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

John A. Forester[a*], Vinh N. Dang[b], Andreas Bye[c], Ron L. Boring[d], Huafei Liao[a], Erasmia Lois[e]

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

Abstract: The International Human Reliability Analysis (HRA) Empirical Study collected data in the 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 to reference data derived from the actual performance of real crews in the scenarios. The comparisons examined both the qualitative and quantitative method predictions. The results provided both general and specific insights about HRA methods and their application. Most of the insights were derived from assessing: 1) whether the methods have the capacity to identify performance drivers and operational details of the performance of the required actions, and 2) whether they have the ability, if applied correctly, to use this information in the evaluation of the human error probabilities (HEPs) in such a way that they reflect the difficulty associated with the performance of the associated actions. This paper discusses some of the final lessons learned from the study, addressing both specific aspects of the methods involved and more general conclusions about HRA methodology. It addresses aspects of HRA methods identified as needing improvement.

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

1. INTRODUCTION

A diverse set of Human Reliability Analysis (HRA) methods are currently available to treat human failure in Probabilistic Risk Assessments (PRAs). Given the differences in the scope of the methods and their underlying models, there is substantial interest in assessing HRA methods, and ultimately in validating the approaches and models underlying them. Such a validation is warranted to assess the credibility of HRA results when decision makers have to use those results to make risk-informed decisions. Thus, the main objectives of the study were to examine how well the different HRA methods could predict or account for operating crew performance in simulated accident scenarios, investigate reasons for variability in the results within and across the different methods, and identify needed improvements in HRA methodology to facilitate making more accurate predictions of crew performance. The focus of this paper is on the conclusions about HRA methods (both specific and general) obtained from the study.

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 to reference data derived from the actual performance of real crews in the scenarios. The comparisons examined both the qualitative and quantitative method predictions. Qualitative predictions include, for instance, the aspects of the scenario or task conditions identified as the driving factors influencing operating crew performance in responding to the scenario. The quantitative comparisons take into account the estimated failure probabilities of the defined Human Failure Events (HFEs) of interest and their correspondence with the observed difficulty of the HFEs.

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

* jforester@comcast.net

This paper discusses some of the final conclusions from the study, addressing both the specific aspects of the methods involved and more general conclusions about HRA methods with common features. It addresses the strengths of HRA methods and aspects identified as needing improvement. 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 [1-4]. A brief summary is provided below. 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. The HRA methods used in the study are listed in Section 3.

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. A total of 9 HFEs were defined for the two SGTR scenarios and 4 HFEs were defined for the two LOFW scenarios.

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 where the teams reported 1) the human error probability (HEP), 2) the driving factors (PSFs), and 3) "operational expressions" or stories. The teams also provided 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 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 operations 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 and serve as the data for comparing with the HRA 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. 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 (team members collecting and analysing the crew data) and the assessment group (those comparing the HRA predictions with the crew data) participated. The difficulty rankings were determined before the comparisons between the HRA and crew data were performed and the resulting rankings remained stable throughout the study.

3. HIGHLIGHTED STRENGTHS AND WEAKNESSES OF INDIVIDUAL HRA METHODS AND APPICATIONS

Before discussing the conclusions about the individual methods, it should be noted that there was only one case in the study where the very same HRA method was applied by different teams. Thus, in some cases it was difficult to clearly separate method specific effects from those introduced by the analysts' application of a given method (i.e., distinguish method from analyst effects on the outcome). However, in spite of this limitation, it was still possible in many cases to see where analysts were varying from the method (e.g., by going beyond the method guidance or not using the method as designed) and where the method itself appeared to be contributing to strengths or shortcomings in the predictive ability of the analysis, particularly in terms of the guidance provided in the methods. It should be kept in mind however, that (with one exception) the results are from one application of each method (one team each) in one study and it is certainly possible that differences might be found in a different study.

In the following, strengths and weaknesses for each method are discussed without detailed descriptions or references to the method. In [1], short descriptions of all the applied methods are available, including all the main references to each method.

3.1 Accident Sequence Precursor (ASEP) method and ASEP/THERP (Technique for Human Error Rate Prediction)

Although two teams used ASEP to some extent in their analyses, there was a difference in the applications of the method between the two teams. Per ASEP guidance, one team mainly used THERP where appropriate to support quantification, whereas the other team only used the specific ASEP guidance for quantification. Simplicity and traceability are the strengths of ASEP. The disconnection between the two team's predictions and the crew data seemed to largely stem from the method's insufficient guidance and coverage of relevant factors. It is interesting to note that contrary to the claim that ASEP generally provides conservative HEPs, apparent optimism due to the method's weaknesses was seen in some cases.

By segmenting total time available for coping with an abnormal event into two independent parts: *allowable diagnosis time* and *allowable post-diagnosis time*, ASEP provides an option to explicitly include and quantify diagnosis or not. However, insufficient guidance is provided as to when to include or exclude diagnosis. The study results indicated that an assumption of successful diagnosis once symptom-based procedures are entered may lead analysts to fail to address operators' cognitive activities and identify some important factors influencing performance as scenarios progress. As a consequence, the final HEPs may be optimistically estimated by assuming a zero diagnosis HEP. In addition, it does not appear that the ASEP diagnosis and post-diagnosis models adequately addresses cognitive activities involved in following and responding to the steps in procedures. It should be noted that although THERP uses multipliers on identified HEPs in quantification of post-diagnosis actions, based on the study results, it does not appear to have adequate guidance to address cognitive activities involved in step-by-step or dynamic actions. In general, ASEP relies heavily on its diagnosis curve with a few PSF adjustments to address diagnosis. This approach limits the method's ability to discover cognitive mechanisms that would lead to human failures, and thus limits its ability to offer insights for error reduction. The method's weaknesses also appeared to include

insufficient guidance to help analysts evaluate PSF scaling for post-diagnosis actions and an inadequate set of factors to reliably evaluate crew behavior.

3.2 A Technique for Human Event Analysis (ATHEANA)

This particular benchmark exercise did not fully test a major feature of performing an ATHEANA analysis, which is the search for a range of Error-Forcing Contexts (EFCs) and unsafe acts (UAs) (i.e., deviation scenarios) that are consistent with the PRA definition of the HFE. It could be argued that much of the value of performing an ATHEANA analysis has not been tested by this exercise, because the EFCs and UAs were essentially predefined. However, it was still possible for the scenarios to evolve in somewhat different ways (particularly from crew actions, timing of actions, etc.), so that the ATHEANA analysis could, in principle, have identified some deviations that would have affected performance on HFEs; and in fact, some of their operational stories did reflect such variations. In addition, even within that constraint of pre-defined HFEs, the method's approach of searching for error modes or mechanisms has been shown to provide some valid predictions, particularly when the error-forcing context is strong (a strength of the method).

The ATHEANA team would also typically include operations experts from the plant being analyzed in performing an HRA, but such experts were not available to form part of the analysis team. It was noted by the ATHEANA team that the crews in the Halden study tended to move more quickly through the procedure steps and there was more variability in performance than would be expected for equivalent US crews. (However, it is impossible without formal comparison of the US and non-US crews to determine the veracity of this observation.) Additional insights on Halden crew performance were gained from the ATHEANA team's experience completing the analysis for the SGTR scenarios and subsequent comparison of their analysis to the actual crew performance. While the ATHEANA analysis did not in most cases produce a good match to the SGTR crew performance in terms of quantitative predictive power, the LOFW analysis was calibrated to Halden crew performance based on feedback in the SGTR round of the study, and the subsequent quantification for the LOFW scenario proved a very close match to the performance data. However, while the ATHEANA method often provided good qualitative analysis and operational stories, there were cases where these did not translate into appropriate HEPs. It is possible that a more structured approach for translating qualitative analysis into the quantification may further improve the ATHEANA HEP results.

3.3 Cause-Based Decision Tree (CBDT) method + THERP

Although referred to as the CBDT+THERP approach in this study [1-2], the EPRI HRA Calculator, which includes CBDT and several other quantification options, was actually used to perform the quantification of the HFEs. Although CBDT and THERP were the primary methods used, the Human Cognitive Reliability/Operator Reliability Experiments (HCR/ORE) methodology was also used in the analysis of the LOFW scenarios to quantify diagnosis. In many cases the method demonstrated the ability to identify factors that were important contributors to performance and the method obtained HEPs that showed sensitivity to the difficulty of the HFEs, even though the application did not always reflect the degree of the differences in the difficulty of some HFEs and did not always detect when error rates would be very high or very low. The derivation of the HEPs within the method and identification of which and how much the different factors contributed to the HEPs is generally traceable.

A potential limitation identified is a modelling option in the method that allows the diagnosis portion of an HFE to be ignored when it can be assumed the crew is simply following through the correct procedure. Not considering the crews' cognitive activities and related potential failure mechanisms while they are working through the procedures, led in some cases to a failure to identify some important negative drivers related to diagnosis and this led to apparent underestimations of HEPs. Another important potential limitation in the CBDT approach identified in the study is that the factors addressed or covered by the CBDT model (and more generally the HRA Calculator) may not always be adequate to identify important driving factors that influence crew performance, i.e. the model did not always guide addressing significant aspects of the scenario. Similarly, even if analysts identify operational conditions in the scenario that could be a problem, the model may not provide a direct means to incorporate this information. This was evidenced to some extent by the fact that a good operational story developed by the analysts and consistent with the data did not always translate into "appropriate" HEPs (at least as suggested by the data). It also appears that in some

cases, the approach may identify some PSFs as important contributors that inappropriately lead to higher HEPs. That is, the PSFs are judged to be at a level that should lead to increased HEPs, when in fact that they have no impact on crew performance. This effect may occur for a number of reasons but likely reflects the need for improved guidance. Taken together, these potential issues with the methodology may have contributed to the lack of differentiation that was seen between some of the HEPs where significant differences in error rates and difficulty rating were obtained in the crew data.

3.4 Commission Errors Search and Assessment – Q (CESA-Q)

The CESA-Q method was developed for errors of commission (EOCs) and was being adjusted for use in this application. Thus, the guidance had not been developed to the level it might be in the future. In addition, since the EOC focused method was intended to be applied in addition to an EOO (error of omission) approach, the method itself did not explicitly address how to treat execution issues or the execution part of HFEs. Apparently ASEP or THERP would generally be used, but were not explicitly used in this application. The method examines whether there is an error-forcing context and evaluates a number of situational factors to support the quantification process, which is eventually determined by comparing the pattern of the factors' evaluations with patterns of catalogued reference EOCs.

The method appears to provide a reasonable set of situational factors to select from to represent important factors in the scenario being analysed (at least in terms of decision-making errors), but some additional ones may be needed to be sufficient for most scenarios. The CESA-Q analysis often identified the main negative drivers reflected in the crew data. In some cases they identified PSFs as negative drivers that either did not have an impact or it could not be determined if they did, but for the most part they were fairly consistent with those identified in the crew data. While the underlying qualitative analysis performed for this study was generally good, it is not clear that without strong analysts to develop such a base (the analysis for the study was performed by the method developers, who were very experienced in PRA/HRA), that the basis for the assessment of the situational factors addressed explicitly by the method would normally be adequate. In addition, some of the results from the study (e.g., HRA team weighting of some factors identified as contributing to HFEs) indicated that additional guidance for scaling the PSFs or situational factors was needed. For example, in the SGTR scenarios the contributing factors were not always weighted negatively enough and the weighting for mild error forcing context (EFC) cases was difficult. The derivation of the HEPs within the method and what is important to performance is generally traceable, but how the various situational factors are weighted in determining the final HEP is not yet traceable.

3.5 Cognitive Reliability and Error Analysis Method (CREAM)

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

While a strength of CREAM is its ability to account for potential errors, much of this advantage is lost in implementation. The main weakness of the Extended CREAM method concerns the assignment of failure types. The assignment of generic error types is subjective (which serve as nominal HEPs), and the process of determining the dominant failure type is complicated. For the effort required to complete this part of the analysis, the result is a list of highly similar nominal HEPs that do not appear to be conservative. Once errors are identified, many are discarded in quantification. The CREAM analysts in the study chose to forego the standard way of completing quantification in CREAM by not downselecting a single, dominant failure type. Instead, they considered all failure types for each analysis. This modified process may have inflated HEP values over those typical of a CREAM analysis, but the analysts saw this as a reasonable compromise to ensure realistically conservative values in CREAM.

3.6 Decision Trees (DT) + ASEP

The DT + ASEP approach used in this study is a combination of a decision tree approach similar to the Cause-Based Decision Tree (CBDT) method and ASEP. It uses decision trees to obtain the failure probability of information processing, and uses the time reliability curve and other rules from ASEP to estimate the failure probability of diagnosis and manipulation. One strength of the approach is that judgments made by the analysts in applying the method and obtaining the HEPs are clearly traceable. With adequate documentation, the basis for judgments regarding which branches to take in the decision trees is traceable. Another strength is that once the factors included in the method are correctly evaluated, the method can provide guidance to facilitate error reduction.

The method's major limitation appears to be in its ability to address complex diagnosis situations. The guidance, influencing factors addressed, and specific questions asked during application of the method do not always seem to be adequate to identify critical issues at a more scenario specific level, particularly with respect to the cognitive aspects. As a result, in some cases the method seems to lack the power (or sensitivity) in terms of HEPs to differentiate HFEs. In several cases, a disconnection between qualitative analysis and quantitative analysis seemed to contribute to optimistic HEPs; a good qualitative analysis did not always translate into a consistent quantitative result.

3.7 Enhanced Bayesian THERP

This method is based on the use of a slightly modified version of the time-reliability curve introduced in THERP and on the adjustment of the time-dependent human error probabilities with expert judgments made about the impact of five PSFs: (1) support from procedures, (2) support from training, (3) feedback from process, (4) need for coordination and communication, and (5) mental load, decision burden (see [1] for more detail). In the analysis of the SGTR scenarios, Enhanced Bayesian THERP succeeded fairly well in the quantitative analysis. Generally the HEP values were within the empirical error limits and the analysis identified the relative difficulties of the different tasks. However, the analysis did not provide a strong differentiation between the easy and difficult HFEs. On the qualitative side there were difficulties in identifying the correct drivers for the scenarios. In this approach, it is argued that in some cases it may be sufficient to have the general effect of the PSFs to be correct, while individual PSF weights might not correspond to the empirical data. The reason for this is that the HEP is driven by the time available for the task and that each of the PSFs is treated the same on the mathematical side. So, the expert panel weighting the PSFs is not necessarily required to identify each PSF correctly to arrive at the 'correct' HEP number; rather, it may be enough that the task analysis with respect to the time available is accurate and that the combined effect of the PSFs reflects the overall difficulty of the HFE. Obviously however, this effect may not always turn out to produce an appropriate HEP or a good understanding of the relevant factors, which was seen in several cases in the study.

As with many of the PSF methods, the method has limited guidance for how to assign the different possible values to the PSFs with only about one sentence for each possible weight/PSF combination. The guidance is also generic in nature, like for example "Mental load is considerable, situation is serious, a serious decision needs to be made," without additional information. While the mathematical side of the method is easy to trace, the assessment of the PSF weights with expert judgment is clear, and the quantitative effect of the PSF weights is explicitly stated, the reasoning behind the PSF weights is dependent on the several different experts, and their reasoning might vary, so there is not necessarily consensus for the qualitative analysis of the scenarios, which limits the ability of the method to support error reduction.

3.8 Human Error Assessment and Reduction Technique (HEART)

In HEART, an HFE is quantified by matching a Generic Task Type (GTTs, of which there are 6) and adjusting the nominal HEP for this task type to account for the effect of Error Producing Conditions (EPCs). The core of the method is the list of EPCs (over 30), each with a maximum multiplier corresponding to the EPCs impact on the HEP when at its most severe. To quantify the GTT's nominal HEP is adjusted by applying a proportion of (the maximum) effect for each EPC. As applied in this study, the qualitative HEART analysis consists of identifying the EPCs relevant to the HFE and justifying each proportion of effect in terms of specific issues. One of HEART's strengths is its focus on identifying these EPCs. By

definition, an EPC is a driving factor so the analysis focuses on identifying driving factors. However, the HEART analyses do not typically explicitly discuss the interaction of the factors in an overall operational expression or failure scenario (or in multiple expressions or scenarios).

Some inherent weaknesses of the current HEART method are a) lack of guidance for the identification of the GTTs, which anchor the quantification of each HFE; b) shortcomings in the description and guidance on the EPCs; and c) lack of guidance for assessing the proportion of the maximum effect, which may lead to difficulties in the reproducibility of the method. As with many other methods, the HEART method does not explicitly include or specify a task analysis method. It can be assumed that combined with an appropriate task analysis method, the qualitative predictive power as well as the potential for obtaining insights for error reduction could be substantially improved. On the other hand, the traceability of the analysis suffers from the lack of guidance for deriving the quantification inputs from the qualitative information, which is an inherent shortcoming of the method. It should be noted that some of these weaknesses of the HEART methodology are well-known, having been identified in previous studies, and that there exist proprietary versions of HEART with additional guidance as well as an effort to develop a new version of HEART, called NARA [12]. Neither of these was available for the Empirical Study.

3.9 Korea Atomic Energy Research Institute (KAERI) K-HRA method

The K-HRA method is a thorough and sound extension of THERP and ASEP for use in the South Korean nuclear industry. It offers a clear decision tree approach that allows the ready extraction of drivers that can contribute to errors. Further, it provides a separate consideration of diagnosis and execution factors, which facilitates consideration of a wide range of error contributors.

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

3.10 MERMOS (Méthode d'Evaluation de la Réalisation des Missions Opérateur pour la Sureté)

The core of the MERMOS method is the development and quantification of one or more failure narratives or HFE failure scenarios for a given HFE. MERMOS is centered on these failure scenarios, described in terms of specific elements of the scenario, crew, and task characteristics and how these operational elements interact to result in the HFE failure. This concept underlies several strengths of the method: plant operations expertise, on how crews are trained and respond to accidents (primarily in simulators) in the plant of interest, can be incorporated directly and explicitly into the HRA; the specificity of the failure scenarios leads to insights that are directly useable for error reduction; and the use of multiple failure scenarios can represent the inherent variability of accident evolutions and the crew responses. MERMOS includes a systematic process for identifying and classifying failure scenarios; the failure narratives that result from this process were traceable and easily attributable to this process. Quantification in MERMOS is based on assessing the probability of each element of the failure scenarios on the scale {very improbable, improbable, probable, very probable} with probabilities {0.01, 0.1, 0.3, 0.9}. In addition, the scenario-based quantification ensures a strong, direct link between the qualitative analysis and the HEPs. As a result, guidance and traceability were assessed positively.

On the other hand, the dominant role of expertise in both the qualitative and quantitative analysis, i.e. to identify the failure narratives and to quantify the HEP, leads to some drawbacks. Although the plausibility and to some extent the comprehensiveness of the failure scenarios can in themselves be evaluated externally,

it is difficult to assess the HRA analysis, its results, and the HEPs without direct access to this expertise. In terms of the qualitative and quantitative predictions assessed in the study, the qualitative predictions were assessed from moderately good (for SGTR) to good (the highest rating on the scale, for LOFW). The failure scenario elements identified in the MERMOS analyses were strongly supported by the empirical evidence, especially for the HFEs where failures or near-failures were observed, i.e. the more difficult HFEs. The failure narratives (predicted interactions of the elements) were also supported by the evidence. The quantitative predictions of MERMOS were assessed as moderately good. The HEPs were mostly consistent with the empirical evidence but the results include some hints for a tendency not to differentiate the HFEs as much as might be expected, which negatively affected the difficulty ranks that were obtained (in particular for the less challenging HFEs). With the exception of the inherent issues associated with the reliance on expertise, the MERMOS method has several very positive characteristics. An issue that is outside the scope of this study is whether the reliance on expert judgment in qualitative and quantitative analysis may negatively impact repeatability (inter-analyst consistency).

3.11 PANAME (French acronym – "new action plan for the improvement of the human reliability analysis model")

In the PANAME methodology, the probability of error during the diagnosis is obtained from curves, which gives the probability that the operator will fail, depending on the time allowed. Several curves are available, and a decision tree helps the analyst to choose the most appropriate curve, depending on the context (determined by a set of PSFs). The probability of error in performing the action is a combination of three factors (a basic probability adjusted by a context factor and a probability of recovery by the team itself, depending on the time allowed to recover, and a later recovery (see [1] for more details). The quantification is easily traceable in PANAME. The performance shaping factors' effect on the quantitative HEP is explicitly stated in the documentation and the mathematics of the method is simple to understand. Depending on six PSFs, a single context factor is decided. The context factor can have a value of 1/3, 1 or 3. While this should allow the analyst to discern between easy and difficult scenarios, the resolution of the method is not that high.

While the guidance is complete and specific in terms of PSFs, it appears that the aspects being addressed may limit the analysis. The specific requirements listed in the guidance for assessing a certain value for a PSF might not be applicable in each case, The PANAME analysis did not capture the negative drivers adequately in either the SGTR or LOFW scenarios. The method itself is not geared towards producing qualitative information. The decision tree approach to analysing scenarios necessarily limits the degree to which qualitative differences of the scenarios can be assessed.

3.12 Standardized Plant Analysis Risk-Human Reliability Analysis (SPAR-H)

This study featured two teams performing separate SPAR-H analyses. The study established that it should *not* be assumed that "procedure driven actions deal with execution type of tasks only". The crews continually perform cognitive activities including situation assessment and decision making while executing procedures. In SPAR-H, one may address two task types: diagnosis and action. The guidance allows for utilizing only task type action when analyzing procedure following aspects after diagnosis of the event has occurred (i.e., after entering the correct procedure). By doing this, there is a danger of being optimistic, since the base probability for the task type action is one order of magnitude lower than for task type diagnosis. The study data suggests that the best strategy would be to always address diagnosis and execution aspects in quantifying an HFE using SPAR-H.

There is limited guidance in SPAR-H regarding the decision as to which PSFs to rate positively and which to rate negatively and this relies heavily on the analysts' judgment. In addition, without documentation on the part of the analysts using SPAR-H, it is not always obvious why the choices are made. How to decide which and how many PSFs to include as negative or positive influences and how to assign the PSF levels, can be a complicated process in SPAR-H, particularly for difficult scenarios. It is also worth noting that several of the PSFs have very high multipliers, which can lead to conservative HEPs that did not seem to align to the crew performance in some cases. Additional guidance as to how to consider the PSFs together and make scaling judgments would be very useful. Improved guidance for the overall qualitative analysis would improve the basis for these judgments. The simplicity of the base probabilities and the adjusting PSF multipliers makes it

very easy and traceable to know where the numbers come from in SPAR-H. Thus, the traceability of the quantification itself is good if good explanations for the PSF judgments are provided.

4. SUMMARY OF KEY LESSONS LEARNED ON HRA METHODS

A review of the strengths and weaknesses of the various methods described above identified a number of common features among groups of methods that are important to note.

Failure to adequately address diagnosis and related activities

Several of the methods, including SPAR-H, ASEP, and CBDT allow analysts to make a modeling decision to not explicitly address the cognitive demands associated with executing emergency procedures, such as the interpretation of cues, interpretation of procedural criteria, and monitoring of relevant plant parameters. While each of these methods includes its own approach to address and quantify the cognitive or diagnosis related aspects of a task, analysts have the option to model HFEs subsequent to the initial HFE or to the identification of the event (e.g., entering the correct procedure) 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 CBDT+THERP analysis it was decided not to use the CBDT to estimate the HEP for some HFEs, but instead included only the execution contribution using THERP. The results of the study clearly showed that failure to adequately consider the crews' cognitive activities and related potential failure mechanisms while they are working through the procedures, can in many cases lead to a failure to identify important influencing factors and result in underestimations of HEPs

Identification of failure mechanism and contextual factors

There was substantial evidence in the study that methods that focus on identifying failure mechanisms (ways the crews could fail a particular task) and the contextual factors that enable them (e.g., CBDT, ATHEANA, CESA, MERMOS), tended to produce richer content in the qualitative analysis than the PSF-focused methods (e.g., SPAR-H, ASEP, Enhanced Bayesian THERP, PANAME and similar methods such as CREAM and HEART) 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, e.g., reliable processes and associated guidance for translating the richer information into HEPs. Nevertheless, it seemed clear that across the variety of possible conditions that can occur in an accident scenario, a thorough assessment of failure mechanisms and context will be needed for reliable results (but see discussion of PSFs below).

Judging the influence of PSFs and choosing the right PSFs

Not surprisingly, in the HRA analyses using PSFs (or similar, such as the common performance conditions used in CREAM and the error producing conditions included in HEART), the evaluation of the degree of influence of the different PSFs considered by the method was an important factor. In both the LOFW and SGTR scenarios, the assessment group identified inconsistencies in the ratings of the performance shaping factors (PSFs) in those HRA methods highly based on PSFs. For example, although the present study only had one case where a single method was used by two different teams (SPAR-H), in a couple of cases the methods were similar (e.g., DT+ASEP and CBDT+THERP, along with ASEP and ASEP/THERP). Observable variations in the HEPs for the same HFEs both in the SGTR and the LOFW scenarios were seen across these methods and differences were seen in both the selection and weighting of the PSFs thought be important. Clearly, in many cases, these judgments can be difficult and the results of some methods were very sensitive to these sometimes subtle judgments. Two aspects of the analysis contributed to these inconsistencies. First, the HRA teams did not develop to the same degree a qualitative understanding of the details of the scenario. Second, there were differences in the interpretation of the scope of the PSFs and in the ratings assigned to the PSF for a given issue or performance condition. In most the HRA methods using PSFs (e.g., SPAR-H, ASEP, ASEP/THERP (THERP itself only to a limited extent), Enhanced Bayesian THERP, K-HRA and PANAME), and other methods such as HEART and CREAM that require similar types of judgments regarding the task types and performance conditions, and CBDT and CESA that require judgments about the level of various conditions, the guidance provided to support these judgments is limited. Consequently, support for consistent transformation of qualitative insights into consistent inputs for

quantification is needed. Of course, the failure cause/context based methods (e.g., ATHEANA, MERMOS) are not immune to this issue, but there is an emphasis in those methods on obtaining additional information to support the judgments. Nevertheless, the study indicated that all of the methods need improvements in the guidance related to judging which factors should be considered and how to evaluate and weight them (e.g., the level or strength of a factor or set of factors relative to an HFE).

Range of PSFs covered

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

5. MAIN CONCLUSION

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 will have an inadequate basis to address the range of conditions possible in PRA scenarios. 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.

Acknowledgements

The authors gratefully acknowledge the contributions of Helena Broberg, Salvatore Massaiu, Michael Hildebrandt, and Per Oivind Braarud, the Halden Reactor Project, Pamela Nelson, Universidad Nacional Autónoma de México, and Ilkka Männistö, VTT for major parts of the experimental work done in the project. The work of the HRA teams, 13 teams providing 14 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 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.

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, OECD 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, OECD Halden Reactor Project, Halden, Norway, (2010) and NUREG/IA-0216, VOL.2, (2011).

[3] 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).

[4] Dang, V.N, Forester J, Boring R, Broberg H, Massaiu S, Julius J, Männistö I, Nelson P, Lois E, and Bye A. *International HRA Empirical Study—Phase 3 Report: Results from Comparing HRA Method Predictions to Simulator Data on LOFW Scenarios.* HWR-951, OECD Halden Reactor Project, Halden, Norway, 2011. To be issued as NUREG/IA-0216 Vol. 3.