Informing HRA by Empirical Data, Halden Reactor Project Lessons Learned and Future Direction

Andreas Bye
OECD Halden Reactor Project, IFE, Halden, Norway

Abstract: Human Reliability Analysis (HRA) needs data to validate HRA models and basic features of methods, and to support consistent use of the HRA methods. Knowledge on how the context can impact human performance will influence the way in which the methods should be applied in the right way.
This paper presents a history of how the Halden Reactor Project has provided data for HRA in various ways; firstly through empirical studies to evaluate HRA methods; secondly with data to support the general knowledge of HRA practitioners through qualitative influential details of scenarios; thirdly with data methods to support basic questions such as to what extent digital systems are more error prone than analog systems. Lessons are drawn from this history with respect to the way in which data may support HRA, and a few questions are posed on future methodologies for improving HRA, such as: Can big data technology and increased amounts of data support HRA in any form?

Keywords: HRA, Empirical Data, Simulator Runs, HRA Empirical Studies

1. INTRODUCTION

The role of data in support of Human Reliability Analysis (HRA) has for a long time been discussed and many actors have requested more data to support HRA. The original need of data arose when first developing the methods. HRA methods quantify the probability of success and failure of Human Failure Events (HFEs), an event in which human action plays a role. Quantification is typically based on a methodology where a basic human error probability (HEP) for a specific or a general task is adjusted by taking into account the context of the task. Thus, two main needs for data occurred: 1) Data to estimate the basic error probability for a task; and 2) data to support the right adjustment of the basic error probability depending on the context for the task. Regarding data for 1), Alan Swain collected data and elaborated on different sources in the initial work on THERP [1] that other HRA methods have been building much of their quantitative modules on, as described by Boring [2]. Data needs in this connection have been to validate the data for the nuclear power plant setting (since a large part of these data were based on other industries), an exercise that was extensively performed in the 1980s, as well as more recently to renew the data for other types of basic tasks, e.g., tasks in modern digital control rooms. Regarding data for 2) the impact of the context on the task in its relevant scenario, work has been going on continuously up to this date, basically as part of development of various HRA methods.

When using an HRA method of the general type as described above, two questions are central to the HRA practitioner. She/he needs to identify the presence of the context (in some methods called PSFs) in the situation, and given this presence, the impact of the context on the task. The latter is partly based on the description and guidance in the method itself, but it is also dependent on the knowledge and experience of the practitioner. Thus, an important role for data in general is also to extend the knowledge of HRA practitioners so that they will know better how to account for the context of a task in the use of the HRA method at hand. The context includes the impact of time, procedures, complexity etc. on the performance of actions by the crew.

Thus two major needs for data have evolved: Data in order to validate HRA models and basic features of methods, and data to support consistent use of the HRA methods. A third case has also arised, especially in the new "big data" era: If we collect enough data, can these be used in a way so that the use of HRA methods can be practically unnecessary? Could we use data from earlier experience and

extrapolate these into probability estimates for predictions in future cases? In a way, this extrapolation is currently done by the use of HRA methods. In Bye et al. [3] we discussed this direct data possibility. We concluded that *"One would have to run an enormous number of scenarios in order to get the validity acceptable, so the practical use of this approach is very limited."* [ibid., pg. 2]. Has the "big data" trend in recent years lead to a change in our position on this? This is discussed in section 3 in this paper.

2. THREE CASES ON DATA FOR HRA

2.1. Empirical studies to evaluate HRA methods

In order to validate HRA models, a major effort was done in the "International HRA Empirical Study" [4] and [5], and followed up in the "U.S. HRA Empirical Study" [6]. In these studies the predicted output from the HRA methods were directly compared to empirical data for crew performance in the same scenarios. The purpose was to identify strengths and weaknesses of the HRA methods. In the international study, the empirical data was human performance measures and observations of 14 licensed operating crews performing scenarios in the HAlden Man-Machine LABoratory (HAMMLAB) simulator. Without knowing the result of the simulator runs, but only with descriptions of scenarios, 13 HRA teams predicted the outcome of the same scenarios. Most HRA methods were only represented by one HRA team, but a few methods were performed by two teams. Thus, on the HRA team side of things, there was not enough data to perform quantitative comparisons, but the outcome was studied in a qualitative way. The operating crews all ran the same scenarios in an experimental set-up, with basic (textbook) variants of the scenarios in addition to complex variants of the same scenarios. Thus, on this side we had a more quantitative basis in the data. The scenarios run were basic and complex variants of steam generator tube rupture (SGTR) and total loss of feedwater (LOFW). There were six HFEs defined for the LOFW, and nine HFEs for the SGTR scenarios. See [4] and [7,8,9] for extensive descriptions of the methodology.

The follow-up study in the U.S. [6] was established in order to look more into variability between HRA teams using the same HRA method, so several HRA teams used the same method in this study. In addition, we looked for confirmation of findings from the international study, so the same scenarios were utilized. A training simulator at a U.S. plant was used, and the methodology was slightly improved and changed relating to the way in which the HRA teams could get information from the plant (e.g., by plant visits and more interviews). However, the treatment and analysis of the empirical data from the operating crews was not very different, and also turned out to be very similar in terms of utility for the exercise. Thus, in this paper I will mainly use the international study and its conclusions as the basis for discussion.

The empirical data had three basic forms, quantitative data as well as two types of qualitative data. We could count the number of failures of the human failure events (HFEs) that the crews did, thus, having a quite good estimation of the human error probability (HEP) of each of the HFEs in question. Most HFEs had 14 crews performing them, while some of them had only 7 crews running the simulation for that HFE. 14 out of 14 crews managing to do an action is not very much information when discussing small probabilities. However, in some of the really complex cases many crews failed our defined HFEs, a few instances were present with all the crews failing the defined HFE. We used Bayesian statistics, and 7 out of 7 failing gives a strong update on the probability. Thus, for the difficult HFEs, where many of the crews failed, the comparison to the predictions could tell quite a lot about whether an HRA method missed the mark. For the easy HFEs, the uncertainty band became very broad, we could not know whether 14 out of 14 crews managing an HFE was a 10E-2 or a 10E-5 number. See the Figure 1 below, in which the uncertainty band is drawn (note that we did not use point estimates, but rather the uncertainty bands in the comparison). Concluding, the numbers gave us stronger evidence as to whether the HRA teams managed to predict the difficult actions (HFEs) with their HRA method than the easy actions. NUREG-2127 concludes [4, pg. 99]:

> *"The human reliability analysis (HRA) Empirical Study is designed around a simulator study with up to 14 licensed operator crews from two units of the participating nuclear power plant.*

In the steam generator tube rupture (SGTR) phase, 14 observations of both scenarios were made, while 10 observations were obtained for both of the loss of feedwater (LOFW) scenarios. While this makes the simulator study remarkably large, the sample size for deriving reference human error probabilities (HEPs) from the evidence remains small. Quantitatively, the data represent a mixture of strong and weak quantitative evidence. When a failure is observed in a small sample, the evidence is strong; on the other hand, when no failures are observed, the evidence for the HEP is weak. Specifically, this means that the simulator data does not provide strong quantitative evidence for HEPs much lower than 0.05 to 0.1; in other words, for human failure events (HFEs) where the performance of the crews easily meets the HFE success criteria, the quantitative observations concerning the number of crews not meeting the success criteria would be "consistent" with a range of several orders of magnitude for predicted HEPs."

Figure 1. Bayesian confidence bounds of the SGTR empirical HEPs.
Adapted from NUREG-2127 [4, Fig. 4-5, pg. 50]

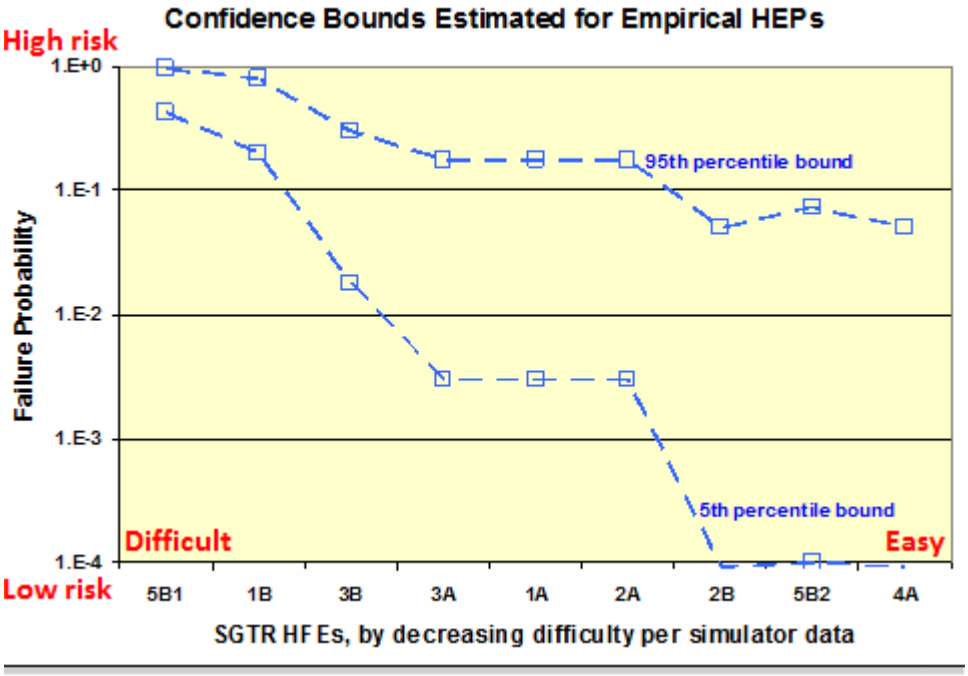

Fig. 1 shows all the nine SGTR HFEs ranked by difficulty, and the ranking was not only decided by quantitative means, but also qualitative [4, pg. 99]:

"In the study, these limitations on the purely quantitative aspects of the reference data were addressed with (1) a combination of benchmarking methodology features that accounted for qualitative evidence as well as a qualitative ranking of the HFEs, and (2) a selection of scenarios and HFEs that included HFEs representing a range of difficulties, and included very adverse, challenging scenario contexts."

On the usefulness of the quantitative and the qualitative parts of the empirical data, NUREG-2127 notes [ibid., pg. 106]:

" it should be noted that the quantitative empirical data and comparisons obtained in the study greatly supplemented the qualitative comparisons and insights identified. This was the case, even though there was a relatively small set of observations (at least from a statistical perspective), in terms of numbers of crews and scenarios. Generally, the quantitative empirical data and comparisons gave a very good starting point for assessing the qualitative predictions of the methods by prioritizing these qualitative findings and providing a measure of the significance of the predicted or observed performance issues. Thus, the present study

demonstrated that important information on HRAs and HRA methods can be obtained without using impractically large numbers of operating crews and scenarios, which is an important achievement."

In the evaluation of the HRA methods, an important point was to see whether the HRA teams managed to identify the difficult HFEs with their HRA method. In case they did not, we needed to understand why they missed out. For this, a deep qualitative analysis of the empirical data was needed. For this purpose, the simulator data was excellent: When studying the crews performing the scenarios in recorded videos, we could find out where exactly they got problems. This was documented in so called "crew stories". For example, we could see whether the crews missed important information and which parts of the procedures that gave them trouble. One recurring theme was procedure – situation mismatch, that the procedures did not cover the situation to a good enough extent and that the crews, even if they did follow procedures, did not manage to solve the situation. A typical example was the HFE-1B, defined as *"Failure of the crew to identify and isolate the ruptured SG in the complex SGTR"* [4, pg. 10]. In the complex case, there were no radiation indications in the secondary circuit, these were all either inhibited (failed signals) or masked by earlier events, in this case a steam line break just before the SGTR. [Ibid., pg. 29] describes the difficulty: *"The crews showed difficulties in identifying the presence of an SGTR, due to the concomitant steam line break and absence of radiation indications. The majority of the successful crews did not transfer to E-3 (SGTR procedure) to follow a transfer condition in the procedure set, but instead diagnosed the situation by interpreting the available indications on the plant status, with a rising SG1 level as the primary cue."* This led to large variations in the procedure paths for the crews.

The HRA teams/methods that did not manage to identify these difficult steps in the scenario development, did not get the numbers right. In order for the HRA teams to identify this, they would have to get their qualitative scenario analysis right. Independent of which HRA method that was used in the quantification, the teams that did not do a proper qualitative scenario analysis did not manage to pinpoint the difficulties that the crews would be expected to experience. Especially, some of the PSF-based methods are prone to error in these examples, since there seldom is any prescriptions in the methods themselves to evaluate the PSFs, such as procedures or complexity, on a too superfluous level and thus not utilizing the right multipliers that would account for these difficulties as needed in the detailed situation. When we evaluated the HRA methods, we especially looked for: Did the method prescribe the right way of exploring the scenarios so that the HRA teams could have a chance to identify the difficult places in the scenarios? As described above, the qualitative depths of the empirical data helped us to evaluate the HRA methods in this way.

The empirical data was also utilized to try to estimate the performing shaping factors (PSFs) that could explain why the crews had failed the various HFEs. Was it due to lack of time, bad procedures, or perhaps a bad human-system interface? This question seems rather simple, and we started out believing that we should be able to easily discover the cause for the failures that we observed in the scenarios. However, this task turned out to be rather difficult. Even though we many times could find an explanatory story for where and when things went wrong, it was not easy to isolate this to a single factor or several clear causal factors. Quite often, several factors were present and seemed to cause the performance, sometimes also the factors interacted. For example, the procedure could be non-optimal, but if they would have had better training, they could have managed anyway. Even though the masking of one symptom due to a faulty sensor was one contributing cause, training and procedures anyway played a role. We often found that training trumped all, since it is actually possible to train workarounds on non-optimal procedures or HSI. Thus, it was not easy to attribute an outcome to a factor, since many factors could be claimed as proximate or distal causes for the failures. This is also a reflection of the fact that a nuclear power plant control room is a very robust system. This challenge in the methodology might not be a complete surprise for the reader skilled in psychological experiments, since in order to find causal explanations one should set up a manipulation for one or two variables and control for other influences in order to claim causation. Our experimental set-up was not made for identifying all of these causal factors. However, there was one manipulation, and that was task

complexity, since each scenario had a basic and a complex version. It turned out that this manipulation had the expected effect, as can be seen in Figure 1.

NUREG-2127 [4, pg. 101-102] concludes:

> *"It should be noted that some of the performance factors used in various predictive HRA methods were problematic in the simulator study. Whereas the crew observations could be used straightforwardly to determine both the presence of procedural guidance issues (applicability in the scenario, clarity of the wording) and their impact on crew performance, the relationship between an adverse performance factor and its impact on the crew responses could not be established for all performance factors. The impact of such factors as "time pressure" and "experience" are frequently not directly observable, even though there is evidence that the factor is adverse. The validity of evaluating such factors in estimating HEPs is not questioned. Instead, the issue is that it can be difficult in this type of simulator study, where the possibility of controlling for these factors is limited, to determine from the empirical evidence whether and how the presence of the adverse factors affected crew performance."*

Although not all factors can be evaluated in each study, when more studies are performed and more cumulative evidence is gained about the impact of PSFs on human performance, HRA method developers can use this accumulated knowledge to update their HRA methods.

The most useful part of applying empirical data for evaluation of HRA methods in the study, appeared to be related to the operational issues and challenges for the crews [ibid., pg. 102]:

> *"Addressing the operational issues and challenges related to the performance factors (i.e., underlying the rating of the performance factors) was very effective in the study. First, it provided a transparent basis for the performance factors, in light of the fact that the definitions of the performance factors (and their scope) varied among the methods. Secondly, it provided more specific information on how the HRA teams understood the HFEs and the expected performance of the crews. For instance, with respect to procedural guidance, the assessments could take into account whether the analysis teams had identified the specific procedure steps and the aspects of these steps that would lead to potential problems for the crews. Addressing the operational issues underlying performance factor ratings ensured an evidence-based assessment of the qualitative predictions."*

This way of using empirical data for evaluating HRA methods, also links to the next use of such data, to inform HRA practitioners about notable aspects of scenarios, and the way in which these impact operating crews. This could actually be seen as a part of training HRA practitioners and improving their knowledge. More about this in the next section.

2.2. Data to support the general knowledge of HRA practitioners through qualitative influential details of scenarios

How can empirical data be used to inform the use of HRA methods, e.g., through general qualitative lessons learned from simulator studies? NUREG-2127 [4, pg. xxiii] concludes:

> *"With respect to improving human performance at the plants, the study provides a strong indication that challenging situations, such as those modeled in a PRA, should be regularly examined to improve plant design features, as well as operational features involving procedures, training, communications, team interactions, and leadership. The study's extensive documentation of crew performance in the simulator is also a rich source of information for practitioners dealing with human performance in general."*

During the years since 2006, the Halden Project has performed many experiments in HAMMLAB and in training simulators at plants. More than ten major data collections, sixty operating crews including 11 U.S. crews, 30+ scenarios including variants forming more than 250 simulator runs have been conducted. Typically, each crew has been in HAMMLAB for one week each running scenarios. The studies have focused on complex scenarios and how the crews solve these in a realistic environment.

Topics have included masking and complexity, as described by Braarud & Johansson in [10]. The main conclusion from this study was that team cognition was more important for diagnosis time in complex situations than in more prototypical situations (like the base scenarios that are normally trained). When team cognition was rated high, the diagnosis time was shorter in the complex situations, while in the base scenarios the diagnosis time was independent of the team cognition rating. Team cognition was in this experiment defined in dimensions including: "*Mission analysis - Cognition beyond procedure guidance; Process of consultation while performing technical work; Distributed leadership (mainly between Supervisor and Reactor operator); Team orientation; Backup and support"*.

Another topic that is related, which we have investigated to a great extent, is the use of emergency operating procedures (EOPs) in complex scenarios. In non-typical conditions there might be a mismatch between the procedures and the plant situation, as described in the former section. The scenario might be so complex that the procedures do not fit, even though the general situation is within design basis. The situation might be characterized by a lack of detailed guidance of the procedures, either since it is difficult to understand which procedure is the correct one to be in, or how to get to the right procedure from the one currently being executed, even though the crew knows that they should be in another procedure based on the general situation.

Massaiu & Holmgren describe this in detail, in the report entitled "*Diagnosis and Decision-Making with Emergency Operating Procedures in Non-Typical Conditions: A HAMMLAB Study with U.S. Operators"* [11]. Aspects of the procedures that did not match the situation were typically caused by: Multiple malfunctions were included in the scenarios; Key indicators referred in the EOPs were unreliable; The situation was not fully covered by the relevant EOP; There were ambiguous guidance and possible conflict between documents. In addition, the mismatches occurred late into the events, making the situation even more complex, so the operating crew could not follow the procedure step-by-step. [11] describes a number of recurring themes in such scenarios, including high crew-to-crew performance variability, even in within-design-basis accidents covered by EOPs. In general the crews had difficulties when the EOPs lacked detailed guidance and required interpretation. This was the more so in non-typical (i.e., textbook) conditions. Degraded indications (instrument failures, overlapping malfunctions, and miscommunications) are extremely challenging to the crews and can seriously affect plant safety. In the scenarios studied, inappropriate handlings occurred in all scenarios, by all the crews. The consequences observed in the simulations due to the inappropriate handlings included: Vessel integrity challenges (e.g., relevant EOP not entered, 1 hour "soak" not respected); Radiation releases outside the plant (e.g., avoidable releases through SG PORVs); Induced equipment failures (e.g., late stop and likely RCPs damage); Inventory depletion (e.g., lack of efforts to identify and isolate RCS leak outside containment).

Massaiu & Holmgren [11] describe challenges and operator strategies, including whether the crews understood the situation; whether they understood the procedure; whether they recognized the procedure – situation mismatch; and how the crews resolve the procedure - situation mismatch.
By these kinds of studies we learned many lessons, useful for many purposes, but especially for training, HRA and for crew organization. For training, Massaiu & Holmgren [11] mention the nature of scenarios to include in training is one of the key learning points. Strict procedure followers generally underperformed in the studied scenarios, meaning that teamwork and decision making strategies must be trained to cope with non-typical scenarios/plant conditions. For HRA, Massaiu [ibid], based on the above lessons, states that the following points should be noted: Do not over-emphasize procedure following in these kinds of scenarios, it is not enough to analyze errors of omission (deviations from the expected procedural progression). After the first hour into an emergency the procedure-situation fit is likely to decrease while fatigue effects arise, and this combination increases the likelihood for operators' autonomous decisions and errors. The second lesson for HRA, is always to include analysis of cognitive aspects, since interpretations may be required also to apparently straightforward steps. Extreme scenarios require lots of cognitive work from the crew, such as: analyzing procedures to identify possible procedure-situation mismatches, since degraded indications will most probably result in mismatches. A deeper understanding of the nature of the

difficulties for the crews is required, hence a thorough scenario analysis is needed for HRA (tasks/procedures). Also, crew teamwork and collective decision making aspects must be analyzed in a complete HRA, not only individual factors. On crew organization, Massaiu emphasizes that team factors and crew cognition are critical for performance in difficult scenarios, e.g., the role of the supervisor, distributed leadership, team orientation, backup and support. It was seen in the scenarios that the quality of teamwork decreases with complexity and fatigue, including less structured meetings and poor quality of briefings and discussions, as well as communication errors. On the role of STA and independence of STA, we saw a tendency of STAs to work mainly as "procedure following double-checker". This was not the intended role of the STA in the first place, they were introduced in the control room crew set-up to perform independent checking of the overall situation and to provide technical expertise for accident mitigation. Lastly, the ability of obtaining local information (e.g., local radiation measurements) in addition to indications available in the control room was found important.

These investigations have thrown light on performance aspects not commonly investigated at training/requalification sessions, and have increased the general understanding for how crews operate in these situations. The scenarios that have been studied the most are SGTR (Steam Generator Tube Rupture) (incl multiple); LOFW (Total Loss of Feedwater) (and combined with SGTR); ISLOCA (Interfacing System Loss of Coolant Accident, LOCA outside containment); and the H.B. Robinson fire.

The results from these studies can be used to inform HRA practitioners. Such knowledge is also important for many other stakeholders: Regulatory authorities can use them for increasing their general knowledge about the ways in which crews are handling scenarios. This gives a good technical basis for guidelines on which things that should be investigated in reviews, and whether licensees include certain topics in their safety evaluations. One of the main reasons that the utilities are interested in participating in HAMMLAB studies, have been that they can get ideas for improving their training and crew organization as well as tools for crew support. Such input comes in the form of knowledge of new types of scenarios, experimental set-ups of new types of teamwork as well as support tools for various activities in the control room. This may especially be strengthened if observers from e.g., the training department join the data collection. Plants can also gain useful ideas for potential future control room solutions, digital interfaces and large screens, since HAMMLAB is a computerized control room and we continuously work with these issues. Many of the participating plants are planning digital upgrades of their control rooms in the years to come.

2.3 Data methods to support basic questions such as to what extent digital systems are more error prone than analog systems

In new power plants and in modernized control rooms, new digital technology is introduced. The question has been posed about whether digital systems are more error prone than analog systems. Is human performance with respect to safety impact similar in analog and computerized control rooms? One may think that it is not the analog/digital dimension, but rather other dimensions, like quality of the solutions, that should be more influential on performance. This question has an impact both for the design itself, how to implement best possible digital systems; and on the analysis, including the HRA, of the design. How can HRA methods and which HRA methods can be used to analyze the new digital systems?

To begin addressing these questions, the Halden Project developed a method to look into the identification/verification tasks that human-system interfaces (HSIs) are used for [12, 13]. Hildebrandt et al. explains the new method called micro-tasks [12, pg. 1]:

> *"One of the fundamental requirements for a human-machine interface (HMI) is to enable the user to perceive information quickly and accurately. Indeed, one of the advantages of digital interfaces is that they provide opportunity for new types of visualizations that can significantly improve information gathering compared to analogue displays. Two important measures of how well an HMI supports information perception are speed of identification and accuracy. We therefore decided to develop a method of interface evaluation that is optimized towards*

these two measures. The direct assessment method should generate high volumes of quantitative performance data in a relatively short amount of time."

The method studies decontextualized tasks, typically identification and verification tasks of operators [12, pg. 1]:

> "*The scope of this kind of method is limited. For instance, insights into the effect of HMI features on situation awareness or higher-level, context-dependent decision-making are limited. Nevertheless, low-level performance data are valuable when seen in the context of a larger evaluation effort and as a foundation for studying the higher-level cognitive processes."*

The method uses simple questions to operators, either in front of a display or simulator with a frozen state of the plant, or utilizing mini-scenarios. Eitrheim et al. [14] tested the method in the HAMMLAB simulator laboratory, and Hildebrandt & Fernandes [13] tested it in a training simulator at a nuclear power plant. In the latter they compared directly digital (tablet displays) and analog solutions (on boards). The questions relate to indications available on the panels/display in front of the participant, and are part of real control room tasks. The questions are easy to understand and quickly answerable individually (less than 20 seconds). The response might be single choice between options or numerical entry. The results so far indicate an intriguing possibility to compare various types of equipment and ways of presenting information, since one may delve into the results based on various patterns. In direct comparisons between concrete solutions, the method can be used to choose among design alternatives. Although the methodology is very similar to the picture-sentence verification paradigm in cognitive psychology, we still need to collect more data in order to validate the method for industrial use.

Massaiu & Fernandes [15] compared results from two micro-task studies with task types in HRA methods (e.g., THERP). They compared Check/reading type of tasks vs calculation tasks, both under the major task type of identification/verification. They found a big difference in error rates in simple checks and in calculation tasks, so the conclusion was

> "*The results show that the task type is a stronger determinant of operator error rate than the HSI"* [15, pg. 1].

Thus, this dimension should be important to focus on in design and analysis of HSIs. For example, comparisons and calculations can be better in (new) digital solutions, and designers might want to focus on those aspects that improves this and then test it with this method.

The preliminary result is that digital displays do not significantly increase the error rates. The nature of the task, that is the type of cognitive tasks (e.g., checking vs. calculating) is a stronger determinant of task accuracy. The question is not only whether the solution is analog or digital, but rather the way in which the solution is implemented, and on what type of cognitive task it is evaluated.

For HRA, one question is whether these kinds of data can contribute to updating HRA methods for digital systems. One may think of using this kind of data to update the basic HEPs for the task types for digital displays, and also to adapt the user guidance on how the HRA methods can be used for the new digital displays.

3. DISCUSSION, FUTURE USE OF DATA FOR HRA

For further supporting HRA with data, the first question is what kinds of data are needed. Is it causal data showing impacts of PSFs on performance? Is it qualitative data, showing details of operational issues and challenges for the operating crews? The studies referred in this paper showed that both qualitative and quantitative data are useful for HRA.

Can big data methods be used to support HRA? This really depends on the specific need, as well as what is meant by the term big data. For typical quantitative data, especially the kind of data where quite a large amount of data may be collected, intuitively big data methods seem to be of use. With the

type of data discussed in the micro-task method in section 2.3 above, one may foresee the use of various big data methods in order to explore quantitative data, looking into the data sets from various angles. E.g., looking into basic data exploring pump data vs flow data, exploring data for one type of cognitive task like identification vs calculation type of tasks, and so on. The micro-task method can collect large amounts of data in a short time and thus, if the data collection is standardized and the dimensions properly defined, one might even think of combining it with machine learning algorithms in order to automate parts of the data treatment. However, one should be aware of the limitations due to the decontextualization of the tasks studied.

The most useful insights from the studies described above are qualitative insights that still depend on qualitative methods and on understanding the problem at hand. There will still be a need for qualitative support in using the HRA methods. Intuitively, big data methods might not be useful for this kind of data. However, this depends on what we mean by the buzz word "big data". As described above, we have collected large amounts of crew stories and qualitative data about a number of basic and complex variants of central scenarios for nuclear power plants that were written by process control and human factors experts working together on the basis of on-line observations, interviews of participating crews and especially reviews of audio/video, interface interaction and simulator log data. One may easily imagine modern types of text recognition algorithms or other types of methods classified as machine learning to be used in order to explore this kind of raw data. This might even be in the form of automatic extraction of important lessons from a rich text-based data repository consisting of narratives in the form of crew stories on one side and the corresponding simulator log data (recording the plant status, the operators' interaction with the plant via the interface and controls, and the effect of the control actions on the plant) on the other side. One must be able to retain the contextual understanding of these data, since that is crucial in order to not misuse the data. Faulty causal explanations might be deduced if the proper methodological basis is not in place. E.g., too much use of correlation statistics may lead to faulty conclusions in such data, where the causal relationships are intricate.

We have collected both quantitative and qualitative data in Halden data repositories and foresee to investigate the use of modern data mining algorithms in order to exploit these data to the most possible extent.

Can Bayesian methods be used on high HEP cases, like in the empirical studies? As described in section 2.1, a few errors in challenging scenarios make up good evidence by utilizing Bayesian methods. It is important to identify these cases though, and one cannot expect these situations to turn up in normal training sessions, since the occurrence of these situations are so rare and usually not part of the standard training programs, where automatic data collection systems could be put in place. Thus, there will still be a need to study these situations in research studies. What about low HEP cases? Usually, these actions and events are considered "easy" for the crews, so that they normally would succeed performing them. Are they easy since they are well trained and occur regularly in training? In that case, we should be able to collect a large amount of data from training sessions and maybe conclude on HEPs based on these. However, how do we know that the identified events and actions are the same from time to time? How do we know that the situation under analysis is similar enough to another case to be able to use the same number? Here the way in which one generalize come into play. If one manages to identify task types and error mechanisms, one can generalize to other situations by applying these dimensions. This still means though, that an analysis in the form of a generalization mechanism, e.g., in the form of an HRA method, is needed.

SACADA [16] has an interesting approach in this respect and early on it might have been envisaged used directly providing numbers for HEPs in HRA. Their approach looking into cognitive dimensions by defining and studying "training objective elements" and error forcing contexts by looking at "situational factors" has divided the problem at hand in interesting and hopefully usable dimensions. However, these do not exactly and directly fit HFEs in specific scenarios in the PRA, so it seems that these data also need an adaption in the form of an HRA method or similar in order to be used for PRA.

4. CONCLUSIONS

The use of the empirical data was very useful in the HRA empirical studies. We could see whether the HRA teams identified the central performance challenges the operating crews would face in the scenarios. Also, we managed to evaluate whether they predicted the HEPs in the right ballpark and if not, why they missed. NUREG-2127 concluded that the empirical data was fit for the purpose [4, pg. 99]:

> *"The results of the study show that the predictive performance of the various HRA methods could be evaluated. Thus, reference data for the benchmarking of the HRA methods can be obtained from a simulator study based on a relatively small number of observations of each HFE."*

[Ibid., pg. 103] concludes:

> *"In general, the promising results from this study encourage and promote the use of simulator data in the future for HRA in many different ways."*

The International and U.S. HRA Empirical Studies are summed up in NUREG-2156, [6, pg. 9-20] regarding data:

> *"The studies have shown that simulator data are highly useful for HRA studies. Although simulator data was used as the empirical basis against HRA predictions, the promising results from this study encourage and promote the use of simulator data in the future, as well as encouraging analysts to use it in different ways. The studies also show the potential of using and aggregating empirical simulator results from multiple studies to strengthen the empirical basis for both method assessment and extending the scope of methods to address some of the identified shortcomings. In summary, while there are other sources of HRA data, this study reinforced the relevance of simulator data for HRA in general."*

The kind of analysis of data described in section 2.2 is of qualitative nature. The biggest benefit for HRA is increased knowledge, either for HRA practitioners, reviewers of HRA submittals so that they can ask the right questions, or for increasing the knowledge for HRA methods people in general. Qualitative data in the form of qualitative findings and insights have in the empirical studies shown to be most useful for evaluation of HRA methods. In addition to the value of the qualitative insights in the other studies discussed in this paper, this would point to the conclusion that it is not advisable to a large extent to use "big data" methods that only look for correlations between variables in the data set. The intricate relations between all the variables in the robust system that a control room environment constitutes should be understood and even though one may find correlations, these do not necessarily render causal relations. Thus, fast conclusions on these matters through correlations only will not be advisable.

Back to the question posed in the introduction of this paper: Can data replace HRA methods in the quantification of HFEs? Given the richness of relations between variables in the control room, and the large amount of spread of scenario developments in difficult scenarios, my conclusion so far is that HRA methods will still be a necessary step. HRA methods are needed in order to extrapolate from data to HEPs of HFEs in the specific scenarios under analysis.

The Halden Project will continue serving data to HRA, both in qualitative and quantitative forms.

Acknowledgements

Thanks to Salvatore Massaiu for inspiring discussions and review.

References

[1] Swain, A.D., & Guttman, H.E. (1983). "*Handbook of human reliability analysis with emphasis on nuclear power plant applications. Final report*." NUREG/CR-1278. Washington, DC: US Nuclear Regulatory Commission.

[2] Boring, R.L. (2012). "*Fifty Years of THERP and Human Reliability Analysis"*, Proceedings of the 11th International Conference on Probabilistic Safety Assessment and Management (PSAM11), June 25-29, 2012, Helsinki, Finland.

[3] A. Bye, K. Laumann, P.Ø. Braarud, S. Massaiu, (2006). "*Methodology For Improving HRA By Simulator Studies*", Proceedings of the 8th International Conference on Probabilistic Safety Assessment and Management (PSAM8), May 14-18, 2006, New Orleans, Louisiana, USA.

[4] J. Forester, V.N. Dang, A. Bye, E. Lois, S. Massaiu, H. Broberg, P.Ø. Braarud, R. Boring, I. Männistö, H. Liao, J. Julius, G. Parry, P. Nelson, (2014). *"The International HRA Empirical Study – Final Report – Lessons Learned from Comparing HRA Methods Predictions to HAMMLAB Simulator Data*", NUREG-2127, U.S. Nuclear Regulatory Commission, Washington D.C.

[5] J. Forester, V.N. Dang, A. Bye, E. Lois, S. Massaiu, H. Broberg, P.Ø. Braarud, R. Boring, I. Männistö, H. Liao, J. Julius, G. Parry, P. Nelson, (2013). *"The International HRA Empirical Study – Final Report – Lessons Learned from Comparing HRA Methods Predictions to HAMMLAB Simulator Data*", Halden Project Report no. 373 (HPR-373). Also published as NUREG-2127.

[6] J. Forester, H. Liao, V.N. Dang, A. Bye, E. Lois, M. Presley, J. Marble, R. Nowell, H. Broberg, M. Hildebrandt, B. Hallbert, T. Morgan (2016). *"The U.S. HRA Empirical Study – Assessment of HRA Method Predictions against Operating Crew Performance on a U.S. Nuclear Power Plant Simulator"*, NUREG-2156, U.S. Nuclear Regulatory Commission, Washington D.C.

[7] Lois, E., Dang, V.N., Forester, J., Broberg, H., Massaiu, S., Hildebrandt, M., Braarud, P.Ø., Parry, G., Julius, J., Boring, R., Männistö, I., Bye A. (2009). "*International HRA Empirical Study Phase 1 Report - Description of Overall Approach and Pilot Phase Results from Comparing HRA Methods to Simulator Performance Data.*" HWR-844; NUREG/IA-0216, Vol. 1, International Agreement Report, U.S. Nuclear Regulatory Commission, Washington D.C.

[8] A. Bye, E. Lois, V.N. Dang, G. Parry, J. Forester, S. Massaiu, R. Boring, P.Ø. Braarud, H. Broberg, J. Julius, I. Männistö, P. Nelson (2010). *"The International HRA Empirical Study – Phase 2 Report – Results From Comparing HRA Methods Predictions to HAMMLAB Simulator Data on SGTR Scenarios*", HWR-915; NUREG/IA-0216, Vol. 2, U.S. Nuclear Regulatory Commission, Washington D.C.

[9] V.N. Dang, J. Forester, R. Boring, H. Broberg, S. Massaiu, J. Julius, I. Männistö, H. Liao, P. Nelson, E. Lois, A. Bye, (2011). *"The International HRA Empirical Study – Phase 3 Report – Results from Comparing HRA Methods Predictions to HAMMLAB Simulator Data on LOFW Scenarios*", HWR-951; NUREG/IA-0216, Vol. 3, U.S. Nuclear Regulatory Commission, Washington D.C.

[10] P.Ø. Braarud, B. Johansson, (2010). "*Team Cognition in a Complex Accident Scenario*". Halden Work Report no. 955 (HWR-955), Halden, Norway.

[11] Massaiu, S. and Holmgren, L. (2014). "*Diagnosis and Decision-Making with Emergency Operating Procedures in Non-typical Conditions: A HAMMLAB Study with U.S. Operators*", Halden Work Report no. 1121 (HWR-1121), Halden, Norway.

[12] Hildebrandt, M., Eitrheim, M.H.R. and Fernandes, A. (2016). "*Pilot Test of a Micro-Task Method for Evaluating Control Room Interfaces*", Halden Work Report no. 1130 (HWR-1130), Halden, Norway.

[13] Hildebrandt, M. and Fernandes, A. (2016). "*Micro Task Evaluation of Innovative and Conventional Process Display Elements at a PWR Training Simulator*", Halden Work Report no. 1169 (HWR-1169), Halden, Norway.

[14] Eitrheim, M.H.R, Fernandes, A., Svengren, H. (2017). "*Evaluation of design features in the HAMBO operator displays*", Halden Work Report no. 1212 (HWR-1212), Halden, Norway.

[15] Massaiu, S. and Fernandes, A. (2017). "*Comparing Operator Reliability in Analog vs. Digital Human-System Interfaces: An Experimental Study on Identification Tasks*". PSAM Topical

Conference on Human Reliability, Quantitative Human Factors, and Risk Management, 7 - 9 June 2017, Munich, Germany.
[16] Chang, Y.J., Bley, D., Criscione, L., Kirwan, B. Mosleh, A., Madary, T., Nowell, R., Richards, R., Roth, E.M., Sieben, S., Zoulis, A. (2014), "*The SACADA Database for Human Reliability and Human Performance*", Reliability Engineering and System Safety, 125(2014) 117-133.