Results and Insights Derived from the Intra-Method Comparisons of the U.S. HRA Empirical Study[1]

Julie Marble[a*], Huafei Liao[b], Mary Presley[c], John Forester[b], Andreas Bye[d], Vinh Dang[e], Erasmia Lois[a]

[a]United States Nuclear Regulatory Commission, Washington, DC, USA
[b]Sandia National Laboratories, Albuquerque, NM, USA
[c]Electric Power Research Institute (EPRI), Palo Alto, CA, USA
[d]OECD Halden Reactor Project, Institute for Energy Technology, IFE, Halden, Norway
[e]Paul Scherrer Institute, Villigen PSI, Switzerland

Abstract: Human reliability analysis (HRA) is an important aspect of PRA which evaluates the contribution of human performance to risk. But HRA is a major contributor to variability in PRA results. Different HRA methods rely on different human performance frameworks and data, and analysts may apply the methods inconsistently. The US Nuclear Regulatory Commission (NRC) first proposed, participated in and supported the International HRA Empirical Study, where HRA predictions of different analysts and methods were compared to crew performance data at the Halden Reactor Project simulator facilities. Only one method in that study was applied by multiple teams; therefore method effects could not be separated from analyst effects. A major objective of the US Empirical Study was to test the consistency of HRA predictions among different analyst teams using the same methods. In this study, at least two different analyst teams applied each method to predict the outcome of the scenarios. We examined the qualitative analyses to identify differences and the extent to which the differences in results were due to analysts versus the methods. This paper discusses the insights for method guidance and the intra-method comparisons. A companion paper discusses the empirical data and overall results [1].

Keywords: HRA, PRA, Benchmarking, Crew Performance.

1. INTRODUCTION

The US Nuclear Regulatory Commission (NRC) implements a risk-informed regulatory framework; risk information (such as that derived from Probabilistic Risk Analysis (PRA)) is part of its decision-making process. HRA is an important aspect of PRA since it addresses the human performance contribution to risk, frequently identified as significant. However, HRA is a major contributor to the variability of PRA results. The use of different HRA methods that rely on different human performance frameworks and data, as well as inconsistent implementation from analysts, is the most common sources of variability.

To address these variability issues, the US NRC participated and supported the International HRA Empirical Study [2, 3, 4], in which HRA predictions of different analysts and methods were compared to observed crew performance data at the Halden Reactor Project HAMMLAB simulator facilities. The International HRA Empirical Study identified important strengths and weaknesses of the various HRA methods used in the study, and an important conclusion was that improving the qualitative analysis aspects of HRA methods could increase their robustness and reduce some of the sources of the variability in results that are seen in applications of different methods. However, since only one of the methods examined in the International Study was applied by multiple teams, it was difficult to clearly separate method-specific effects from variability created by the analysts' application of a given method. Thus, in addition to examining differences across methods, a major objective of the present study was to test the consistency and accuracy of HRA predictions among different analyst teams using the same methods. The study was performed on a US nuclear power plant (NPP) simulator and is thus referred to as the US HRA Empirical Study.

Nine teams participated in the study and each team applied an HRA method to predict the Human Failure Events (HFEs) in the simulator scenarios. Two teams used ATHEANA, two teams used SPAR-H, two teams used ASEP, two teams used the HRA Calculator (with CBDT, HCR/ORE and THERP), and one team used methods in the HRA Calculator without using the actual software. The experimental design, criteria used

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

* Julie.Marble@nrc.gov

and descriptions of the scenarios and HFEs are described in [1]. The HRA teams' predictions are compared to the empirical crew data from the simulator. The aggregated performance of the crews' HFE related actions is described in the following three ways, which correspond to those in which the HRA teams were asked to report their predictions and serve as the data for comparing with the HRA predictions.

- 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.

2. OVERALL EMPIRICAL CREW RESULTS

The data were collected on a US plant full-scope training simulator with a conventional control room. Four crews of five licensed crew members performed three scenarios; however, one crew was unable to complete Scenario 3 (SGTR) due to a simulator problem. Five HFEs were defined in the three scenarios, 1A, 1B, 1C, 2A and 3A, see [1] for a description. The HFEs were ranked in terms of problems experienced by the crews in diagnosing and completing the action based on the empirical data, as well as by three of the four Unit Supervisors who participated in the study, and a consistent ranking was obtained. The crew failure rates and difficulty ranking are listed in Table 1 (see [1] for more information). It should be noted that HFE 1B was not ranked since there was no empirical data for this HFE.

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)

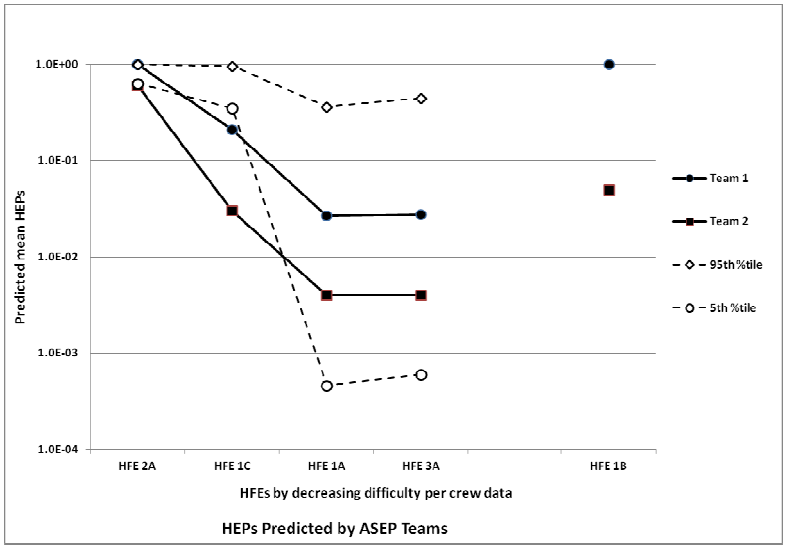
3. INSIGHTS ABOUT METHODS FROM THE INTRA-METHOD COMPARISONS

The HEPs predicted by each method are plotted in Figure 1 alongside the Bayesian uncertainty bounds derived from the crew data. On the horizontal-axes, the HFEs are ordered by their difficulty ranking. In the following sub-sections, the intra-method comparison results are summarized for each method.

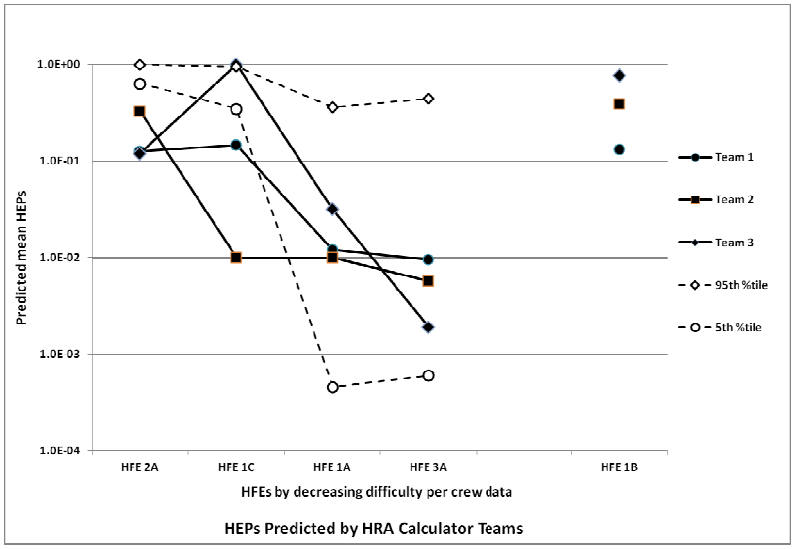
3.1 ASEP (Accident Sequence Evaluation Program Human Reliability Analysis Procedure)

Two teams (ASEP Team 1 and Team 2) performed analyses with ASEP. In most cases where a discrepancy occurred, the qualitative analysis of Team 1 tended to be more consistent with the empirical crew data in terms of performance drivers and operational stories. Overall this appears to be due to a more detailed qualitative analysis by Team 1 that seems to go beyond the ASEP methodology. In comparison, Team 2 tended to stick relatively close to the specific guidance in ASEP. The payoff of Team 1's detailed qualitative analysis was particularly apparent as the team was able to identify how complications in the scenarios could influence operators' diagnosis. Their detailed analysis can also be illustrated in their consideration of the role of operating procedures in operators' diagnosis. Such consideration is not explicitly addressed in ASEP, but it enabled Team 1 to identify the difficulties in complicated scenarios that would prevent operators from making a timely diagnosis or delay them in reaching relevant procedure steps. In contrast, Team 2 only considered whether post-diagnosis actions were covered in procedures, per the ASEP guidance.

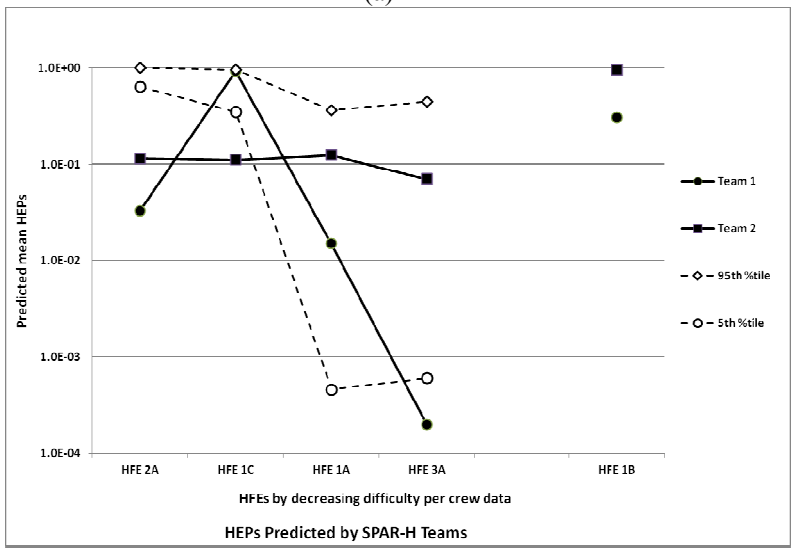
The findings discussed above, to some degree, can be related to one of the methodological features of ASEP. For estimation of diagnosis HEP, ASEP relies on its Nominal Diagnosis Model (i.e., time reliability curve (TRC); ASEP Figure 8-1) with a few PSF adjustments. The benefit of the model is its simplicity; however, by ignoring the details of how PSFs interact with operators' cognitive behavior (as with many other first generation HRA methods), such an approach is likely to limit the method's ability to identify important plant conditions or operational situations that would negatively impact operators' behavior, and thus produce few insights for error reduction.



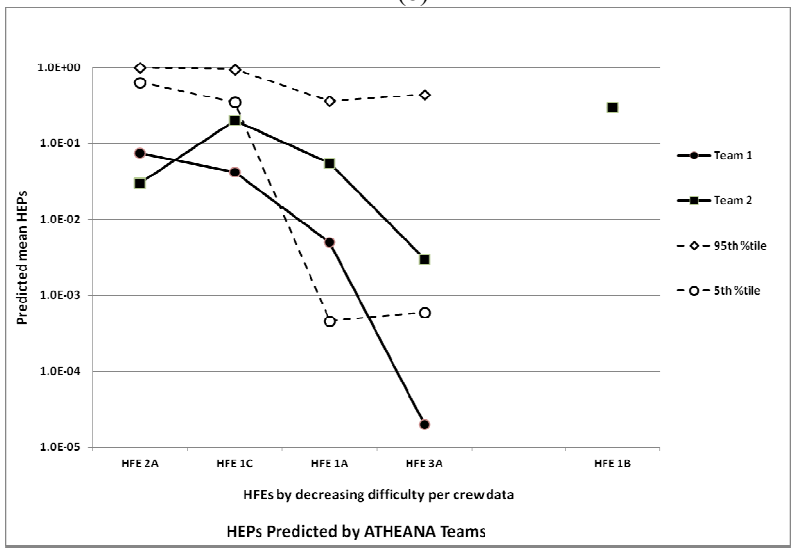
(a)



(b)



(c)



(d)

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

As shown in Figure 1.a, the HEP curves derived by the two teams are close to parallel and, on a general basis, consistent with the HFE difficulty ranking. Unlike the qualitative analysis, although the HEPs estimated by Team 1 are one order of magnitude larger than those by Team 2 for most HFEs, it is difficult to tell which team's HEP predictions tend to be more consistent with the crew data because it is difficult to determine the true bounds due to the small sample of crew performance data.

For most cases, the diagnosis HEPs by Team 1 were larger than those by Team 2. The differences seem to be largely caused by the differences between the two teams in their execution time estimation, which affected the estimates of available time for diagnosis. As discussed above, the differences can be partly attributable to the detailed analysis of Team 1, which enabled them to identify plant conditions or operational situations that would delay operators in reaching relevant procedural steps. As another example to illustrate the detailed analysis by Team 1, Team 1 considered the time needed to finish some procedural steps that did not seem to be explicitly considered in the analysis of Team 2. Moreover, Team 1, in a couple of cases, showed more conservatism in estimating the time for some actions than did Team 2. The discrepancies seem to suggest that more specific guidance for execution time estimation is necessary to reduce variability in HRA teams' analyses. In addition, Team 2's analysis seems to indicate that there is inadequate guidance to address how operators' training and experience may impact their selection of procedural paths and thus affect execution time.

It should be noted that the two teams seemed to have obtained different information/impression from their interviews with operators for the most difficult HFE (HFE 2A) with respect to crews' knowledge and experience. The information obtained by Team 2 led them to believe that crews' experience would help their diagnosis and thus the team selected the nominal diagnosis HEP, whereas the information obtained by Team 1 led them to believe that crews' training and experience was not adequate and thus they selected the upper bound diagnosis HEP. Although the two teams' HEPs are not substantially different, this, to some extent, can illustrate how the information obtained from the interviews can impact HRA results.

The HFE 1B analyses showed how the HRA analysts' judgment about the information from interviews led to variability in their analyses. The two teams obtained the same information from interviews about crew training and experience but their analyses and predictions differed. Team 1 decided to discount training and experience due to the high difficulty of the scenario and obtained a high diagnosis HEP, while Team 2 decided to credit crews' training and experience and obtained a small diagnosis HEP. The HEP difference (i.e., very large vs. relatively low) is large enough to lead to different conclusions. Evaluating which prediction is more accurate was not possible due to the lack of empirical data for this HFE.

The execution HEPs of Team 1 are consistently about one order of magnitude larger than those of Team 2. One contributing factor to the numerical difference seems to be the different quantification approaches used by the two teams. Team 2 used ASEP rules, whereas the Team 1 followed the guidance in Item 2 in ASEP Table 8.5 to use THERP as they decided that there was enough information to perform task analysis. This finding may sound contradictory to the claim that ASEP generally provides conservative HEPs, and can be explained by the fact that Team 1 quantified more steps.

The impact of experience and training is, to some degree, addressed by PSF adjustment; however, there does not seem to be explicit and/or adequate guidance to help analysts address knowledge-based behaviour in identifying critical actions and procedural paths. In one case, Team 1 considered procedures that the crews did not actually enter; as a result, they conservatively estimated the time required to finish post-diagnosis actions, and obtained a larger execution HEP compared to Team 2.

3.2. HRA Calculator

Two teams (HRA Calculator Team 1 and Team 2) performed an analysis of the five HFEs using the EPRI HRA Calculator. A third team (HRA Calculator Team 3) performed the analysis using a hybrid CBDT+ASEP method without using the actual software package (CBDT and ASEP are a key aspect of the HRA Calculator). The EPRI HRA Calculator quantifies the cognitive contribution to the final HEP with CBDT (if time was not critical) or the maximum value of CBDT and HCR/ORE methods (if time appeared to be critical), and the execution contribution with THERP. In contrast, the analysts using the CBDT +

ASEP approach quantified the cognitive contribution with the sum of the values from CBDT and the TRC from ASEP, and quantified the execution contribution with ASEP.

Overall the three teams performed well against the empirical data. While at a high level the operational stories were generally consistent between teams, there are many instances of inconsistencies in the details of the operational stories and performance drivers; however due to the significant variability in the quality of the documentation available for the qualitative analysis, it is difficult to pinpoint the root cause of inconsistencies between teams. For instance, the operational stories were not always detailed enough to understand the expected progression of the scenarios (e.g., the expected procedural path, critical decision point, cues and other activity). Some teams provided a full description of the scenarios, while others provided only the decision tree paths, with minimal qualitative assessments. The greatest variation in the analyses was seen in cases where multiple procedural or knowledge-based success paths existed; the empirical data shows much crew variability in these cases as well. This may be explained by one of the inherent weaknesses of the methods. The HRA Calculator is comprised of three parts: CBDT branch points, HCR/ORE formula and THERP tables; dependency and recovery factors are then applied based on pre-defined rules. This structured approach provides a clear way to trace between the input values and the final HEP; however it does not provide a specific format for the qualitative analysis, so the area of least traceability is the operations story, cues and timing analysis. Consequently, *why* each input value (e.g., timing, branch points, and recovery designations) was chosen is not always discernable – it depends on the quality of the analysts' documentation.

With respect to prediction of the performance drivers of all HFEs, drivers often matched at high level, but the reasoning for selection of the driver did not consistently match the empirical data. For example, procedural guidance came up several times as a negative driver in the empirical data. In these cases, some of the analyses would also find procedural guidance as a negative driver. However, the empirical data refers to difficulty with the content and progression of the procedures, while the analyses found procedural guidance to be a negative factor based on the format of the procedure (e.g., existence of NOT or AND/OR statements), which was not supported by the data. Also, the magnitude of the effect of drivers was not always consistent between what was discussed in the qualitative analysis and what was manifest in the quantitative analysis.

The basic assumption behind this method is that operators will follow procedures and generally trust their cues. That means that little to no credit is given for actions involving knowledge-based actions, so significant analyst judgment had to be applied to quantify these cases and make the trees "fit" their understanding of the operational story. Furthermore, in CBDT, procedural guidance only covers procedure format and whether or not it matches with the cues; the clarity of the workflow of the procedures are not addressed. Similarly, substantial complexity and/or teamwork is not dealt crisply within CBDT and mismatch between training and scenario is not evaluated except in very specific circumstances. This aspect of the method contributed to the variability across analysts.

The HEP curves derived by the three teams are shown in Figure 1.b. Overall, total HEPs match scenarios moderately well in rank and level of differentiation; however, the cognitive vs. execution contribution to the HEPs were not always consistent with what would be expected from the empirical data. For more difficult HFEs, HEPs were generally underestimated and there was more variability in the judgment of contributions to the final HEPs (although there is not significant variability in the final HEPs). The results on HFE 1C were least consistent, which seemed to be congruent with the observation that the HFE had the least defined procedural path and the most crew variability in the empirical data.

Inconsistencies across these teams' analyses stem mainly from the following three major sources:

- Timing estimates (for HCR/ORE) for diagnosis and execution.

Timing was consistently a discrepancy across scenarios. There were both large variations from analyst to analyst and between the analyses and the empirical data, with the assessed timing fairly consistently optimistic compared to the data. Team 1's documentation did not include timelines for each HFE. Although clear guidance is provided on definition of timing points, no specific guidance exists on how to assess those values; therefore the timing estimate is the most subjective portion of the HRA Calculator. Differences in timing estimates came from differences in the analysts' assumptions and what was included in the estimate

(e.g., steps, cues, factors that might add delays). For many teams, the timing seemed to be based strictly on the time it would take to get cues and get through the steps without explicitly accounting for elongation of the time frame due to distractions, parallel actions, etc.

Moreover, cases with high complexity and limited time do not seem to be well accounted by the HRA Calculator. The CBDT method only accounts for complexity since a prerequisite for using it is that time is sufficient. The HCR/ORE method accounts for time explicitly, and other PSFs (e.g., complexity) are supposed to be reflected in the time estimates. However, in this application, time estimates were provided as point estimates derived from analyst or operator opinion. Furthermore, the assumptions underlying the timeline are not clearly defined (e.g., did the obtained timeline for diagnosis consider distractions that the operators may be facing?). For HCR/ORE, lack of traceability in developing the timeline translates directly into a lack of traceability in the final HEP.

Additionally, the analysts obtained their timing from different sources: one team relied primarily on analyst judgment while the other team used input directly from the interviewed trainers/operators. The variability seems to suggest that additional guidance for timing estimation is necessary to reduce subjectivity.

- Philosophy of method application/analyst judgment

In method application there were two areas where differences arise: 1) conservatism vs. best estimate and 2) verbatim application of the method (this was what was asked of the analysts) vs. analyst judgment. For example, one team used the Calculator and applied recovery to the cognitive and execution portions as per the calculator, while the two other teams "conservatively" did not credit recovery. Moreover, the differences in conservatism led the teams to select different decision tree branches based on the same qualitative analysis input. Other less quantitatively significant versions of differences in analyst subjective judgment were peppered through variability in PSF scaling and procedure interpretation. Such findings seem to indicate inadequate method guidance for analysts to perform quantitative analysis and how to drive quantitative analysis with qualitative analysis.

- Decomposition of the HFE

There is little guidance on how or when to break up HFE into subcomponents (e.g., different cognitive failure mechanisms); as a result nearly all of the HFEs were decomposed differently between teams. For some HFEs, one team defined multiple cognitive failure mechanisms – each quantified separately, while the other two teams quantified these as one failure. Similarly, different teams made different decisions on which execution steps to be quantified.

Although two teams used THERP in the quantification of the execution steps, inadequate guidance led to a difference in their approaches: for each execution step quantified, one team applied *either* the Error of Commission (EOC) *or* the Error of Omission (EOM), while the other team applied both.

3.3. SPAR-H (Standardized Plant Analysis Risk HRA)

Two teams (SPAR-H Team 1 and Team 2) employed different approaches to the qualitative analysis. The analysis of Team 1 evaluated the events at the level of the HFEs with a focus on PSFs, while the analysis of Team 2 used CRTs (Crew Response Trees) to determine the decision points and basic events for the analysis based on the break points in the procedures. The CRTs are a qualitative analysis tool being developed as part of an ongoing project (see [5]), which results in a detailed decomposition of actions and events within one HFE. The freedom in the choice of qualitative approaches stems from the fact that SPAR-H, as explicitly labeled in its documentation, is a quantitative method. It does not provide insight or detail on 'conventional' modeling considerations, such as the level of decomposition of a scenario or how sub-tasks should be combined for analysis, as the method assumes that these are situation and analyst specific. The method states that the analyst may break the scenario into subtasks for analysis, and suggests that the ATHEANA decomposition process be used, but other event trees may be created. Therefore, the choices of both teams on qualitative analysis approaches are in keeping with the SPAR-H method.

SPAR-H basically assumes that differences in task decomposition will not make large effects on the estimated HEPs. The study suggested that it may have an impact on the quantitative analysis to break down

an HFE at a minute level as was in the case of Team 2. SPAR-H requires a scaling factor to be used if more than three PSFs are used at the HFE level; however, there is no clear instruction on how to apply or interpret this rule when two or three PSFs are applied to the sub-events of an HFE and then summed up to the HFE level. The method guidance should be updated by taking such cases into account. Whether this was the reason for e.g., the over-estimation of HFE 3A or whether Team 2 over-estimated the complexity of this HFE needs to be more closely investigated.

Compared to Team 1, Team 2's use of the CRTs allowed them to identify detailed failure and success paths and provide a greater qualitative insight in the scenarios, and potentially greater qualitative insight into possible operational stories and performance drivers. This would be expected to yield a better basis for error reduction with Team 2's analysis. Note that Team 2's analysis benefited from the experience of the team with nuclear operations; however, it may also have had less experience in developing HEPs than Team 1.

The differences in the analysis of the procedures between the two teams can be illustrated in the following. Team 1 did not discuss transitions in procedures and where exactly in the procedures the complications in the scenarios would have an impact, although they discussed some details of the developing scenarios, and this discussion seemed to be built on good scenario insight. In contrast, transitions in the procedures were a significant point in Team 2's analysis and served to distinguish basic events as subdivided in the CRTs.

Team 1 was optimistic on the most difficult HFE (HFE 2A), mainly based on the information from interviews with operators. It appears that Team 1 could have improved their qualitative analysis with a more detailed analysis including procedure branching and timing, and/or with a more detailed interview. It seems that an HRA method should include an interview process that guides analysts into the details of scenario progression, which in this case may have revealed the difficulties on the complexity of the situation combined with the timing issues of the procedure following. Moreover, a detailed scenario understanding seemed to contribute to Team 1's good predictions of other difficult HFEs. For easy ("vanilla") scenarios, Team 1's overall analysis seemed to be sufficient to make a fairly good assessment of the qualitative drivers. This finding seems to suggest that a detailed qualitative analysis, in order to capture the nature of the scenarios, is more essential to the analysis of more difficult scenarios. However, it may be difficult to know which scenario requires such a detailed analysis without performing it.

SPAR-H has a set of pre-defined PSFs with associated levels of effect. As with other methods, there is overlap between the definitions of the PSFs. The overlap in dimensions allows analysts to select how to account for a factor in different ways. In the current study, Team 1 accounted for the lack of plant indications (cues) under complexity, while Team 2 accounted for it under HMI: Missing Indicator, obtaining a much greater multiplier for the same factor. The mapping of qualitative analysis results to PSFs and choosing between PSF levels may be a difficult process in all PSF based methods. The guidance for how to do this in SPAR-H could be improved, e.g., with particular examples or reference cases.

The quantification process of SPAR-H is very transparent and is highly traceable, given the PSF multipliers. However the traceability of the choice of a multiplier for a particular PSF will largely depend on the documentation of the complications of scenarios in operational terms (especially for complexity issues). The HEP estimates by the two SPAR-H analyses are presented in Figure 1.c. In general, the HEP estimates for the scenarios predicted by Team 1 follow the difficulty ranking by the crews, except the most difficult HFE as mentioned above. It is interesting to note that despite a detailed qualitative analysis, the HEP estimates predicted by Team 2 hardly distinguish among the HFEs, nor do they follow the ranking based on the empirical data. At this point, the reasons for this outcome are not clear and are being investigated.

3.4. ATHEANA (A Technique for Human Event Analysis)

Two teams utilized ATHEANA to predict the results in this empirical study shown in Figure 1.d (ATHEANA Team 1 and Team 2). For both teams, the HEPs of the most difficult HFEs (HFEs 2A and 1C) appear to be under-predicted against the empirical data. With the exception of Team 2's evaluation of HFE 2A, the HEPs of both teams are consistent with the difficulty ranking, with a significant differentiation among the HFEs. However, Team 1's estimates are consistently nearly an order of magnitude less than Team 2's estimates for those HFEs.

Many discrepancies between the two teams seem to arise from the differences in application of the method. It appears that although much of the guidance was effectively used by Team 2, they omitted or performed some elements in less detail than the method guidance would suggest. This, to some extent, may be related to one weakness of the method. ATHEANA has a heavy emphasis on a good qualitative analysis and the full implementation of the method can be resource intensive. Compared to Team 2, Team 1 took a more detailed approach at the analysis. However, it took Team 1 250 planned man-hours to perform the analysis versus 90 planned man-hours for Team 2.

For each scenario, Team 1 developed an expected response, including alternative procedural paths and sequences of actions in the performance of the HFE task. These alternatives are based on variations in the duration of operator subtasks or in the sequence of these subtasks, which in turn affect the plant response and the timing of cues. During the expert elicitation, these scenario maps were refined and for every branch point and timing estimate a distribution was estimated. The distributions include potential effects of various PSFs or other sources of uncertainty. For each branch, Monte Carlo trials were run to come up with a probability of success for that branch (where "success" is when the time required is less than time available). Note that this Monte Carlo treatment of the paths and their duration is to some degree an innovation relative to previous ATHEANA applications from the literature. In two instances, where the distribution for time required was always less than the time available, the analysts developed a separate distribution to account for failures due to "unexplained reasons". After all the HFEs were quantified, the analysts performed a sanity check by ensuring the numerical ranking of the HFEs were consistent with the analysts' understanding of relative degree of difficulty.

Team 2's analysis was more streamlined. They made the following simplifications, which led to the differences in the analyses of the two teams.

- The ATHEANA method requires analysts to define the expected scenario progression and then search for deviation scenarios and factors which could produce an Error Forcing Context (EFC) leading to Unsafe Actions (UAs) that can ultimately lead to failure in that scenario. Team 2 did not explore alternative paths, which seemed to partly contribute to the team's failure to capture some of the main negative drivers for the most difficult HFE (HFE 2A). For this study, the EFCs were essentially pre-defined by the scenarios that were run.
- Although Team 2 broke some of the HFEs down into UAs, they did not attempt to quantify the UAs separately. Instead, they estimated probabilities for overall UAs corresponding to the HFEs themselves. As a result, it is more difficult to see where analysts' judgment came in and what was behind that judgement. However, it is not clear from the available information whether or to what extent the qualitative analysis and quantitative results might have been different had it been possible to follow through more fully with the ATHEANA process.
- The quantification process of Team 2 appears to have been somewhat less formal, with each of three experts providing his or her inputs in a more holistic framework for the HFE. Moreover, the analysis stopped with the definition of the initial three points of a distribution (1%-tile, 99%-tile and most likely value), without following through further developing the distribution. In addition, because of time constraints, there was no check performed after all the HFEs were quantified to ensure that the ranking of the HFEs was consistent with the analysts' expectations; this "sanity check" is a normal part of the process. It is not possible to assess the impacts that would have resulted from following the available guidance more explicitly.
- Team 2 only produced point-estimate in timing estimation (vs. Team 1's distributions for everything). Team 1's analysis seemed much more focused on whether sufficient time was available given different conditions, and nearly all the dominant failures stemmed from finding a context where there was not enough time. Team 2's approach did not exhibit the same rigor in accounting for delays in the timeline.

It should be noted that despite the simplifications mentioned above, Team 2 was generally quite effective in identifying potential causes for the HFEs in terms of operational expressions. This was the case even for the difficult HFEs where the team was relatively less successful in identifying important drivers.

Team 1's approach (using detailed scenario maps) and documentation provided a very clear and traceable link between the qualitative analysis and the resultant HEPs. In contrast, although it was relatively easy to understand Team 2's thought processes behind their qualitative analysis, the translation of the qualitative

information into quantitative estimates was somewhat less transparent. ATHEANA, to a large extent, relies on expert judgment. It is challenging to provide direct correlations of qualitative inputs to expert judgment estimates, as expert judgment depends in great part on individual's background and experience. This challenge generally makes reproducibility of quantitatively results inherently more difficult, and may have been magnified somewhat in Team 2's analysis because they were not able to follow the process as fully as the ATHEANA guidance might have called for. It should be noted that for the most difficult HFE (HFE 2A), Team 2 seemed to be biased by their experience to substantially underestimated the impact of the negative drivers. As discussed above, the bias may have been tempered with consideration of possible deviation scenarios.

Another weakness of ATHEANA illustrated in the study seems to be the inadequate guidance to drive quantitative analysis with qualitative analysis. For Team 2, PSF rankings were used to guide their judgments, but they were not explicitly factored into quantitative results. For the most difficult HFE (HFE 2A), insufficient emphasis was placed in the quantitative analysis on the qualitative factors identified as potential drivers.

4. CONCLUSIONS

The International HRA Empirical Study recognized that significant variability can occur in the results of different HRA methods for the same HFE due to the limitations in the methods' technical and methodological bases [2, 3, 4]. The differences between the analyst teams applying the same method observed in this study underscore the need to enhance the guidance for the application of the methods. Furthermore, it suggests that piloting of the methods (and of this guidance) in view of analyst-to-analyst reproducibility would be warranted. The implication of the study is that in addition to inherent method-driven factors, analyst-driven factors and the interaction between the analyst-driven factors and method-driven factors can also cause significant variability in the HRA results.

The variability across analysts using the same method seems to stem largely from analysts' decisions about how to apply various aspects of the method. As seen in the study, 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 have to, more or less, rely on their own, subjective interpretation of the guidance. Sometimes the methods allow analysts to deviate from method guidance. The sources for analysts' subjectivity can be illustrated in the following aspects.

- Choices and uses of qualitative analysis approaches. Some methods (e.g., SPAR-H) do not have specific guidance for qualitative analysis, and thus the decisions of qualitative approaches are completely left to analysts. Some other methods (e.g., ATHEANA) do have guidance for qualitative analysis, but the guidance is somewhat open-ended and not well-structured for translating the information into HEPs in a consistent manner (also see below).
- Decisions on task decomposition. Most HRA methods do not have a consistent approach for task decomposition. Insufficient task decomposition may cause analysts to fail to understand the difficulty in a scenario, especially for complicated scenarios. It can also cause analysts to ignore cognitive activities involved in step-by-step actions (e.g., ASEP). In addition, the level of task decomposition may result in different groupings of tasks and thus affect the dependency between tasks, which may have a further effect on HEP estimation.
- Decisions on performance shaping factors (PSFs) and associated scaling levels. 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. Overlap in the definitions of the pre-defined PSFs and inadequate guidance on determining the level or strength of a PSF can cause observable variations in analysts' judgment.
- Translation of qualitative analysis to HEPs. A broad qualitative analysis in evaluating likely crew performance does not necessarily lead to HEPs that are consistent with crew failure rates. 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. 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.

HFE 3A was a standard SGTR scenario. Based on the difficulty ranking in Table 1, it was the easiest and probably far easier than the other HFEs. All of the HRA teams also agreed that it was the easiest HFE; however, it is interesting to note that there is a large amount of variability in the HEPs. A closer investigation is needed to shed some light on the source of the variability.

For the most difficult HFE (HFE 2A), the HEPs are comparable to each other within the same method, but all the methods seemed to underestimate the HEPs except ASEP given the 100% failure rate. Since the HEPs of difficult HFEs are normally driven by challenges in diagnosis and/or insufficient time, the optimistic HEPs may suggest the inefficacy of these methods to assist analysts to fully understand those challenges and then make appropriate judgment to address them.

In summary, the study revealed that while a good qualitative analysis is a relative strength of some methods, qualitative analysis is a shared weakness across all methods (i.e., they all can be improved). Detailed qualitative analyses can lead to good HRA predictions in terms of performance drivers and operational stories both across and within methods. However, good qualitative analyses do not necessarily lead to good quantitative analysis, so additional guidance on translating the information into HEPs is also needed.

The intra-method comparison performed in this study has examined the qualitative and quantitative comparisons in order to identify method strengths and weaknesses independent of analyst specific effects and useful results have been obtained. However, a caution should be raised about the relatively few data points regarding number of HRA teams per HRA method in this study and the limited number of HFEs and scenarios. With two or three teams per method we obviously cannot conclude that any one method is better than another, even though Figure 1 suggests that some methods provided somewhat more consistent results than others. However, we hope we have identified useful information about the methods that should be worth being aware of when using them.

Acknowledgements

The authors gratefully acknowledge the contributions of Helena Broberg, Salvatore Massaiu, and Michael Hildebrandt, the Halden Reactor Project, Bruce Hallbert and Tommy Morgan, Idaho National Laboratory (INL), 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 the additional assessment team members, Alysia Bone, USNRC, Stuart Lewis, Electric Power Research Institute (EPRI) and Katrina Groth, Sandia National Laboratories (SNL). Very special thanks goes to the US NPP 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 NPPs.

This study is a collaborative effort of the Joint Programme of the OECD Halden Reactor Project, the USNRC, the Swiss Federal Nuclear Inspectorate (DIS-Vertrag Nr. 82610) and the US EPRI. In addition, parts of this work were performed at SNL and INL with funding from the USNRC. SNL is a multi-program laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the US Department of Energy (DOE) under Contract DE-AC04-94AL85000. INL is a multiprogram laboratory operated by Battelle Energy Alliance LLC, for the US DOE under Contract DE-AC07-05ID14517.

References

[1] Bye A, Dang V N, Forester J, Hildebrandt M, Marble J, Liao H, Lois E. 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.

[2] Lois E, Dang V N, Forester J, Broberg H, Massaiu S, Hildebrandt M, Braarud P Ø, Parry G, Julius J, Boring R, Männistö I, and Bye A. 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, Washington, DC, 2009.

[3] Bye A, Lois E, Dang V N, Parry G, Forester J, Massaiu S, Boring R, Braarud P Ø, Broberg H, Julius J, Männistö I, and Nelson P. 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, Washington, DC, 2011.

[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.

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