing and additional information may emerge.

3.1. Observations from Qualitative Analyses

3.1.1. Nature of the qualitative analysis

The nature of the qualitative analysis required to support the quantification is different from method to method and these differences sometimes impacted the results. For example, at one extreme, the qualitative assessment is focused on identifying failure mechanisms, including the contextual factors that enable them (e.g., CBDT, ATHEANA, CESA, MERMOS). At the other, it is focused on determining the strength of a PSF that is then used to modify a basic HEP, without an explicit assessment of the failure mechanisms (e.g., SPAR-H, THERP). For those methods that are based on the identification of failure mechanisms and context, the qualitative analysis performed tended to be richer in content than the PSF-focused methods and 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 more accurate HEPs, so other factors are involved.

A related issue concerns the extent to which methods might require or imply the need for the use of a job task analysis (e.g., THERP) as a qualitative analysis. However, not all methods do. For those that do so, the guidance does not necessarily suggest consideration of the cognitive demands in connection with the execution of a task, such as the interpretation of cues, interpretation of procedural criteria, and monitoring of relevant plant parameters. The lack of consideration of cognitive activities was most clearly discernible in the SPAR-H, ASEP, and CBDT applications of the SGTR scenarios. Each of these methods includes its own approach to addressing the cognitive aspect of a task, but in this benchmark, these applications modelled several of the HFEs subsequent to the initial HFEs as purely task oriented. For example, for some HFEs, the SPAR-H and ASEP analyses did not include the explicit diagnosis contribution to the HEP; and in the EPRI CBDT analysis it was decided not to use the CBDT to estimate the HEP for some HFEs, but instead included only the execution contribution. This has an effect on the analyses in two ways: 1) A task analysis that addresses cognitive aspects would result in a greatly improved HRA analysis for many HFEs in all these methods (related to the paragraph above). 2) The classification of a task as being only execution-oriented rather than also being of a cognitive nature has a direct impact on the assessment of the HEP itself. This is particularly evident for SPAR-H, which uses a base HEP for diagnosis actions that is ten times the base HEP for execution, and therefore, its inclusion has a significant impact on the resulting HEP. The empirical data did show that for some of the HFEs for which the cognitive aspects were not addressed by these methods, there was some cognitive activity (e.g., monitoring level, temperature, and pressure, choosing a response strategy) that had an impact on the effectiveness of response. Although the performance of these cognitive activities by the crews did not necessarily result in failure as defined by the success criteria, delays or difficulties were sometimes associated with the performance of these activities. These observations highlight the importance of addressing this aspect of the response, since they do indicate that the cognitive aspects could under other circumstances contribute to HFE failure.

The performance of a task analysis, in particular one that includes cognitive tasks, can also be useful to identify potential recovery mechanisms. A good example from the study involved an HFE where taking into account the primary indication, i.e., the indication that the PORV is closed (which was faulty), would lead the crews to conclude that no action was necessary. However, a recovery path existed through the monitoring of RPV pressure. At least one of the method applications applied this recovery mechanism. However, in the benchmark exercise, the secondary indications were not strong enough to lead to a successful recovery within the time allotted for this response action. Nevertheless, it is an indication that such a recovery should be considered. Similarly, another HFE was moderately challenging because of the need to monitor plant parameters while executing the procedure. The implication is that a richer qualitative analysis could help understand that concurrent behaviours could affect the likelihood of errors.

Although there may be several contributing factors, an interesting finding from the study was that, when the results of the different methods are taken together, the HEPs for the LOFW scenarios

suggested a tendency towards pessimism whereas some the HEPs from the SGTR scenarios showed some optimism (in other cases, the reference data is inconclusive). This occurred even though the HFEs in the LOFW scenarios were not obviously more difficult. The LOFW analyses were completed after the SGTR analyses and analysts had the benefit of knowing the results of the SGTR comparisons before submitting their final LOFW analyses. This led at least some HRA teams to make sure they addressed the cognitive portion of the actions (which appeared to be important based on the results) and more generally to do a better qualitative analysis, even given the limitations of the method guidance. While it could not be determined that the increased pessimism (and possibly more realism) was necessarily due to the improved qualitative analysis in all cases, it did appear to be related (but treatment of dependencies [section 3.1.4] may also have played a role in some cases). The improved qualitative analyses performed by some methods certainly led to better predictions of the operational stories for some of the LOFW events (e.g., CBDT and ASEP). Another possibility is that the HFEs in the SGTR scenario challenged the crews and the HRA methods in different ways than the HFEs in the LOFW scenario. For instance, the implementation element of the SGTR HFEs involved more control of the plant and, as noted, were more cognitively demanding.

An important conclusion that can be drawn is that the guidance for performing a qualitative assessment that is systematic and thorough enough to provide a meaningful assessment of the PSFs or other method-specific influencing factors appears to be inadequate for most methods. It was clearly important for the HRA teams to understand the context for the crew actions, including its dynamic aspects, and adequate guidance is not always provided. One of the consequences of this lack of guidance is the risk of a lack of reproducibility and traceability of the analysis, along with concerns about the validity of the results.

3.1.2. Issues Associated with Judgments about PSFs and Important Factors

The results of the study suggest that not all HRA methods cover an adequate range of PSFs or causal factors in attempting to predict operating crew performance for all circumstances. In other words, there were important aspects of accident scenarios that were not captured by consideration of the PSFs. The different methods often have somewhat different and limited sets of PSFs or causal factors that they normally consider, and even if they do perform a broader qualitative analysis in evaluating likely crew performance, there is limited guidance in the methods for how to translate such information into HEPs. Clearly, this could produce variability in results both within and across methods

Similarly, for PSF methods (e.g., SPAR-H, THERP) and other methods (e.g., CBDT) where judgements about the specific levels of factors (e.g., high vs. low workload) relative to a given scenario/HFE must be made, variability in making those judgments can occur and this can lead to variability in results within and across methods. The present study only had one case where a single method was used by two different teams (SPAR-H), but in a couple of cases the methods were similar (e.g., NRI DT+ASEP and CBDT [+THERP], along with ASEP and THERP/ASEP). 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 be important. Of course, other judgment issues could also be important to the methods oriented to failure causes or mechanisms, such as ATHEANA.

Due to these issues, it would appear that additional guidance is needed for making judgments on the level or strength of a PSF relative to an HFE, determining which PSFs are the most relevant, and determining how to integrate the role of factors not explicitly covered by a method in determining HEPs.

3.1.3. Crew-to Crew Variability

Evidence from the study indicated crew characteristics such as team dynamics, work processes, communication strategies, sense of urgency, and willingness to take knowledge-based actions can

have significant effects on individual crew performance. In addition, different crews may adopt different operational strategies or modes to address the scenario conditions, which may result in different scenario evolutions. The effects from these factors can be positive for some crews and negative for others within the same accident scenario, depending on its characteristics. Crew-to-crew variability is not explicitly considered for many methods. Several methods (e.g., SPAR-H, ASEP, HEART, CBDT) consider the "average" crew characteristics. Time reliability curve (TRC) approaches (e.g., diagnostic curve of ASEP, HCR/ORE) by contrast can be interpreting the TRC as a reflection of the variability of crew performance which could include crew to crew variability. "Sub-scenario" or "detailed context" based methods (e.g., ATHEANA, MERMOS) could also address crew-to-crew variability in estimating the HEP if they chose to. In fact this option is possible with any other method by developing different HEPs for different PSFs that reflect the impact of crew characteristics, and performing a weighted sum of the HEPs. Of course, an approach for assessing crew characteristic related PSFs would need to be developed.

These factors are not normally evaluated by most HRA methods and even if they are, it is often difficult to observe enough crews to make reasonable inferences about the effects of such factors across crews (i.e., are the effects systematic across crews or is there a great deal of variability across crews that makes the effects hard to track). Nevertheless, it is clear that such factors can have an important impact in some scenarios and are certainly worth investigating in the context of an HRA when their impacts could be significant. Improved guidance for addressing such factors could improve HRA analyses, but for practical reasons, the effects may often have to be addressed using sensitivity analyses on the HRA results.

3.1.4. Treatment of Dependency

In the LOFW scenarios, HRA analysts were asked to estimate the HEP for the crews failing to initiate Bleed and feed (B&F) prior Steam Generator (SG) dryout (HFE 1A and 1B, with A and B referring to the simple and complex versions of the scenario) and for initiating B&F before core damage (HFEs 2A and 2B). If the crews succeed in initiating B&F prior to dryout, then the second HFE (2A or 2B) is irrelevant. However, if they fail to initiate prior to dryout, there will still be time available to complete the response before core damage (CD). The HRA teams were asked to quantify both cases (dryout and CD) and to provide the joint probability of the two events. Interestingly, different methods or applications addressed the dependency issue between the second case and the first case (i.e., 2A|1A and 2B|1B) in different ways.

Some applications computed a conditional probability for the second event given failure of the 1st, often using the THERP dependency model, and then multiplied the two probabilities to obtain the joint probability. One approach quantified the probability of failing to take both actions as one event (a joint probability) and then divided by the probability of failing the first action to obtain the conditional probability. Another approach computed the HEP for the first event and then calculated the second probability as a recovery action. That is, they applied a recovery value to the first event to produce the joint probability of the two events, with the recovery credit reflecting the probability of the second event.

One result observed is that the assumption of neglecting negative dependence in the THERP dependence model and dependence models derived from this model, may be overly conservative in some cases. (Negative dependence refers to cases where the failure of an HFE decreases the failure probability of a subsequent HFE.) In the observed performances of the complex LOFW scenario, all crews that failed the first HFE succeeded on the subsequent HFE (B&F prior to core damage). Several HRA teams thought that the THERP model was producing conservative values for these cases and modified the approach to obtain what they suspected would be more realistic conditional probabilities. Thus, although there may have been some positive dependency effects present, the THERP model may still lead to inappropriately high conditional HEPs in cases where negative dependence is also present. Given failure on the 1st event, the THERP model can only leave the probability of the second HFE

unmodified or increase it due to (positive) dependence factors. The main lesson learned from these observations is that HRA methods need better guidance for addressing dependency in HRA.

3.2. Observations from Quantitative Comparisons - Trends in the HEP Predictions

Despite the care taken to provide a detailed description of the scenarios and definitions of the HFEs, along with consistent information to all HRA teams, the HEPs provided by the HRA teams show significant variability from method to method (see companion paper on quantitative results for this conference [5]) for both the SGTR and LOFW HFEs. The variability was present for both the easy HFEs (i.e., those with expected low HEPs) and the difficult (i.e., those with expected high HEPs). The variability is not correlated across the HFEs in the sense that the same HRA method did not result in consistently producing the highest (or the lowest) HEP for the set of HFEs. In other words, none of the methods were systematically more conservative or optimistic than the other methods. In addition, the ranking of the HEPs was not consistent from method to method. Compared to the actual performance of the crews and the uncertainty bounds from the Bayesian analysis [5]), the HEPs for the SGTR scenarios appeared to tend somewhat toward optimism, while those for the LOFW seemed to be more pessimistic or conservative about the probability of failure.

Another thing to note about the methods is that many of the applications did not exhibit much variation among the HEPs; in other words, the range of HEPs for the set of HFEs was rather narrow, in some cases, less than an order of magnitude, even though many of the HFEs differed significantly in difficulty. In general, this was more of a trend in the SGTR scenarios than in LOFW. The method applications that resulted in little variation among the HEPs in the SGTR analysis also appeared to provide optimistic assessments of the HEPs associated with the HFEs assessed to provide the greatest challenge, when compared with the HEPs provided by other method applications. However, in the LOFW scenarios, there was better overall differentiation among the different HEPs and no systematic tendencies toward optimism (if anything they tended toward pessimism). Thus, it would not appear that the trends toward optimism in the SGTR scenarios for some methods necessarily reflect inherent characteristics of the methods.

3.3. Understanding the Sources of Variability Among Methods

Variability should not be unexpected since the methods have very different theoretical bases and approaches for quantification. For example, there are differences related to:

- Whether failure mechanisms are identified and at what level of detail. Those methods that address these at a fairly detailed level include, for example, ATHEANA [error forcing context and unsafe acts], MERMOS [stories], CBDT [failure mechanisms]
- The use of generic failure types, for example, CREAM, HEART
- Whether the methods decompose tasks, for example, as THERP and ASEP do
- The scope of the PSFs that are addressed and the scales for these, for example, SPAR-H and CREAM

Given the differences in the methods, the factors that can affect the variability in predictions can be grouped into the following types:

Method Driven

These include:

- The capability of the method to capture the significant influences on behaviour
- The depth of qualitative analysis required by the method, and the degree to which it leads to an understanding of the underlying dynamics of the scenario and driving factors.
- Any inherent pessimism or optimism of the method

- The capability of the method to accommodate the analysts' knowledge and understanding in a way that allows a characterization of the relative difficulty of the actions associated with the HFEs

Analyst Driven

These include:

- Whether the method has been applied as intended
- The depth of 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 has limited capability to cast light on all of these factors. Certainly the last two items are not easily testable. In general, it is difficult to distinguish between the effect of the method and the effect of the analysts. A different study would be required to validate many of these aspects. In this study, we have focused on investigating the analyses by qualitative means in order to cast light on possible strengths and weaknesses of the methods that have enabled or hindered analysts to make good analyses.

While the quantitative comparisons have also contributed important information about the methods, the results of this simulator experiment are not directly able to validate the assessment of human error probabilities, for the following reasons.

- The definitions of failure for the purposes of identifying failures in the empirical data were not necessarily defined in the same way as they would be for a PRA. In a PRA, failure would be defined as failure to perform the required action in time to prevent an irreversible change in plant state. In the experiment, failure was sometimes defined in terms of a somewhat arbitrary time, which was based on reasonable expectations of crew performance based on their training. The HRA teams understood they were trying to predict performance with respect to the corresponding time window, but use of these failure criteria may have been a little confusing.
- The empirical HEPs were estimated on the basis of a sample of at most 14. When there are a significant number of observed failures, this can provide a reasonable estimate of the failure probability. However, for many of the HFEs there were no observed failures, and therefore it is not possible to derive reliable HEP estimates and therefore a reliable empirical ranking of the HEPs on purely statistical grounds.

However, an attempt was made to use all the evidence from the experiment to assess the relative challenge that the actions would pose to the operators; this was used to rank the HFEs with respect to difficulty. It was taken as a premise that the ranking with respect to difficulty should be reflected in the methods' predicted ranking of the HFEs based on their HEPs, and this information did prove useful.

4. CONCLUSION

The International Empirical Study has confirmed that simulator exercises that are well designed with extensive documentation and analysis can provide significant insights to support HRA method benchmarking and development. Most of the insights were derived from assessing: 1) whether the methods have the capacity to identify operational details of the performance of the required actions, and 2) whether they have the ability to use this information in the evaluation of the HEPs in such a way that they reflect the difficulty associated with the performance of the associated actions.

Based on the lessons learned described above, it is clear that the qualitative analysis performed to support HRA quantification is an important contributor to the adequacy of HRA predictions. The various methods vary significantly in the nature and degree of the qualitative analysis performed. While a good qualitative analysis (including a task analysis) is a relative strength of some methods, it is clear that all of the methods could use improvement in this area. This conclusion is based on a number of findings which were discussed above, but the main one is that even the methods with strong guidance for qualitative analysis did not always provide acceptable predictions of HEPs. Nevertheless, it was shown that without a good qualitative analysis that covers a thorough set of conditions and influencing factors, the methods have an inadequate basis for their predictions. This was particularly demonstrated when method applications did not address the cognitive aspects of performance in implementing procedures even though the initial diagnosis had been completed.

While a number of areas where qualitative analysis could be improved were discussed in Section 3 above, future empirical studies should take steps to obtain additional information. One limitation of the present study was that the experimental design made it difficult to separate method vs. 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 the specific aspects of the different methods qualitative analysis that lead to shortcomings in their predictive validity.

Acknowledgements

The authors gratefully acknowledge the contributions of Helena Broberg, Salvatore Massaiu, Michael Hildebradt, and Per Oivind Braarud, the Halden Project, Ron Boring, Idaho National Lab, Pamela Nelson, Universidad Nacional Autónoma de México, and Ilkka Männistö, VTT for major parts of the experimental work done so far in the project. The work of the HRA teams, 13 teams providing 13 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.

This study is a collaborative effort of the Joint Programme of the OECD Halden Reactor Project and in particular Halden's signatory organizations who provided the analysts teams, the U.S. Nuclear Regulatory Commission (USNRC), the Swiss Federal Nuclear Inspectorate (DIS-Vertrag Nr. 82610) and the U.S. Electric Power Research Institute. In addition, parts of this work were performed at Sandia National Laboratories with funding from the USNRC. 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 opinions expressed in this paper are those of the authors and not those of the USNRC or of the authors' organizations.

References

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

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

[3] A. Bye, S. Massaiu, V.N. Dang, J. A Forester, *The International Empirical Study – Utilizing HAMMLAB Simulator Data to Evaluate HRA Methods,* OECD NEA Workshop on Simulator Studies for HRA, Budapest, Hungary, 4-6 November, 2009 (proceedings to be published in 2010).

[4] V.N. Dang, A. Bye, E. Lois, J.A. Forester, Per Øivind Braarud. *Benchmarking HRA Methods Against Simulator Data – Design and Organization of the International HRA Empirical Study,* Proc. 9th Int. Conf. on Probabilistic Safety Assessment and Management (PSAM9), Hong Kong, China, May 18-23, 2008, CD-ROM (2008).

[5] V.N. Dang, A. Bye, J.A. Forester, S. Massaiu, *Quantitative Results of the HRA Empirical Study and the Role of Quantitative Data in Benchmarking*, Proceedings of the 10th Int. Conf. on Probabilistic Safety Assessment and Management (PSAM10), Seattle, WA, June 6-11, 2010, CD-ROM (2010).

[6] H. Broberg, S. Massaiu, J. Julius, and B. Johansson, *The International HRA Empirical Study: Simulator results,* Proceedings of the 10th Int. Conf. on Probabilistic Safety Assessment and Management (PSAM10), Seattle, WA, June 6-11, 2010, CD-ROM (2010).

[7] R.L Boring et al., *Lessons Learned on Benchmarking from the International HRA Empirical Study,* Proceedings of the 10th Int. Conf. on Probabilistic Safety Assessment and Management (PSAM10), Seattle, WA, June 6- 11, 2010, CD-ROM (2010).

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