SAND2007-3552C

Human Reliability Analysis (HRA) in the Context of HRA Validation/Benchmarking*

J. Forester[1], A. Kolaczkowski[2], V.N. Dang[3], E. Lois[4]
[1]Sandia National Laboratories, Albuquerque, NM, jafores@sandia.gov
[2] SAIC, Eugene, OR, USA
[3]Paul Scherrer Institute, Switzerland
[4] United States Nuclear Regulatory Commission, Rockville MD, USA

Abstract— **Given the significant differences in the scope, approach, and underlying models of a relatively wide range of existing HRA methods, there has been a growing interest on the part of HRA method developers and users to empirically validate and test the reliability of the various methods. To this end, there is an ongoing international effort to begin this process by testing the application of HRA methods to nuclear power plant operating crew performance in the HAMMLAB simulators at the Halden Reactor Project in Norway. Initial efforts in designing and implementing these studies have identified a number of issues associated with structuring the studies in order to allow an adequate and appropriate test of the different methods. This paper will focus on issues associated with applying HRA methods in the context of a validation study, particularly when a research simulator is used for data collection. Example issues include: method scope differences, differences in the use of simulator exercises to support the analysis, impact of experimental controls on application of methods, and given the low probability of human failure events typically modelled in nuclear power plant Probabilistic Risk/Safety Assessments (PRAs/PSAs), the correlated need for method users to represent the results of their analysis in a somewhat different form than usual. These and other issues related to applying HRA methods in the context of validation studies will be discussed and potential resolutions will be addressed.**

I. Introduction

Given the significant differences in the scope, approach, and underlying models of a relatively wide range of existing human reliability analysis (HRA) methods, there has been a growing interest on the part of HRA method developers and users (including PRA analysts that typically rely on HRA results) to empirically validate and test the consistency/reliability of the various methods. In addition, such validation and testing is vital to adding credibility in HRA results when decision makers have to use those results to make risk-informed decisions. To this end, there is an ongoing international effort to begin this process by testing the application of HRA methods to nuclear power plant operating crew performance in the HAMMLAB simulators at the Halden Reactor Project in Norway.

Initial efforts in designing and implementing these studies (a pilot study is under way) have identified a number of issues associated with structuring the studies in order to allow an adequate and appropriate test of the different methods. These issues can impact (a) the ability to test the consistency of HRA results across the different methods and across the same methods using different analysis teams, and (b) the ability to test the validity or accuracy of HRA results by comparing the output of the different methods to the results of crew performance in the simulator. This paper focuses on the issues associated with analysis teams applying HRA methods in the context of a validation study, particularly when a research simulator is used for data collection. That is, the paper addresses issues associated with teams applying HRA methods and producing useful results that allow for fair comparisons between the methods, with respect to the performance of the operating crews in the simulations of nuclear power plant scenarios. The study design adopted in this work represents an attempt to resolve these issues and the resolutions are discussed in the corresponding sections.

II. Method Scope Differences

A. *HRA Processes Covered by the Methods*

Most existing HRA methods were initially developed to support the consideration of human performance in the context of PRAs/PSAs of nuclear power plants. However, not all of the methods address all aspects of performing an HRA in this context. For example, several HRA methods emphasize and provide guidance for identifying and incorporating human failure events (including errors of commission) into the PRA models, in addition to providing guidance for quantifying these human failure events (HFEs). Most other HRA methods only

*This work was funded by the U.S. Nuclear Regulatory Commission (USNRC) and performed in part at Sandia National Laboratories. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000. The work was also supported by the Swiss Federal Nuclear Inspectorate (DIS-Vertag Nr. 82610). The opinions expressed in this paper are those of the authors and not necessarily those of the USNRC or of the authors' organizations.

address the quantification process, leaving other parts of the HRA process to be covered by other sources as needed. While it might therefore seem that most of the methods can only (or at least) be straightforwardly tested and compared on the quantification process, which they all generally cover, some methods claim that performing these other parts of the analysis are important to the validity of the method's quantification process.

For example, the ATHEANA method [1] places significant importance on the processes of

- identifying the HFEs to be modeled in a particular scenario,
- deciding how to represent them in the scenario,
- identifying potential errors of commission (EOCs) that should be modeled, and
- identifying potential deviation scenarios that could affect the likelihood of crew success for a given HFE.

From the ATHEANA perspective, these analyses can have an important impact on the eventual quantification of an HFE. Users of the MERMOS [2] method would likely make related arguments.

However, in the context of an experimental study and the corresponding need for adequate controls to ensure that differences in method results can be understood, it may be necessary to initially constrain what the methods are allowed to do. That is, in order to do a reliable test of the validity of the quantification process, the scope differences of the methods may need to be managed to some extent. For example, to ensure that all of the methods are quantifying the same event and conditions, it may be necessary to carefully define the HFE and its context, which may be inconsistent with the usual application of some methods.

The only clear solution to this type of problem is to perform a series of studies (or create separate experimental conditions if practical) where such scope differences between methods are systematically allowed to come into play. In this way the effects of the scope differences can be tracked. For example, if the emphasis is on the validity of the quantification process of the methods, then it may be necessary to initially constrain methods such as ATHEANA and MERMOS to their basic quantification process, without the benefit of using their other processes. In later studies, these aspects could be systematically introduced to assess their affects on the validity of the results.

B. Information Collection

Another method scope issue relates to the fact that the different methods tend to have different bases and make different underlying assumptions about human behavior. These basic method differences can lead to differences in the information needed to perform the analysis. Thus, different teams may ask different questions and may need different levels of support from key plant personnel such as operators and trainers. On the face of it, one might argue that this is fine in the context of a validation study because the goal is to assess which methods provide sound results. However, there are constraints involved in using a simulator test bed (see our other paper) and providing adequate controls for this type of research, that raises concerns about the different methods using different information. Differences in method results should be due to the methods per se, and not due to artifacts of the difficulties of experimental design for this kind of research. One of the earliest HRA benchmarking studies [3] was criticized on the grounds that the teams produced different results because they used different assumptions that may have been related to the study design rather than to the methods per se.

Similar to item A above, good experimental design to address this issue will require a series of conditions or converging studies to be performed. In the current pilot study, the decision was made to provide the same information to all HRA teams. An initial set of information for each HFE to be quantified was provided to the teams and they were then allowed to ask questions. However, all questions and answers were shared among the teams. This approach at least ensures that all of the teams have the same information available and that they would tend to make consistent assumptions. While some of the methods may not be able to directly use all of the information available or benefit from the questions asked by different teams, differences in the method results will be more directly attributable to the methods (or at least to the teams) rather than to potential inconsistencies in the information available caused by other reasons. To some extent, this approach strives to maximize the potential for obtaining consistent results across the methods. Once a baseline of method results is obtained in this way (with information level held constant), follow-on studies can begin to allow the different methods to use only the additional information they request, so that method differences can be more explicitly tracked.

III. ACCESS TO PLANT PERSONNEL AND OBSERVATION OF SIMULATOR EXERCISES TO SUPPORT METHOD APPLICATION

Although related to method scope differences in some ways, the use of the HAMMLB simulator in Norway, along with crews from different countries visiting for limited periods of time, can limit the ability of some teams to do all that they would normally do in an application of their method. For example, some methods would ideally want to observe crews performing the actual scenario they are trying to model (e.g., MERMOS [2] and HCR/ORE [4]) and most methods would at least like to observe some crews operating in sample scenarios. In addition, many methods emphasize the importance of being able to interview (debrief) crews after observing their performance in sample scenarios. Methods like ATHEANA [1] enlist operators and trainers from the target plant to participate directly in the quantification process. In the context of the HAMMLAB simulations, such activities are not always possible for multiple practical reasons.

While there may not always be perfect solutions to some of the problems, some steps can be taken to various degrees depending on the resources available. Ideally, all of the HRA teams could go to Halden at important points during the study and collect the desired information. One partial and more practical solution is to have the study team provide much of the information that would be obtained from simulations and debriefs as part of the information input to the HRA teams. In addition, actual plant personnel can be asked to respond to questions from the HRA teams at a later time and when possible, tapes of crews participating in the simulations can also be distributed. And as needed, personnel from similar plants or with similar experience could be asked to support aspects of the analysis, e.g., the quantification process. Yet, even in the context of a plant specific PRA/HRA, it is seldom possible for analysts to view crews in all of the scenarios being modeled or even to discuss all of the potential scenarios. Similarly, it is usually not possible to have operators and trainers participating in all aspects of an HRA.

Thus, although numerous steps can be taken to address such gaps in the experimental setting, it should be remembered that HRA is seldom performed under ideal conditions. There are always going to be some limitations in a given application. Given this is the case, it would seem that different HRA methods should still be able to produce reasonable results in spite of some limitations, albeit with somewhat more variability in results than might be desirable. If the methods cannot handle some degree of variation in the inputs received, it could be concluded that the methods will not be very reliable or valid.

IV. Form of HRA Predictions

A. Effect of Highly Reliable Perfromance

Human performance levels in nuclear power plant operations are generally high because of the comprehensive analysis of potential accident scenarios, the support provided to the operators by abnormal and emergency procedures, and the extensive, on-going training programs. As a result, operator crews are expected to make few errors in most scenarios; moreover, un-recovered errors are expected to be even less likely (the probability of HFEs in PRAs/PSAs in general accounts for recovery factors, for instance, that a fellow crew member detects an omission).

The scenarios that are selected for HRA-related studies at Halden are designed to present a significant challenge to the operators, relative to, for instance, design basis accidents. Still, this increased level of difficulty results mainly in more variability of performance that is generally successful. The high levels of human performance in these scenarios means that the observations in Hammlab cannot generally be used to obtain an "observed" human error probability (HEP) that could generally be compared to the HEP predicted by one HRA team with a given method.

Naturally, it would in principle be possible to design scenarios that are so difficult that a significant number of the operating crews could be expected to fail on the same task (HFE). In this context, a significant number of the crews means 20 to 50% or more of the crews, corresponding to HEPs of 0.1 or larger. Such scenarios, however, may not be especially interesting for HRA; the consensus is that most methods would yield HEPs of this magnitude for these types of scenarios. An additional problem is that the probability of such scenarios is generally very low; in other words, they often require unlikely combinations of independent failures.

Thus, the highly reliable performance expected in HRA-related scenarios means that a comparison of an observed HEP for an HFE with predicted HEPs may not always be possible. While a comparison of the HEPs obtained by the different teams using different methods on a given HFE is of course possible, the desire to validate the methods against actual data necessitates that more qualitative predictions of the HRA analyses are also included and compared with the results observed in the simulator.

One of the more obvious types of qualitative information produced by HRA methods are the performance shaping factors (PSFs) identified as driving performance in a given scenario and contributing to the HEP. In principle, the HRA teams can be asked to document the PSFs identified as being important and to discuss how and why these factors should affect crew performance in the context of a scenario. They can also be asked to describe why a scenario might be difficult or easy for the crews and why the scenarios might be confusing in terms of the specifics of the scenarios. To the extent that the results from the crew simulations can be described in ways that will allow comparisons with the output of the methods, the validity of the methods' predictions can be assessed.

However, different methods sometimes address different PSFs (arguing that their set of PSFs are adequate to predict performance) and in many cases, the methods themselves do not ask the analysts to produce the results in the ways described above (i.e., that are needed to be able to compare results from the simulator or between the different methods). Thus, in the context of a validation study, in some cases the HRA teams may be asked to represent the results of their analysis in ways they may not do so normally.

While asking for their results in this way may require the teams to go beyond the specifics of the method to some extent, it can be argued that doing so does not exceed what is normally expected from the application of an HRA method. Certainly a major goal of a HRA is to identify when support for the crews' response is adequate (leading to high success rates) and when potential problems might arise and why. To be valid, HRA methods should be able to provide reasonable understandings of what should occur in accident scenarios. Thus, it is expected that HRA teams should in general be able to provide the needed discussions and that this information can be compared to the simulator results and the results of the other methods.

Nevertheless, to provide additional support for assessing the validity of the HRA methods, additional steps are possible. One is to provide the HRA teams with a common template for representing the results of their analysis. In the current pilot study, several aspects of the HERA database [5] taxonomy were included in a form for the HRA teams to fill out. In particular, the wide range of PSFs and categories for describing their impact on performance were included from HERA. The goal was to provide a common terminology in which the different teams could represent their results and support their ability to describe what is expected to happen in a given scenario. Not only will this allow for more straightforward comparisons between the results of the methods (which often use somewhat different terminology), but by having the Halden experimenters fill out the same forms based on their observations of crew performance, a more direct comparison between the results from the methods and those from the simulator is possible.

B. *Use of Extraneous Human Perfomance Models*

A related issue is that some analysts may use their own or other human performance models to support their application of a given HRA method, while others may be very method bound (i.e., very constrained) in terms of the factors considered and evaluated. In some cases, such differences may be due to the experience levels of the HRA teams or to the differences in the expertise of some of the members (e.g., researchers in the area of HRA versus those that are primarily users). To some extent, such diffferences may be controlled for by requiring comparable levels of expertise and background across the different HRA teams. However, the scarcity of available team members that have experience in applying a given method can sometimes limit the ability to match teams along all relevant dimensions. In addition, it can be argued that appropriate use of an HRA method should involve thinking beyond the method to some degree, so that a clear understanding of what might go on in a scenario can be obtained. The method can then be used to transfer the obtained information into HEPs.

Another way to potentially help control for this issue is to provide (as was done in the current pilot study) a common taxonomy for all of the HRA teams to represent not only the output of the method, but also a means to address a broader range of performance descriptors, whether from an extraneous human performance model or from a direct evaluation of the factors in the common taxonomy. And as noted above, this may allow a better comparison of the results of an HRA analysis with the results from the simulator. Whether such manipulations provide adequate controls will be an important finding from the current pilot study.

V. Representaion of Scenarios and HFEs

The final issue to be discussed related to applying HRA methods in the context of a validation study, concerns questions or disagreements from the HRA teams on how the scenarios and HFEs are represented by the experimental design team. In order to test the HRA methods, the experimenters must identify reasonable scenarios and appropriate HFEs for the HRA teams to model. However, in performing PRAs, there may be different models possible for a given initiating event and different ways to represent the success criteria for a particular HFE. Thus, it becomes possible that not all of the HRA teams will concur with how a given scenario and its success criteria are represented. While the HRA teams can be asked to address the scenarios and HFEs as modeled, it is possible that they may think that the representation is inconsistent with how they expect the crews to approach the scenario, thus affecting the way they would model the scenario. In other words, the fact the experimenters chose a representation and particular success criteria, does necessarily mean that it reflects how the crews will consider the scenarios.

To help avoid creating such problems for the HRA teams and for the study itself, a potential solution to the problem is for the experimenters to seek confirmation from knowledgeable plant personnel and PRA modelers that the scenarios are consistent with how the participating crews will think about the scenarios and what their expectations for the success criteria will be. While the results of such steps may not always prevent problems, it is important that experimenters do all that they can to identify the scenarios and HFEs in ways consistent with the characteristics of the plant modeled in the simulator and the participating crews' training..

VI. Conclusions

The goal of this paper was to discuss some of the issues associated with the application of HRA methods in the context of a HRA validation study, particularly when a research simulator is used for data collection. As described above, there are various adjustments to the study design that can be made to alleviate the impact of some of the issues and allow reasonable conclusions to be made from a given experiment. However, perhaps the most important conclusion is that because of the complexity involved in conducting this type of empirical research, it is clear that a single experiment is not going to provide a final set of answers regarding the validity and reliability of the various HRA methods. A set of converging operations will be required to control for the range of factors that can affect the application of the HRA methods and to control for the potential impact of other experimental design issues. In other words, in order to be able to adequately control for the range of factors that could confound the conclusions from the study, a series of experiments will be required to cover the relevant conditions. The results from initial experiments can be used to guide the design of later experiments as needed to allow empirically based conclusions about the validity and reliability of the methods. Although the discussions in this paper point out the complexity of the problem, it is thought that reasonable conclusions can eventually be made. The results of the ongoing pilot study should provide a good baseline for the potential of achieving the goal of validating HRA methods (at least partially) and testing their reliability, along with an idea of the level of resources needed to achieve the goal.

Acknowledgment

We acknowledge the contributions of Andreas Bye, Helena Broberg, Per Øivind Braarud and Salvatore Massaiu, the Halden Project, Bruce Hallbert and Ron Boring, Idaho National Laboratory, Jeff Julius, Scientech, and Gareth Parry of the US Nuclear Regulatory Commission, for major parts of the work done so far in the project. The views expressed in this paper are only those of the authors and do not necessarily represent the views the organizations they represent.

References

[1] "Technical Basis and Implementation Guidelines for A Technique for Human Event Analysis (ATHEANA)," NUREG-1624, Rev. 1., USNRC, Washington, D.C., May 2000B.

[2] C. Bieder, P. Le-Bot, E. Desmares, J.-L. Bonnet, and F. Cara, "MERMOS: EDF's new advanced HRA method," in Probabilistic Safety Assessment and Management (PSAM 4), A. Mosleh and R.A. Bari (Eds), Springer-Verlag, New York, 1998.

[3] A. Poucet, “The European benchmark exercise on human reliability analysis.” Proceedings of the American Nuclear Society International Topical Meeting on Probability, Reliability, and Safety Assessment (PSA ’89), 103-110, 1989.

[4] G. Parry, B.O.Y. Lydell, A.J. Spurgin, P. Moieni, A. Beare, “An Approach to the Analysis of Operator Actions in PRA,” EPRI TR-100259, Electric Power Research Institute, Palo Alto, CA, 1992..

[5] R. B. Hallbert, R. Boring, D. Gertman, D. Dudenhoeffer, A. Whaley, J. Marble, J. Joe1, E. Lois, “Human Event Repository and Analysis (HERA) System – Overview”, NUREG/CR-6903, Vol. 1, INL/EXT-06-11528, prepared by Idahol National Laboratory for the U.S. Nuclear Regulatory Commission, Washington, DC, USA, July 2006.