SAND2012-1680C

CHAPTER XX

Challenges in Human Reliability Analysis (HRA): A Reflection on the Accident Sequence Evaluation Program (ASEP) HRA Procedure[1]

Huafei Liao[1], Alysia Bone[2], Kevin Coyne[2], John Forester[1]

[1] Sandia National Laboratories, USA
[2] U.S. Nuclear Regulatory Commission, USA

ABSTRACT

Human reliability analysis (HRA) is used in the context of probabilistic risk assessment (PRA) to provide risk information regarding human performance to support risk-informed decision-making with respect to high-reliability industries. In the current state of the art of HRA, variability in HRA results is still a significant issue, which in turn contributes to uncertainty in PRA results. The existence and use of different HRA methods that rely on different assumptions, human performance frameworks, quantification algorithms, and data, as well as inconsistent implementation from analysts, appear to be the most common sources for the issue, and such issue has raised concerns over the robustness of HRA methods. In two large scale empirical studies (Bye et al., 2012; Forester et al., 2012), the Accident Sequence Evaluation Program (ASEP) HRA Procedure, along with other HRA methods, was used to obtain HRA predictions for the human failure

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

events (HFEs) in accident scenarios. The predictions were then compared with empirical crew performance data from nuclear power plant (NPP) simulators by independent assessors to examine the accuracy of the predictions. This paper first provides a brief overview of the study methodology and results, and then discusses the study findings with respect to ASEP and their implications in the context of challenges to HRA in general.

Keywords: human reliability analysis (HRA), ASEP, THERP, probabilistic risk assessment (PRA), crew performance

1 INTRODUCTION

Human reliability analysis (HRA) can be defined as the use of systems engineering and behavioral science methods in order to render a complete description of the human contribution to risk and to identify ways to reduce that risk. It is used in the context of probabilistic risk assessment (PRA) to provide risk information regarding human performance to support risk-informed decision-making with respect to high-reliability industries. For example, risk information from HRAs is an important input to the U.S. Nuclear Regulatory Commission (NRC) for their licensing and regulatory decisions.

In the current state of the art of HRA, variability in HRA results is still a significant issue, which in turn contributes to uncertainty in PRA results. The existence and use of different HRA methods that rely on different assumptions, human performance frameworks, quantification algorithms, and data, as well as inconsistent implementation from analysts, appear to be the most common sources for the issue, and such issue has raised concerns over the robustness of HRA methods.

The Accident Sequence Evaluation Program (ASEP) HRA Procedure (referred to as "ASEP" in this article) (Swain, 1987) is based heavily on the Technique for Human Error Rate Prediction (THERP) (Swain and Guttman, 1983) method with simplified human performance models and guidance, and it was developed to enable systems analysts at reasonable cost, with minimum support and guidance from experts in HRA, to make estimates of human error probabilities (HEPs) and other human performance characteristics that are sufficiently accurate for many PRAs. The analysts essentially quantify HEPs by first estimating the response times required to perform some critical actions and evaluating factors prescribed by the ASEP guidance and relevant to the human failure events (HFEs) being addressed. Then, they select appropriate HEPs (with uncertainty bounds) from the tables and curves provided in ASEP based on the assessment of the factors. Like THERP, ASEP relies on a time reliability curve (TRC) for quantifying the probability of failure in the diagnosis portion of human actions. Although ASEP allows use of THERP to support quantification of post-diagnosis actions, ASEP is almost entirely self-contained as an HRA method; the analysts do not need to be familiar with THERP and are required to use any of the THERP models or data.

2 OVERVIEW OF THE STUDY METHODOLOGY AND RESULTS

Two large scale empirical studies were conducted to develop an empirically based understanding of the performance, strengths, and weaknesses of different HRA methods used to model human response to accident sequences in PRAs. The empirical basis was developed through simulator runs with real crews responding to accident situations similar to those modeled in PRAs. The two studies are referred to as the International HRA Empirical Study (Lois et al., 2009; Bye et al.,2011; Dang et al, 2011; Forester et al., 2012) and the U.S. HRA Empirical Study (Bye et al., 2012; Marble et al., 2012), and were a multinational collaborative effort involving a number of organizations from 10 countries. The U.S. NRC, in particular, played a major role in supporting the preparation and execution of the studies.

In either study, two teams employed ASEP to obtain HRA predictions for the HFEs in the accident scenarios. The predictions were then compared with the crew data by independent assessors to examine the accuracy of the predictions. 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.

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 HFEs of interest and their correspondence with the observed difficulty of the HFEs (Bye et al., 2012; Forester et al., 2012).

In the International Study, two categories of scenarios were performed: steam generator tube rupture (SGTR) and loss of feedwater (LOFW) scenarios. Nine HFEs were defined in the SGTR scenarios and four HFEs in the LOFW scenarios. Fourteen European nuclear power plant (NPP) operator crews participated in study, and their performance data were collected in the Halden Reactor Project's HAMMLAB (HAlden huMan-Machine LABoratory) NPP simulator facility (Lois et al., 2009; Bye et al.,2011; Dang et al, 2011). It should be noted that some crews did not complete the LOFW scenarios due to simulator problems.

In the US Study, there were three scenarios. Scenario 1 was a total LOFW followed by an SGTR, for which three HFEs were defined. Scenario 2 was a loss of component cooling water (CCW) and reactor cooling pump (RCP) sealwater, for which one HFE was defined. Scenario 3 was an SGTR scenario without further

complications, which one HFE was defined. Four crews from a participating U.S. NPP participated in the U.S. Study, and their performance data of four HFE scenarios were collected on the plant full-scope training simulator (Bye et al., 2012; Marble et al., 2012). It should be noted that one crew was unable to complete Scenario 3 (SGTR) due to a simulator problem.

The HFEs were ranked in terms of problems experienced by the crews in diagnosing and completing the actions based on the empirical data, as well as by opinions of the crew members who participated in the study. The crew failure rates and difficulty rankings for both studies are listed in Tables 1 to 3. It should be noted that HFEs 5B1 and 5B2 in the SGTR scenarios are exclusive, meaning that data were available for only one HFE for a particular crew.

Table 1. Crew Failure Rates and HFE Difficulty Ranking for the SGTR Scenarios in the International HRA Empirical Study

HFE	Failure Rate	Difficulty Ranking
HFE 5B1	7/7	5 (Very difficult)
HFE 1B	7/14	4 (Difficult)
HFE 3B	2/14	3.5 (Somewhat difficult)
HFE 3A	1/14	3 (Somewhat difficult)
HFE 1A	1/14	2.5 (Easy to Somewhat difficult)
HFE 2A	1/14	2.5 (Easy to Somewhat difficult)
HFE 2B	0/14	2.5 (Easy to Somewhat difficult)
HFE 5B2	0/7	2 (Easy)
HFE 4A	0/14	1 (Very easy)

Table 2. Crew Failure Rates and HFE Difficulty Ranking for the LOFW Scenarios in the International HRA Empirical Study*

HFE	Failure Rate	Difficulty Ranking
HFE 1B	7/10	5 (Very difficult)
HFE 2B	0/7	3.5 (Somewhat difficult to difficult)
HFE 1A	0/10	2.5 (Easy to somewhat difficult)

*Empirical data were not available for HFE 2A.

Table 3. Crew Failure Rates and HFE Difficulty Ranking for the Scenarios in the US HRA Empirical Study*

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)

*Empirical data were not available for HFE 1B.

Alongside the Bayesian uncertainty bounds derived from the crew data, the HEPs predicted by the two analyst teams are plotted in Figures 1 and 2 for the SGTR and LOFW scenarios, respectively. On the horizontal-axes, the HFEs are ordered by their difficulty ranking.

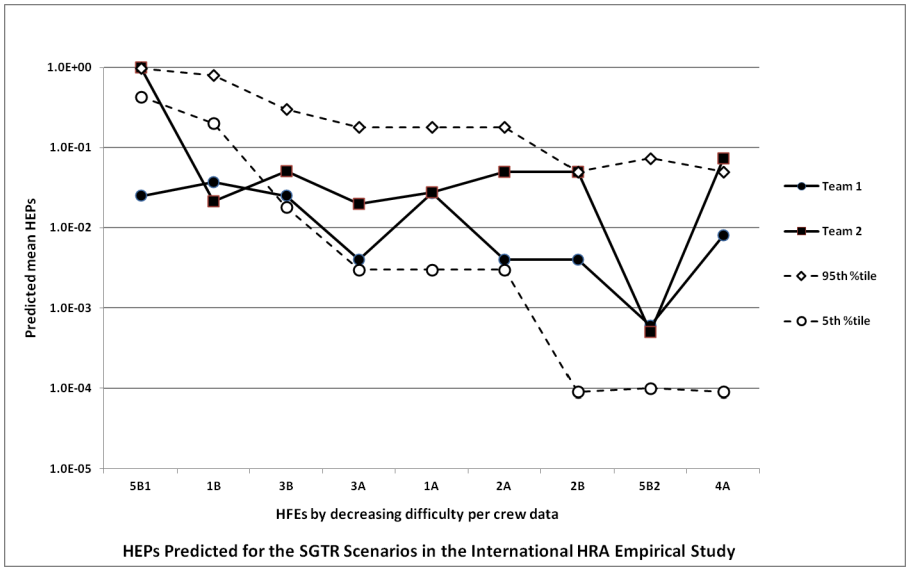

Figure 1. Predicted Mean HEPs with Bayesian Uncertainty Bounds for the SGTR Scenarios in the International HRA Empirical Study

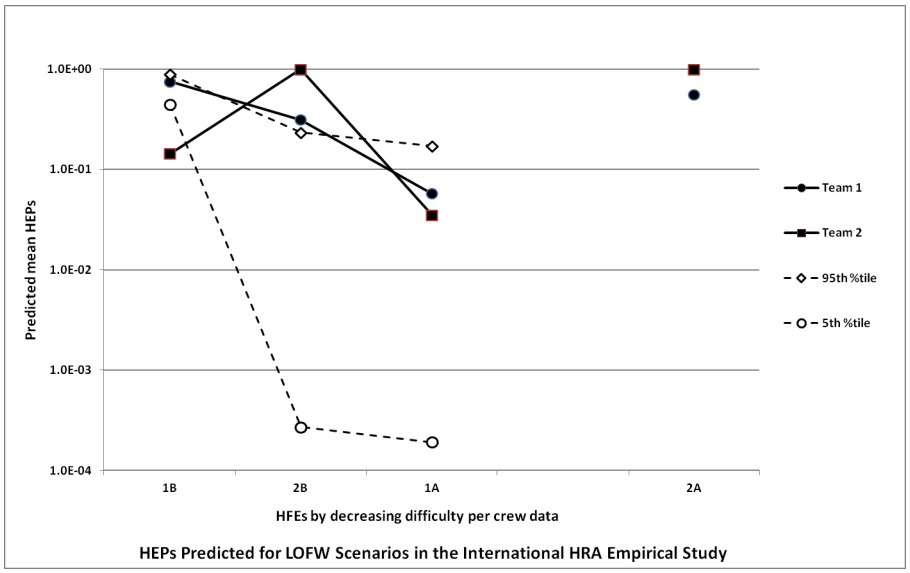

Figure 2. Predicted Mean HEPs with Bayesian Uncertainty Bounds for the LOFW Scenarios in the International HRA Empirical Study

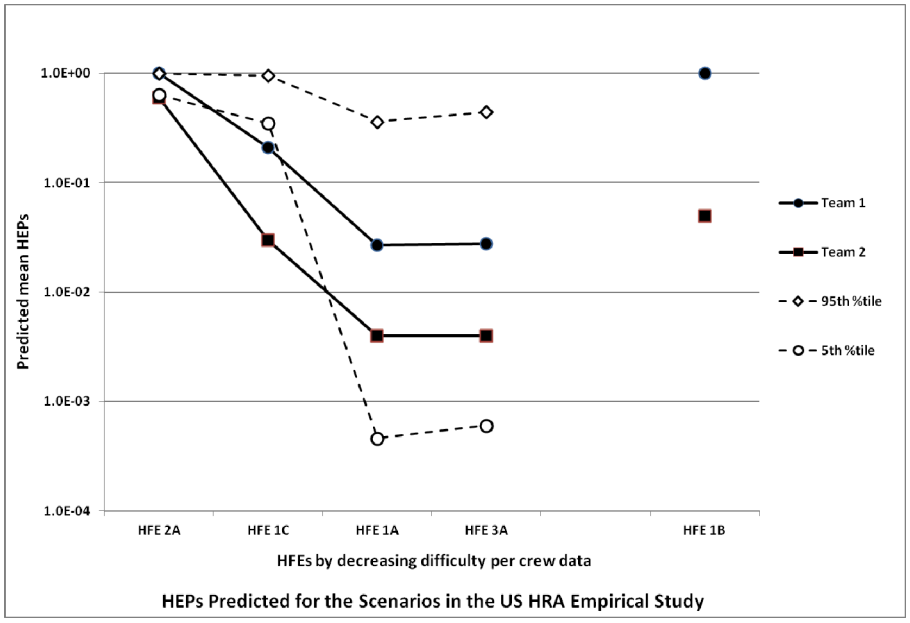

Figure 3. Predicted Mean HEPs with Bayesian Uncertainty Bounds for the Scenarios in the US HRA Empirical Study

3 INSIGHTS FROM THE STUDIES

In the two studies, ASEP was assessed to explore 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, and where analysts were varying in their implementation of the method (e.g., by going beyond the method guidance or not using the method as designed). The strengths and weaknesses for ASEP are summarized below.

3.1 Strenths

1) Simplicity

One strength of ASEP is ease of use due to its simplicity, such as simplifying its human performance model by separating diagnosis from post-diagnosis actions, estimation of the diagnosis HEP only with the TRC with a few PSF adjustments, and focus only on the major procedural steps without examining potential complexities in the sub-steps given the conditions of the scenario. On one hand, the simplifications make the method easy to use; on the other hand, they seem to contribute to the weaknesses discussed below. The simple analysis is justified by the developer by claiming that conservative HEPs will generally be obtained. However, apparent optimism due to the method's weaknesses was seen in some

cases (e.g., the analysis of HFEs 1B and 5B1 in the SGTR scenarios by Team 1 in the International Study) (see discussion below). The implication is that the tradeoffs between simplicity and thorough analysis need to be weighed before applications of the method.

2) Traceability

At some level, another strength of the method is its traceability. The estimation of allowable diagnosis time and allowable post-diagnosis time, the derivation of the HEP within the method, and what is important to performance given the factors considered is generally traceable, and how the various factors are weighted in determining the final HEP can be determined. However, how analysts might bias or alter the rating or level of the factors considered in applying the method, based on other information identified that is not covered by the method would be difficult to trace if the analysts do not document their decision process well.

3.2 Weaknesses

1) Insufficient guidance on when to consider cognitive demands in connection with the execution of a task.

By segmenting total time available for coping with an abnormal event into two artificially independent parts: *allowable diagnosis time* and *allowable post-diagnosis time (i.e., related to response execution 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. In the International Study, Team 1 assumed that no diagnosis was required once the crews entered symptom-based procedures in SGTR scenarios, and Team 2 made such an assumption in both SGTR and LOFW scenarios. The crew data showed that such an assumption was inappropriate as crews had to assess the situations and/or make new response plans while the scenarios progressed. Failure to address diagnosis seemed to be a major contributing factor to the team making predictions inconsistent with the empirical data in terms of performance drivers and operational stories. Apparently, the decision to skip the diagnosis part of crew response may have precluded the opportunity to address operators' cognitive activities, examine the difficult conditions operators would be facing, and identify some important factors influencing performance. As a consequence, the HRA teams only obtained a partial picture of the dynamic nature of the accident scenarios, which was seen in both teams' analyses in the International Study. For example, by focusing mainly on crews working through the procedures, Team 1 in the International Study did not really distinguish the strong difference between the conditions for HFEs 5B1 and 5B2 in the SGTR scenarios. In addition, except for the easiest HFEs 5B2 and 4A where there seems to be a good agreement between the predicted drivers and those identified from the crew data, the predicted negative drivers rarely matched those identified from the crew data in the SGTR scenarios. In contrast, the team identified many of the important drivers that would influence performance in the LOFW scenarios. Although this seems to be partly due to team's experience from

conducting HRAs for the SGTR scenarios (there might be a learning effect for Team 1 in the LOFW scenarios, because HRAs for the LOFW scenarios were conducted after the team saw the study results for the SGTR scenarios) and effort that went beyond ASEP guidance, addressing diagnosis in terms of the ASEP TRC did lead the team to obtain a good understanding of what would be going on in the scenarios and consider the potential impact of time available on the diagnosis.

Quantitatively, the final HEP in ASEP is the sum of the diagnosis and execution HEPs. Under the assumption of successful diagnosis, the final HEP is only determined by the probability of making an error in executing post-diagnosis actions, and thus can be optimistically estimated. Although it is difficult to estimate what the true HEPs are given the limited data, the optimism in the HEPs of the SGTR scenarios by Team 1 in the International Study is well illustrated in the HEP pattern (see Figure 1). For the most difficult HFEs 5B1 and 1B, the HEPs are below the lower Bayesian uncertainty bound. In particular, the HEP for HFE 5B1 is 0.025, which shows a large disconnection with the fact the all crews failed that HFE. In addition, the HEPs for HFEs 3B (0.025) and 3A (0.004) appear to be smaller than the actual crew failure rates (2 out of 14 crews failed in HFE 3B and 1 out of 14 crews failed in HFE 3A).

2) Inadequate guidance for estimating time requirements

The above trend toward optimism in Team 1's HEPs for SGTR scenarios in the International Study is interesting in that ASEP claims to provide generally conservative HEP values. However, where diagnosis was addressed, the HEPs for the LOFW scenarios in the International Study do seem to suggest conservatism. In particular, the HEP for HFE 2B (0.312) not only is above the upper Bayesian uncertainty bound, but also seems to be more conservative than appropriate given a zero crew failure rate. In this case, the main contributor to the conservatism seems to be the conservative assumption about the allowable diagnosis time in conjunction with the use of the ASEP TRC. It appears that more guidance on estimating time requirements and considering factors that could influence time requirements (e.g., concurrent activities) would strengthen the method.

3) Inadequate guidance to examine low level cognitive activities

Having learned the lesson from the International Study, the teams in the U.S. Study considered diagnosis in their analyses, and that seemed to help the teams' qualitative analyses. However, the U.S. Study has revealed that even if the analysts decide to explicitly address diagnosis in their analyses, there still is a chance for the analysts to overlook difficulties in cognitive activities. This is caused by the focus of ASEP on procedural steps at a high level (e.g., identification of the initiating event and entry into the appropriate EOP), rather than the diagnosis and cognitive activities involved in following and responding to the steps in the EOPs. That is, lower level cognitive activities, such as interpreting the plant status in the context of the step by step procedures and associated time-limiting conditions need more attention than given in evaluating post-diagnosis tasks. As a consequence, HRA predictions are likely to be limited to the crew's interaction with the main

procedural steps and lead to optimistic HRA results by ignoring the difficulties operators would face at the sub-step level.

4) Inadequate set of performance shaping factors (PSFs)

One primary purpose of qualitative analysis is to understand fully all possible sources of error and the underlying PSFs that impact the reliability of human performance. It could be argued that the analyses in the International Study might have been improved if the HRA teams had explicitly addressed diagnosis. However, even if diagnosis is explicitly included, the method still shows an inability to guide analysts to examine an adequate set of factors that could influence crew behavior for all circumstances. For example, the guidance to evaluate diagnosis/cognitive activities is minimal, and the method relies heavily on its diagnosis TRC with adjustment for only a few PSFs.

5) Inadequate guidance for choosing PSF levels

When addressing post-diagnosis actions, whether using ASEP or THERP (as mentioned above, ASEP allows the use of THERP in quantification with respect to post-diagnosis actions), decisions need to be made regarding the specific levels of PSFs relative to a given scenario/HFE must be made (e.g., stress levels and execution complexity). In both studies, differences across analysts were seen in the selection of PSF levels, and the differences led to observable variations in the HEPs for the same HFEs. The guidance on those decisions appears to be limited for some situations, which may explain why the teams' decisions on those factors did not appear to correspond well to the factors and conditions observed from crew performance data.

6) Limited insight for error reduction

In general, ASEP can be considered as a PSF-focused method. As mentioned above, it relies heavily on its diagnosis TRC with a few PSF adjustments to address diagnosis. This approach limits the method's ability to discover cognitive mechanisms and/or contextual factors that would lead to human failures, and thus limits its ability to offer insights for error reduction.

4 CONCLUSIONS

Based on discussion above, it is clear that the qualitative analysis performed to support HRA quantification is an important contributor to the adequacy of HRA predictions. This was particularly demonstrated when the method applications did not address the cognitive aspects of performance in implementing procedures even though the initial diagnosis had been completed.

The variability across analysts using ASEP seems to largely stem from analysts' decisions about how to apply various aspects of the method. As seen in the studies, analysts are often called upon to make decisions in their analyses, and the guidance of the method is not sufficient or specific enough, so that analysts may have to,

more or less, rely on their subjective judgment in interpretation of the guidance.

The differences between the analyst teams observed in the studies underscore the need to enhance the guidance for the application of ASEP. Furthermore, it suggests that piloting of the method (and of this guidance) in view of analyst-to-analyst reproducibility would be warranted.

REFERENCES

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

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

Dang, V. N., J. Forester, R. Boring, H. Broberg, S. Massaiu, J. Julius, I. Männistö, P. Nelson, E. Lois, and A. Bye. 2011. *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. To be issued as NUREG/IA-0216, Vol. 3.

Forester, J., V. N. Dang, A. Bye, R. Boring, H. Liao, and E. Lois. 2012. Conclusions on Human Reliability Analysis (HRA) Methods from the International HRA Empirical Study. Proceedings of the 11th International Probabilistic Safety Assessment and Management Conference, Helsinki, Finland.

Forester, J., Hildebrandt, M., Broberg, H., Nowell, R., Liao, H., Dang, V.N., Presley, M., Bye, A., Marble, J., Lois, E., Hallbert, B., and Morgan, T. (2011). *US HRA Empirical Study -- Comparison of HRA Method Predictions to Operating Crew Performance in a US Nuclear Power Plant Simulator and an Assessment of the Consistency of HRA Method Predictions* (Draft Report).

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

Marble, J., H. Liao, J. Forester, A. Bye, V. N. Dang, M. Presley, and E. Lois. 2012. Results and Insights Derived from the Intra-Method Comparisons of the US HRA Empirical Study. Proceedings of the 11th International Probabilistic Safety Assessment and Management Conference, Helsinki, Finland.

Swain, A. D., and H. E. Guttman. 1983. *Handbook of human reliability analysis with emphasis on nuclear power plant applications*. NUREG/CR-1278-F, U.S. Nuclear Regulatory Commission.

Swain, A.D. (1987). *Accident Sequence Evaluation Program Human Reliability Analysis Procedure*. NUREG/CR-4772/SAND86-1996, Sandia National Laboratories for the U.S. Nuclear Regulatory Commission, Washington, DC, February 1987.