Lessons Learned from the US HRA Empirical Study

Huafei Liao[a*], John Forester[a,b], Vinh N. Dang[c], Andreas Bye[d], Erasmia Lois[e], Y. James Chang[e]

[a] Sandia National Laboratories, Albuquerque, NM, USA
[b] Idaho National Laboratory, Idaho Falls, ID, USA
[c] Paul Scherrer Institute, Villigen PSI, Switzerland
[d] OECD Halden Reactor Project, Institute for Energy Technology, IFE, Halden, Norway
[e] U.S. Nuclear Regulatory Commission, Washington, DC, USA

Abstract: The US Human Reliability Analysis (HRA) Empirical Study (referred to as the US Study in the article) was conducted to confirm and expand on the insights developed from the International HRA Empirical Study (referred to as the International Study). Similar to the International Study, the US Study evaluated the performance of different HRA methods by comparing method predictions to actual crew performance in simulated accident scenarios conducted in a US nuclear power plant (NPP) simulator. In addition to identification of some new HRA and method related issues, the study design of the US Study allowed insights to be obtained on some issues that were not addressed in the International Study. In particular, because multiple HRA teams applied each method in the US Study, comparing their analyses and predictions allowed separation of analyst effects from method effects and allowed conclusions to be drawn on aspects of methods that are susceptible to different application or usage by different analysts that may lead to differences in results. The findings serve as a strong basis for improving the consistency and robustness of HRA, which in turn facilitates identification of mechanisms for improving operating crew performance in NPPs.

Keywords: HRA, simulator data, nuclear power plant, performance shaping factor.

1. INTRODUCTION

As an effort to improve the robustness of human reliability analysis (HRA), the US Nuclear Regulatory Commission (NRC) participated in and supported the International HRA Empirical Study [1-4] (referred to as the International Study hereafter), in which HRA predictions of different analysts and methods were compared to observed crew performance data at Halden Reactor Project's HAMMLAB (HAlden huMan-Machine LABoratory) simulator facilities. Many HRA methods were tested in the study; however since, with one exception, there was one HRA team applying each method, this limited our ability to make decisive conclusions concerning how the HRA analysts' applications of a given method could contribute to the variability in HRA results and how the methods themselves contributed to the variability.

In contrast, the HRA methods evaluated in the US HRA Empirical Study [5-7] (referred to as the US Study hereafter) were applied by multiple HRA teams. This allowed us to obtain new insights on factors contributing to variability in HRA results as well as strengths and weaknesses of the chosen HRA methods. Another notable difference from the International Study was that the HRA analyst teams in the US Study were able to visit the plant, observe a crew in a simulator training scenario, and interview training personnel to collect information needed to perform HRA. Third, the US Study was performed on a US nuclear power plant (NPP) training simulator, which allowed us, to some extent, to evaluate the generalizability of Halden human performance studies to US applications.

In the following sections, the study methodology, simulator data, and predictive quantitative results will be presented. Then, the findings on the contributing factors to variability in HRA results will be discussed.

[*] Corresponding author.
Email address: hnliao@sandia.gov

2. STUDY METHODOLOGY OVERVIEW

The US Study capitalized on the design and methodology of the International Study. It focused on control room personnel actions required in the response to initiating events typically modelled nuclear power plant (NPP) probabilistic risk assessments (PRAs). Three scenarios were developed and five human failure events (HFEs) were defined (see [5] for detailed description of scenarios and HFEs). Scenario 1 was a total loss of feedwater (LOFW) followed by a steam generator tube rupture (SGTR), for which three HFEs (HFEs 1A, 1B, and 1C) were defined. Scenario 2 was a loss of component cooling water (CCW) and reactor cooling pump (RCP) sealwater, for which one HFE (HFE 2A) was defined. Scenario 3 was an SGTR scenario without further complications, in which one HFE (HFE 3A) was defined. HFE 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 many HFEs in standard PRA scenarios. The reason for this is that although the HFE success criteria used in the study should relate to those commonly used in PRA/HRA, for the purposes of this study, it was important that they be clearly observable in the simulated scenarios.

Four crews of five licensed operators from a participating US NPP simulated the scenarios on the plant full-scope pressurized water reactor (PWR) training simulator. Their performance data were analyzed and described in the following three ways.

- Performance on the HFE related actions expressed in operational terms ("operational descriptions");
- Assessment of the performance shaping factors (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

Nine HRA teams participated in the US Study and each team applied an HRA method to predict performance relative to the HFEs defined in the study. Two teams used ATHEANA (A Technique for Human Event Analysis) [8], two teams used SPAR-H (Standardized Plant Analysis Risk-Human Reliability Analysis) [9], two teams used ASEP ((Accident Sequence Evaluation Program Human Reliability Analysis Procedure) [10], two teams used the EPRI HRA Methodology (implemented with the HRA Calculator version 4.1.1) [11] and used CBDT (Cause-Based Decision Tree) [12] and HCR/ORE [12] for the diagnosis portion of the response and THERP (Technique for Human Error Rate Prediction) [13] for response execution, and one team used a hybrid CBDT+THERP+ASEP method with similarities to the EPRI HRA method. The predictions were compared to crew performance data to assess the predictive power of the method applications and examine aspects such as traceability, method guidance, and insights for error reduction.

Predictive power was assessed in terms of qualitative predictive power and quantitative predictive power. Qualitative predictive power was evaluated based on a comparison of predicted operational expressions and observed operational descriptions, as well as a comparison of predicted PSFs and observed performance drivers. Quantitative predictive power was assessed in terms of the absolute values of the HEPs predicted by each HRA team and the ranking of the HFEs based on the magnitude of the predicted HEPs. Given the small sample of observations, the accuracy of the predicted HEPs is difficult to assess. Therefore, although the quantitative predictive power was assessed to the extent possible in light of the small number of data points, method assessment was conducted primarily from a qualitative analysis perspective. In practice the assessment of quantitative predictive power were performed as a prelude to analysis of qualitative issues.

As mentioned above, the HRA methods in the US Study were applied by multiple HRA analyst teams. Therefore, the predictions of the different teams using the same method were compared to separate analyst effects from method effects to the extent possible and gain insights on the factors that can

contribute to variability in HRA results. The intra-method comparison focused on the differences in qualitative predictions, quantitative results, analysis approaches and assumptions.

3. EMPIRICAL AND PREDICTIVE QUANTITATIVE RESULTS

Simulator data were collected for four HFEs. The crew failure rates and HFE difficulty ranking are listed in Table 1. The difficulty ranking of the HFEs was determined in terms of the crew failure rates and the challenges experienced by the crews in diagnosing plant status and executing manual actions to bring the plant to a stable state.

Table 1. Crew Failure Rates and HFE Difficulty Ranking

HFE	Failure Rate	Difficulty Ranking
HFE 2A	4/4	1 (Very difficult)
HFE 1C	3/4	2 (Difficult)
HFE 1A	0/4	3 (Fairly difficult to difficult)
HFE 3A	0/3	4 (Easy)

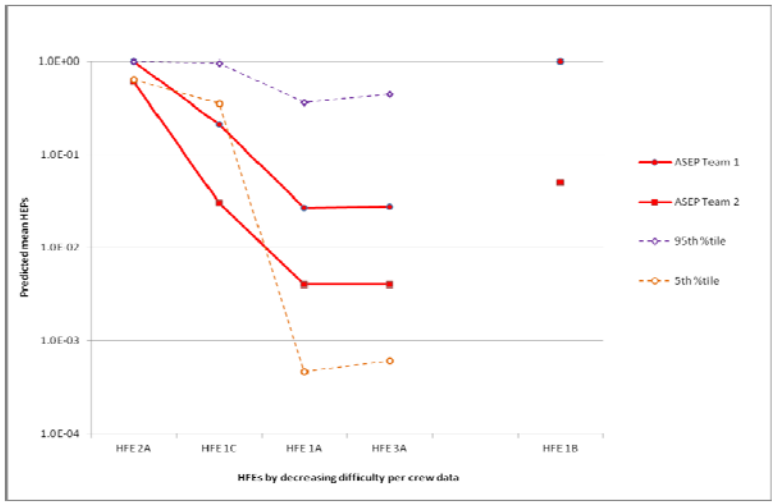
* No data were collected for HFE 1B.

The mean HEPs predicted with each method are presented in Figure 1 alongside the Bayesian uncertainty bounds derived from the simulator data with a non-informative prior (Jeffrey's prior). The HFEs are ordered by their difficulty ranking on the horizontal-axes, and the HEPs are presented on the vertical-axes in a logarithmic scale. The following observations are made from the HEP curves (see [5-6] for more detailed discussion).

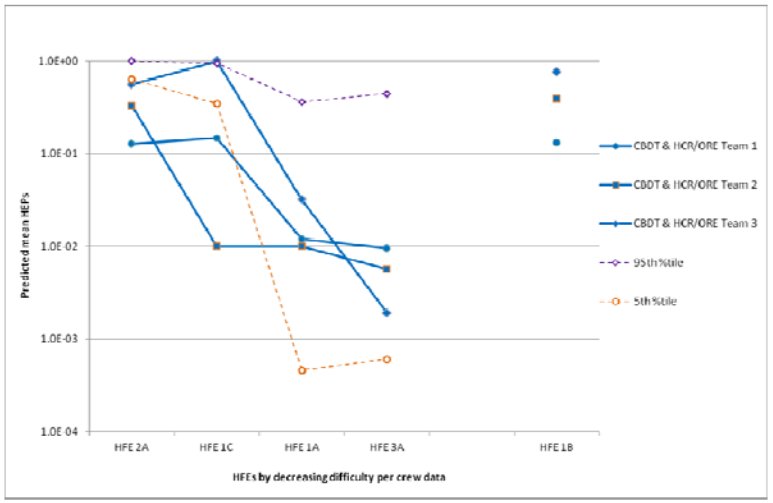
- Overall, most methods identified the correct HFE difficulty ranking with a few exceptions in terms of the relation between HFEs 2A and 1C.
- For most HFEs, there is about one order of magnitude or less difference across the HRA analyst teams using the same method, especially if we consider the HEP of HFE 1C predicted by the CBDT & HCR/ORE Team 2 was caused by a misunderstanding of the HFE definition.
- Across the four methods, it seems that ASEP, ATHEANA, and CBDT & HCR/ORE produced relatively more consistent quantitative results than the SPAR-H.
- Many teams underestimated the most difficult HFE 2A. This seems to be caused either by insufficient qualitative analysis to understand the scenario dynamics, or by inappropriate assumptions or interpretations based on information obtained from the interviews with plant personnel.
- Although all of the HRA analyst teams concluded that HFE 3A was the easiest, there is significant variability in the HEPs with a couple of estimates being much lower than the other HEPs. This seems to indicate that there is a lack of consensus in HRA in terms of what the baseline HEP for generally good conditions should be.

It could be argued that with the exception of HFE 3A, there was less variability in the predicted HEPs across the methods in the US Study compared to the International Study. In addition, the difference in the within method HEP predictions was less than might have been expected based on general HRA performance in the International Study and on the couple cases where similar methods were used. That is, one might argue that the analysts using the same methods in the US Study did a relatively good job in many cases and often corresponded relatively well in their predicted difficulty rankings of the HFEs. Potential reasons for these results include the following:

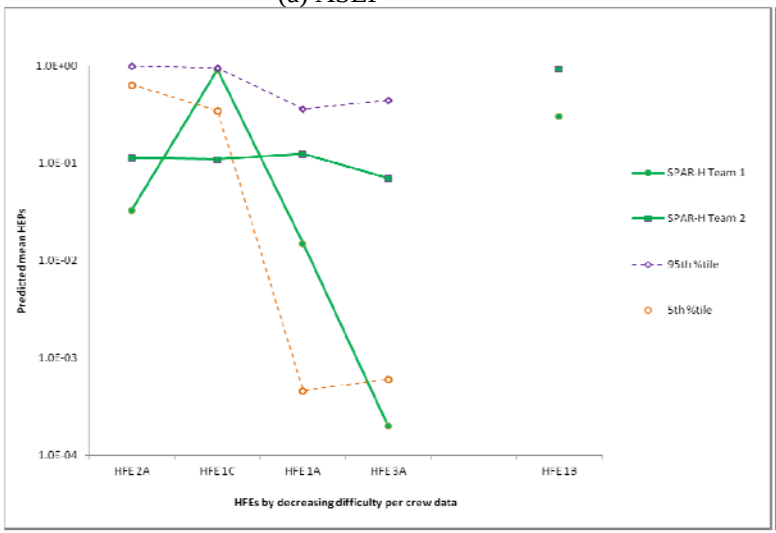
- There may have been some learning effects between the International Study and the US Study. Some of the HRA analysts participated in both studies and most participants in the US Study were familiar with the results of the International Study. Thus, the lessons learned may have improved the HRA team's applications in that they had a better idea of what they needed to do to perform a better analysis with the method they were using. There also appears to be some evidence of the learning effects between the different phases of the International Study.



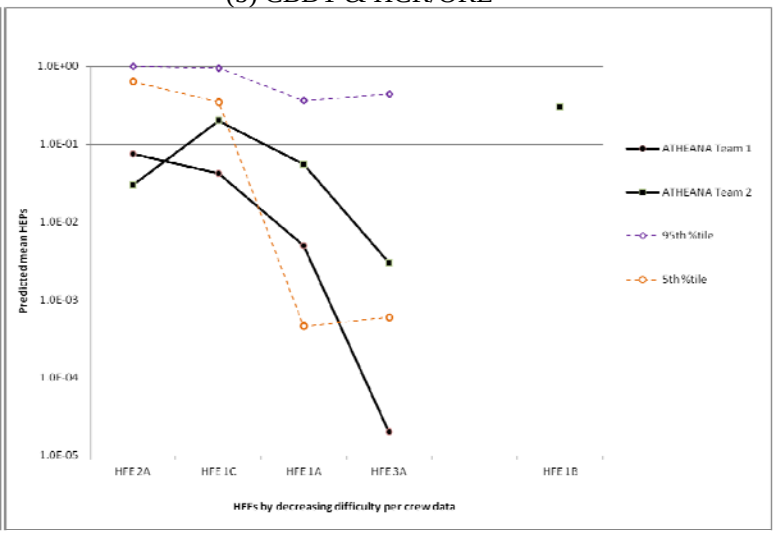
(a) ASEP



(b) CBDT & HCR/ORE



(c) SPAR-H



(d) ATHEANA

Figure 1. Predicted Mean HEPs by HRA Methods with Bayesian Uncertainty Bounds

- It is also possible that the HRA teams were, in fact, better at predicting the performance of US crews on a US simulator than the performance of foreign crews. That is, their previous experience with how US crews work and interact with procedures etc. may have facilitated their performance.
- Similarly, their ability to visit the plant, interact with plant personnel, and obtain the usual HRA related information may also have contributed to improved performance.

4. FACTORS CONTRIBUTING TO HRA PREDICTIVE DIFFERENCES

The International Study identified that significant variability can occur in the results of applying different HRA methods for the same HFE due to the differences and limitations in the methods' technical and methodological bases [1-4] and in the associated guidance provided for applying the methods. In addition, although it appeared that the analysts themselves could contribute to differences in results (analyst driven factors), the design of the International Study did not allow clear separation of method effects from analyst effects because there was only one case where two teams used the same method. With at least two teams per method, the US Study was able to identify that in addition to inherent method-driven factors, analyst-driven factors and the interactions between the analyst-driven factors and method-driven factors can also cause significant variability in the HRA results. Overall, the US Study revealed that a major source of variability across analysts using the same method was analyst decisions about how to apply various aspects of the method. Analysts are often called upon to make decisions in their analyses, and the guidance of the HRA methods are not sufficient or specific enough, so that analysts may, to some degree, have to rely on their own subjective interpretation of the guidance. Moreover, the methods sometimes allow analysts to apply different analysis or modeling options without providing clear criteria for when to use the different options. The factors leading to analysts' subjectivity and thus contributing to HRA predictive differences are discussed below.

4.1. Factors Contributing to Differences in Qualitative Analysis

- *Differences in qualitative analysis approaches, scope and depth.* One reason for the differences is the degree to which a specific method framework specifies or guides the qualitative analysis process. Some methods (e.g., SPAR-H) do not provide specific or complete guidance for performing the qualitative analysis, and thus HRA analysts are left to decide how they will perform the qualitative analysis and its level of detail. Although some other methods (e.g., ATHEANA) support the search and treatment of a more comprehensive set of performance drivers at a more detailed level, the guidance is somewhat open-ended and not always well-structured for translating the information into HEPs in a consistent manner; thus, to some extent leaving the level of detail up to the analysts. Another reason is the analysts' level of effort devoted to their qualitative analyses. Some analysts may undertake a detailed qualitative analysis by even going beyond method guidance as they see necessary. Although more detailed qualitative analyses tend to lead to a better understanding of scenario dynamics, it does not necessarily lead to improved quantitative predictive performance.
- *Differences in task decomposition approaches.* Most HRA methods do not have a consistent approach or provide much guidance for task analysis and decomposition. Insufficient task analysis to understand scenario conditions, procedural interactions, and key manual actions may cause analysts to fail to appreciate the task complexity, especially for complicated scenarios. This can lead to different task groupings and dependency modeling, which, in turn, can affect quantitative results. It can also cause analysts to ignore cognitive activities involved in step-by-step actions. In addition, the level of task decomposition may impact the application and traceability of the quantitative analysis.
- *Timeline analysis.* Timing is either explicitly or implicitly considered in various HRA methods; hence the uncertainties in timing analysis can affect HEP estimates. For methods strongly based on time reliability correlations (TRCs) (e.g., ASEP and HCR/ORE), HEPs are estimated as a function of the time available for operators to respond to accident scenarios and

the HFEs of interest. Considering the characteristics of TRCs, the HEPs can be sensitive to timing analysis results, particularly for the HFEs that are time critical, that is, small differences in time windows can lead to large differences in HEPs. For some methods that are not TRC-based (e.g., SPAR-H), the availability of time is often considered as a PSF. The differences in the rating of this PSF can also lead to variability in HEPs, which can be as large as that obtained with TRCs. The variability in timing analysis can occur because: (1) analysts need to make decisions on which procedural steps to include in the timing analysis; (2) some methods do not provide specific guidance on assessing the values of timing points (e.g., how to assess the time required for an action); and (3) analysts' judgment is needed to account for the impact of distractions, delays, and competing task demands on timing analysis.

- *Difficulty in understanding and treating complexity*. HRA methods differ in their ability to deal with complexity. It can be difficult for the analysts to identify and treat the issues in complex scenarios that are beyond the scope of a given method. Although narrative-based methods (e.g., ATHEANA) appear to present some advantages over PSF-based methods in addressing complex HFEs, there are some aspects that are still left to the analysts. While the analysts can always go beyond the methods in qualitative analysis for HFEs that are broader than the scope of the factors explicitly treated by a given method, such an extended analysis is subject to variability because the analysts will determine the appropriate scope of performance drivers without a common basis for this judgment. Additionally, as mentioned above, the results of this extended analysis may not be easily incorporated into the quantification portion of the method to produce improved quantitative predictions.

4.2. Factors Contributing to Differences in Quantitative Analysis

- *Judgment about when to credit recovery*. Although some methods provide an option to credit error recovery, it is often the analysts' decisions whether or not to consider it in an HRA application and whether the criteria are adequately met. Some analysts may choose not to credit recovery just for a preference toward a more conservative result. Different biases between analysts toward obtaining conservative results can obviously lead to variability in results. This type of variability is not necessarily a problem as long as the bases for the decisions are made clear and analysts have the opportunity to follow up on the estimates and do more detailed and realistic analysis for important contributors.
- *Attempt at compensation for inadequate range of PSFs*. It has been observed in the US Study that in some situations the PSF-based methods were inadequate in coping with some HFE-specific performance drivers identified in the crew data simply because they were not addressed by the method. As a result, some analysts had to compensate for such limitations by stretching the method to fit the situation based on their experience. This led to quantitative differences in HRA results between the applications. This finding suggests that to be able to reliably predict performance, the PSFs need to have a sufficient coverage. Nevertheless, as was also shown and discussed in the International Study, the US Study showed that the PSF-based methods sometimes produced reasonable HEPs without identifying all performance drivers, particularly for the easy HFEs. It could not be determined whether this reflects an inherent characteristic of the methods or whether it was just a coincidental effect.
- *Decision on PSF selection and rating*. For some methods (e.g., SPAR-H, ASEP and CBDT), judgment about the relevance of a particular factor and the specific level of that factor in a given scenario must be made, and for others (e.g., ATHEANA) the analyst must determine what factors are present and characterize them, including the strength of their impact. Definition overlap between PSFs and inadequate guidance on determining the PSF status can cause variations in analysts' interpretation of the scope of the PSFs and in the ratings assigned to the PSF for a given issue or performance condition. This underscores the importance of addressing such issues for PSF-based methods.
- *Inadequate coupling between qualitative analysis and quantification model*. A broad qualitative analysis in evaluating likely crew performance does not necessarily lead to HEPs that are consistent with the observed crew performance. For some methods, the guidance on quantification of the impact of PSFs on crew performance is limited, and to varying degree

left to analysts' judgment, particularly when the analysts' qualitative analysis goes beyond the method guidance. In addition, it seems that not all HRA methods cover an adequate range of PSFs to predict operating crew performance for all circumstances; as a result, analysts may have to rely on their judgment to decide how to integrate the role of factors not explicitly covered by a method in HEP estimation, which can obviously lead to variability in results. A good tie or dovetailing between the qualitative analysis and the quantification approach is important for consistency in the results of HRA applications within methods.

- *Efforts to compensate for poor treatment of diagnosis*. Methods that strongly rely on TRCs to quantify diagnosis (e.g., ASEP) appear to be poorly equipped to address the difficulties in operators' cognitive activities. In addition, it does not appear that all the HRA methods in the study are well equipped to address the full scope of cognitive activities related to operators' overall response to the scenarios. This can lead analysts to attempt to compensate for the method's shortcomings by doing more qualitative analysis than is directed by the method and trying to incorporate the information into the quantification approach by adjusting the method based on their own experience, which can introduce variability in results. Moreover, when the methods do address diagnosis, analysts tend to focus on operators' cognitive activities in understanding the plant situation to decide the appropriate response plan (i.e., initial diagnosis). The cognitive activities in executing the response plan are typically omitted. This has significant effects when the response plan is more complex than simple skill-of-the-craft. The study results have showed that inadequate consideration of operators' cognitive challenges in working through procedures can lead to failure in identifying important performance drivers and result in underestimations of HEPs. Although analysts may compensate for methods' inadequacy in diagnosis treatment based on their experience, it can lead to quantitative differences in HRA results.

4.3. Factors Related to HRA Practices

- *Different levels of reliance on interview information*. Information from interviews with plant personnel was used for several purposes, such as estimating the time required for diagnosis and execution and evaluation of training and experience. Most of the time, the information provided valuable insights for analysts to understand the scenario dynamics. However, the analysts differed in the extent to which the interview information was used in their analyses. Some analysts tended to rely on input directly from the interviews while others tended to rely on their own analysis and judgment with interview information as a supplement. The US Study has shown that over-reliance on trainer or operator opinions can sometimes negatively impact HRA results. Furthermore, it also seems to suggest that the differences in plant personnel's opinions may increase with the increase of scenario complexity. Obtaining consensus among multiple trainers or operators and/or more detailed qualitative analysis may help reduce the effects.
- *Different approaches to plant personnel interview*. It was observed that the HRA analysts differed in the scope and level of detail of the qualitative scenario analysis conducted before the interviews. They also differed in their interview techniques. Some analysts focused on questions regarding whether operators would take some actions. In contrast, some analysts asked the plant personnel to do a detailed talk-through together with them and focused on questions regarding timelines and the interactions between operators and procedures. The US Study seemed to indicate that sufficient preparation before interviews and the talk-through based interview technique would provide a better understanding of scenario dynamics and complexity.
- *Reasonableness check of HEPs*. The HEPs resulting from the application of each HRA method in the US Study were assessed with respect to their relative values (rank order of HFEs by failure probabilities) and the overall differentiation among these probabilities. Overall, the HEPs showed reasonable differentiation. However, in some cases, the HEPs for the two most difficult HFEs (2A and 1C) were not consistent with the difficulty ranking and/or fell into a narrow range. Particularly, the HEPs of SPAR-H Team 2 did not show differentiation expected from their qualitative analysis. This suggested that the team did not

check the reasonableness of the obtained probabilities or performed an inadequate check in view of their qualitative findings. This is especially remarkable for their analysis of HFE 3A.

Although HEPs are often checked for reasonableness in external reviews of the PRA/HRAs, where each individual HEP cannot be reviewed in detail and emphasis is placed on the relative values of the HEPs, there appears to be little documented guidance on how to perform reasonableness checks. Some of the factors to be considered in terms of similarity and levels of challenge include:

- Available time (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 performance shaping factors, and failure modes identified for the HFE

Comparing the related HFEs (for the same tasks performed in different scenarios) or HFEs with similar performance conditions typically leads analysts to review their differences in HEPs to determine whether the HEPs correspond to 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 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.

5. FINDINGS ON HRA GENERAL ISSUES

There was significant agreement in the findings between the International and US Studies. The conclusions from both studies about HRA in general and the identified needed improvements are summarized below.

- *Consideration of cognitive activities.* Both studies agreed that the consideration of operator cognitive activities is an important contributor to the adequacy of HRA predictions. It is especially beneficial to understand the difficulties in operators' assessing the situations and/or making new response plans in complex scenarios. However, the US Study has also revealed that even when diagnosis is explicitly considered, the methods still show some limitations in the ability to assess crews' cognitive activities in order to adequately support the understanding and identification of failure mechanisms and the HEP quantifications, particularly for the more difficult HFEs.
- *Explicit guidance and framework to support structured and consistent qualitative analysis.* While a good qualitative analysis is a relative strength of some methods (e.g., ATHEANA), one conclusion from both studies is that qualitative analysis is a shared weakness across all methods. The variability across HRA applications is not unexpected given the differences in the technical bases and methodologies of the methods. However, it seems that the variability also has its root in the fact that the methods do not provide sufficient guidance or an explicit framework for analysts to conduct a structured and consistent qualitative analysis. This is clearly evidenced by the variability in timing analysis, HFE decomposition, the range of factors considered, and the treatment of complexity.
- *Method improvement and extended qualitative analysis to treatment of complexity*. There is a need for method improvement to cover a broader scope of performance drivers in all methods. Given that complex scenarios normally involve relatively more cognitive challenges than easy scenarios, one priority of method improvement should focus on providing means and frameworks for analysts to identify and characterize contextual factors and mechanisms that can cause failures at the cognitive level and provide a structured and systematic way to incorporate the information into the quantification process. However, it should be realized that HRA applications will always rely to some extent on analysts' experience and expertise.

Nevertheless, the goal should be to provide as much structure and guidance as possible to support analysts at differing levels of expertise.

- *Coherent coupling between qualitative analysis and quantitative model.* As discussed above, extended qualitative analysis can help analysts uncover scenario-specific performance drivers. However, it may not necessarily lead to appropriate HEP estimates in all cases because of the difficulties in translating qualitative analysis into HEP impact, especially for complex scenarios.
- *Adherence to good practices with improved guidance.* Improved guidance is needed for performing plant visits and personnel interviews. Additionally, there is little documented guidance on how to perform reasonableness checks on HEPs and the consistency between qualitative findings and quantitative results in terms of performance conditions.

6. ACHIEVEMENTS AND OVERALL CONCLUSIONS

The US Study and the International Study are two large-scale systematic data collection and HRA method evaluation efforts. The achievements from the studies are summarized below:

- The US Study and the International Study demonstrated the feasibility of using simulator data to evaluate HRA methods. The methodological tools developed in the studies, such as (1) the development of the experimental design, focusing on evaluating HRA methods, (2) the methodology for collecting simulator data, (3) the methodology for analyzing simulator data for PRA-type of scenarios and tasks, and (4) the methodology for data-to-method comparisons, were tailored to HRA needs and are proving to be very useful achievements. They have also demonstrated that important information on HRA and HRA methods can be obtained without using impractically large numbers of operating crews and scenarios.
- The studies have shown that simulator data are highly useful for HRA studies. Although simulator data was used as the empirical basis against HRA methods' predictions, the promising results from this study encourage and promote the use of simulator data in the future, as well as encouraging analysts to use it in different ways. 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. In summary, while there are other sources of HRA data, this study reinforced the relevance of simulator data for HRA in general. We also saw similar crew performance in the US Study as in the International Study, indicating that the results from the HAMMLAB simulator are applicable to the human performance in NPPs in other countries.
- The scenarios developed in the studies are similar to those modelled in PRA and represent difficulty levels from basic to highly complex. They can be used as standard scenarios for other HRA benchmarking studies. Complex scenarios can be used to determine whether an HRA method may lead to HEP underestimation. Basic scenarios can be used to establish a baseline performance. The difficulty levels can be used to test whether a method can produce HEPs with appropriate differentiation.

Acknowledgements

The authors gratefully acknowledge the contributions of Helena Broberg and Salvatore Massaiu, the Halden Reactor Project, Bruce Hallbert and Tommy Morgan, Idaho National Laboratory, and Amy D'Agostino, USNRC for major parts of the experimental work done in the project. The work of the nine HRA teams has of course been of invaluable importance, as was that of the additional assessment team members, Alysia Bone, USNRC, Katrina Groth, Sandia National Laboratories, and Stuart Lewis, Electric Power Research Institute. Very special thanks goes to the US nuclear power plant that supported the study with their training simulator, operating crews, instructor support in designing the scenarios, and multiple staff supporting the data collection and analysis. The plant support is obviously a major contribution to supporting the improvement of HRA and the safety of nuclear power plants.

This study is a collaborative effort of the Joint Programme of the OECD Halden Reactor Project, 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 and Idaho National Laboratory (INL) with funding from the USNRC. Sandia is a multi-program laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000. INL is a multiprogram laboratory operated by Battelle Energy Alliance LLC, for the United States Department of Energy under Contract DE-AC07-05ID14517. 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. Forester, H. Broberg, S. Massaiu, M. Hildebrandt, P.Ø. Braarud, G. Parry, J. Julius, R. Boring, I. Männistö, and A. Bye. "*International HRA Empirical Study—Phase 1 Report: Description of Overall Approach and Pilot Phase Results from Comparing HRA Methods to Simulator Data. NUREG/IA-0216, Vol. 1,*" US Nuclear Regulatory Commission, 2009, Washington, DC.

[2] A. Bye, E. Lois, V.N. Dang, G. Parry, J. Forester, S. Massaiu, M. Hildebrandt, P.Ø. Braarud, H. Broberg, J. Julius, I. Männistö, and P. Neslson. "*International HRA Empirical Study—Phase 2 Report: Results from Comparing HRA Method Predictions to Simulator Data from SGTR Scenarios. NUREG/IA-0216, Vol. 2,*" US Nuclear Regulatory Commission, 2011, Washington, DC.

[3] V.N. Dang, J. Forester, R. Boring, H. Broberg, S. Massaiu, J. Julius, I. Männistö, H. Liao, P. Neslson, E. Lois, and A. Bye. "*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,*" US Nuclear Regulatory Commission, 2011, Washington, DC.

[4] J. Forester, V.N. Dang, A. Bye, E. Lois, S. Massaiu, H. Broberg, P. Broberg, R. Boring, I. Männistö, H. Liao, J. Julius, G. Parry, and P. Neslson. "*The International HRA Empirical Study—Final Report: Lessons Learned from Comparing HRA Methods Predictions to HAMMLAB Simulator Data NUREG-2127,*" US Nuclear Regulatory Commission, 2013, Washington, DC.

[5] J. Forester, H. Liao, V.N. Dang, A. Bye, M. Presley, J. Marble, H. Broberg, M. Hildebrandt, E. Lois, B. Hallbert, and T. Morgan. "*The US HRA Empirical Study – Assessment of HRA Method Predictions against Operating Crew Performance on a US Nuclear Power Plant Simulator. NUREG-2156,*" US Nuclear Regulatory Commission, 2014, Washington, DC.

[6] J. Marble, H. Liao, J. Forester, A. Bye, V.N. Dang, M. Presley, and E. Lois. "*Results and Insights Derived from the Intra-Method Comparisons of the US Empirical HRA Study,*" Proc. PSAM11&ESREL2012, June 25-29, 2012, Helsinki, Finland.

[7] A. Bye, V.N. Dang, J. Forester, M. Hildebrandt, J. Marble, H. Liao, and E. Lois. "*Overview and First Results of the US Empirical HRA Study,*" Proceedings of the 11th International Probabilistic Safety Assessment and Management Conference, June 25-29, 2012, Helsinki, Finland.

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

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

[10] A.D. Swain. "*Accident Sequence Evaluation Program Human Reliability Analysis Procedure. NUREG/CR-4772/SAND86-1996,*" Sandia National Laboratories for the U.S. Nuclear Regulatory Commission, 1987, Washington, D.C.

[11] J. Julius, J. Grobbelaar, D. Spiegel, and F. Rahn. "*The EPRI HRA Calculator® User's Manual, Version 3.0, Product ID #1008238,*" Electric Power Research Institute, 2005, Palo Alto, CA.

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

[13] A.D. Swain and H. E. Guttman. "*Handbook of Human Reliability Analysis with Emphasis on Nuclear Power Plant Applications NUREG/CR-1278-F,*" U.S. Nuclear Regulatory Commission, 1983, Washington, D.C.