Title	Enhanced Bayesian THERP - Lessons learnt from HRA benchmarking
Author(s)	Holmberg, Jan-Erik; Bladh, Kent; Oxtrand, Johanna; Pyy, Pekka
Citation	Proceedings of PSAM 10 International Probabilistic Safety Assessment & Management Conference, 7-11 June 2010, Seattle, Washington, USA, pp. Paper 52
Date	2010
Rights	This article may be downloaded for personal use only

Enhanced Bayesian THERP — Lessons learnt from HRA benchmarking

Jan-Erik Holmberg[a*], Kent Bladh[b], Johanna Oxtrand[c], and Pekka Pyy[d]
[a]VTT, Espoo, Finland
[b]Vattenfall Power Consultant, Malmö, Sweden
[c]Ringhals AB, Väröbacka, Sweden
[d]Teollisuuden Voima Oy, Helsinki, Finland

Abstract: The Enhanced Bayesian THERP (Technique for Human Reliability Analysis) method uses as its basis the time-reliability curve introduced in the Swain's human reliability analysis (HRA) handbook. It differs from the Swain's Handbook via a transparent adjustment of the time-dependent human error probabilities by use of five performance shaping factors (PSFs): (1) support from procedures, (2) support from training, (3) feedback from process, (4) need for co-ordination and communication, (5) mental load, decision burden. In order to better know the characteristics of the Enhanced Bayesian THERP from a more international perspective, the method has been subject to evaluation within the framework of the international "HRA Methods Empirical Study Using Simulator Data". Without knowledge of the crews' performances, several HRA analysis teams from different countries, using different methods, performed predictive analyses of four scenarios. This paper gives an overview of the method with major findings from the benchmarking. The empirical comparison gives confidence that the time reliability curve is a feasible and cost effective method to estimate human error probabilities when the time window is well defined and relatively short. The comparison of empirical observations with predictions was found as a useful exercise to identify areas of improvements in the HRA method.

Keywords: Human reliability analysis, human performance, simulator study, nuclear power plant safety

1. INTRODUCTION

The modeling and quantification of human interactions is widely acknowledged as a challenging task of probabilistic safety assessment (PSA). Methods for human reliability analysis (HRA) are based on a systematic task analysis combined with a human error probability quantification method. The quantification typically relies on expert judgments, which have rarely been validated by statistical data.

In order to compare different HRA methods an international study "HRA Methods Empirical Study Using Simulator Data" has been carried out using actual simulator data as reference for the comparison [1]. The overall goal of the international HRA method evaluation study was to develop an empirically-based understanding of the performance, strengths, and weaknesses of the HRA methods. It is expected that the results of this work will provide the technical basis for the development of improved HRA guidance and, if necessary, improved HRA methods. Results from the empirical comparison are now available, while the drawing of conclusions and reporting are in progress.

This paper gives an overview of the enhanced Bayesian THERP method with major findings from the benchmarking study. Presentation of the overall outcomes of the international study is outside the scope of this paper.

* jan-erik.holmberg@vtt.fi

2. EMPIRICAL STUDY SET-UP

2.1. Overview

The empirical study was based on a set of simulator experiments run in the Halden Reactor Project's HAMMLAB (HAlden huMan-Machine LABoratory) simulator facility. Fourteen operating crews from an operating nuclear power plant (a pressurized water reactor) participated in a series of performance shaping factor/masking experiments. Without knowledge of the crews' performances, several HRA analysis teams from different countries, using different methods, performed predictive analyses of the scenarios.

Two scenarios were experimented both with a basic variant and a complex variant. The first scenario was a steam generator tube rupture (SGTR) and the second one a loss of feedwater (LOFW). Both scenarios are considered in a PSA for a corresponding nuclear power plant.

In the basic variant of scenarios, no additional difficulties were implemented in the scenario, meaning that they represented a familiar and routinely practiced case for the crews. The complex variants included additional failure events and misleading indications of conditions at the plant, making it considerably more difficult for the operators to diagnose the situation. In fact, the initial purpose of the simulator studies was to study the effect of masking to the crew performance, and the same simulator runs were then used to HRA methods comparison purpose as well.

2.2. Scenario descriptions

2.2.1. Steam generator tube rupture

In the basic case, the event sequence starts with a steam generator tube rupture after which reactor trip will be actuated manually or automatically. In general, what is expected is that the crew will perform four primary tasks corresponding to the human failure events (HFE) defined for the base SGTR scenario. These tasks include:

#1 Identifying which SG is ruptured and isolating it.
#2 Cooling down the reactor coolant system (RCS) expeditiously by dumping steam.
#3 Depressurizing the RCS expeditiously using the pressurizer sprays but also likely by using a pressurizer relief valve (PORV) to expedite the depressurization.
#4 Stopping safety injection (SI) upon indication that the SI termination criteria are met.

The complex scenario differs from the basic scenario in several important features. The event starts off with a major steamline break downstream of the main steam isolation valves (MSIVs) with a nearly coincident steam generator tube rupture that occurs when the MSIVs close. This will cause an immediate automatic scram. The secondary radiation indication fails independently to show increased radiation level. A busbar failure later in the scenario makes the pressurizer sprays ineffective in performing the RCS depressurization so that the crew is forced to use a pressurizer PORV to perform the depressurization. The PORV fails to close completely when the operators reclose the PORV upon achieving the desired RCS depressurization. Half of the crews were given an erroneous closed indication for the PORV and half of the crews were given a correct open indication for the PORV. All the PORV block valves failed open by virtue of the simulation design but unknown by the crew so that none of the PORV block valves will actually close. Tasks #1, #2 and #3 are same as in the base scenario, task #4 is not relevant to the complex scenario, and an additional task #5 was defined related to the identification of the failed closure of the PORV.

2.2.2. Loss of feedwater scenario

All normal feed water to the steam generators is lost in the beginning of the scenario. The steam generator levels drop rapidly and the reactor is tripped automatically, if the crew doesn't trip it manually before. The auxiliary feed water (AFW) pumps fail to start due to various technical failures. After the reactor is tripped, the crew should detect that they don’t have any feedwater to the SGs. Two main actions of the emergency operating procedure are:

1) Attempt restoration of feed flow to steam generators (decay heat removal from the secondary side, from a variety of systems
2) Initiation of RCS bleed and feed heat removal (decay heat removal from the primary side, after the secondary heat sink is failed)

The difference of the complex scenario is that the failure causes for loss of main and auxiliary feedwater systems are more difficult to identify, and the operators may e.g. spend more time in trying to use the condensate pumps to restore feedwater. In addition the wide range (WR) level measurements are failing in two of the three SGs (there is only one WR level measurement in each SG) making it more difficult to detect the criteria to start bleed and feed.

The task of the HRA teams was to characterize the human failure event and the human error probability associated with the event "Feed and Bleed before SG dryout" (denoted as X4). Additionally, the task was to evaluate the probability of the human failure event associated with the "Late Recovery Before Core Damage" (denoted as X4L), which is the same operative action (Feed and Bleed) but implemented after SG dryout, when a rise in RCS pressure causes operators to re-examine decay heat removal.

2.3. HRA assessment phases

In order to facilitate the human performance predictions, the organizers of the experiment prepared an extensive information package for the HRA analysis teams including descriptions of the scenarios, description of the simulator and its man-machine interface, differences between the simulator and the home plant of the crews, procedures used in the simulator, characterization of the crews, their work practices and training. The task of the HRA analysis teams was to predict failure probabilities of operator actions defined, e.g., isolation of the ruptured steam generator, and to qualitatively assess which performance shaping factors (PSF) affect positively or negatively to success or failure of the crew.

The scenarios were evaluated in three phases. In the pilot phase, the SG isolation action of the SGTR scenario was analysed by HRA teams and compared against empirical results. In the second step, the rest of the actions of the SGTR scenario were under consideration. Finally, the LOFW scenario was under consideration. Generally, all three phases of the study were carried out quite similarly, but the HRA teams were allowed to change their approaches between the phases. The HRA team using the enhanced Bayesian THERP performed their assessment similarly in all phases.

On the empirical side, time was mainly used as the criterion for defining success/failure of the crew performance. The empirical identification of PSFs was based on a detailed analysis of simulator performances. Analysts viewed the video und transcribed key communications and events, and used also additional data sources, such as crew interview, crew PSF questionnaire, and observer comments. Finally, the analyst summarized the observed episode in the form of an operational story, highlighting performance characteristics, drivers, and key problems.

A specific method was used to empirically rate the PSFs. In the pilot phase, seven-item scale was used [1]: "very good", "good, "somewhat good", "nominal/average", "somewhat poor", "poor", "very poor". For certain PSFs, the scale was one-sided (can only have nominal or negative effect): "nominal", "somewhat high", "high", "very high". In the second and third phases, the empirical rating was done in two ways [2]. Firstly, an observational group rating was done using a 9 points scale: -2, -1,5, -1, ..., +1,5, +2. Secondly, the observational ratings were mapped on the so called HRA ratings: MND = Main negative driver, ND = Negative driver, 0 = Not a driver, N/P = Nominal/Positive (contributes to the overall assessment of the HEP being small).

In both pilot phase and later, the rating was made first for each crew performance and the crew specific rates were aggregated to general rates. Obviously, there existed variability between crews, and different drivers were present for different crews. Crew specific rates were therefore first grouped into well performing and less well performing crews, and group specific rating was performed for both groups. Finally, PSF aggregations for well performing crews were contrasted with aggregations for less well performing crews, in order to produce the overall observed PSF rating for each HFE [2].

3. ENHANCED BAYESIAN THERP

The Enhanced Bayesian THERP (Technique for Human Reliability Analysis) method is based on the use of the time-reliability curve introduced in the Swain's human reliability analysis (HRA) handbook [3] and on the adjustment of the time-dependent human error probabilities with performance shaping factors (PSFs) [4]. The method is divided into a qualitative and quantitative analysis part.

3.1. Qualitative analysis

The qualitative analysis consists of a modelling of the scenario with a block diagram and a description of the basic information of each operator action. The purpose of the block diagram is to define the operator actions in relation to relevant process events. The block diagram representation is close to a PSA event tree but is usually a somewhat more detailed model than an event tree. The purpose of the description of the basic information of each operator action is to consistently characterize main aspects of an operator action, e.g., initiating event, scenario, time windows, support from procedures and MMI, practical maneuvers needed in the action and other noteworthy information.

The block diagram is also used to show the assumed dependencies between operator actions belonging to the same scenario. The blocks used in the diagram should have exact correspondence with functional events (in event trees or system fault trees) of the PSA-model. This is important in cases where operator action basic events are modeled in system fault trees so that the link to event tree branches is not obvious. In this study, the operator actions were given, so that the construction of the block diagrams did not serve as defining the operator action basic events.

Figure 1 describes the block diagram of the basic LOFW scenario. Each block is an event, either a process event or an operator action. Arrow to the right corresponds the success path and downwards arrow the failure path. Orange blocks are the operator action events for which error probabilities were to be estimated. Light yellow blocks are operator actions out of the scope of the study. White blocks are process events and gray blocks are end states of the sequences. Timelines and emergency operator procedure (EOP) references shown in the top of the diagram were given to the HRA teams by the organizers of the study.

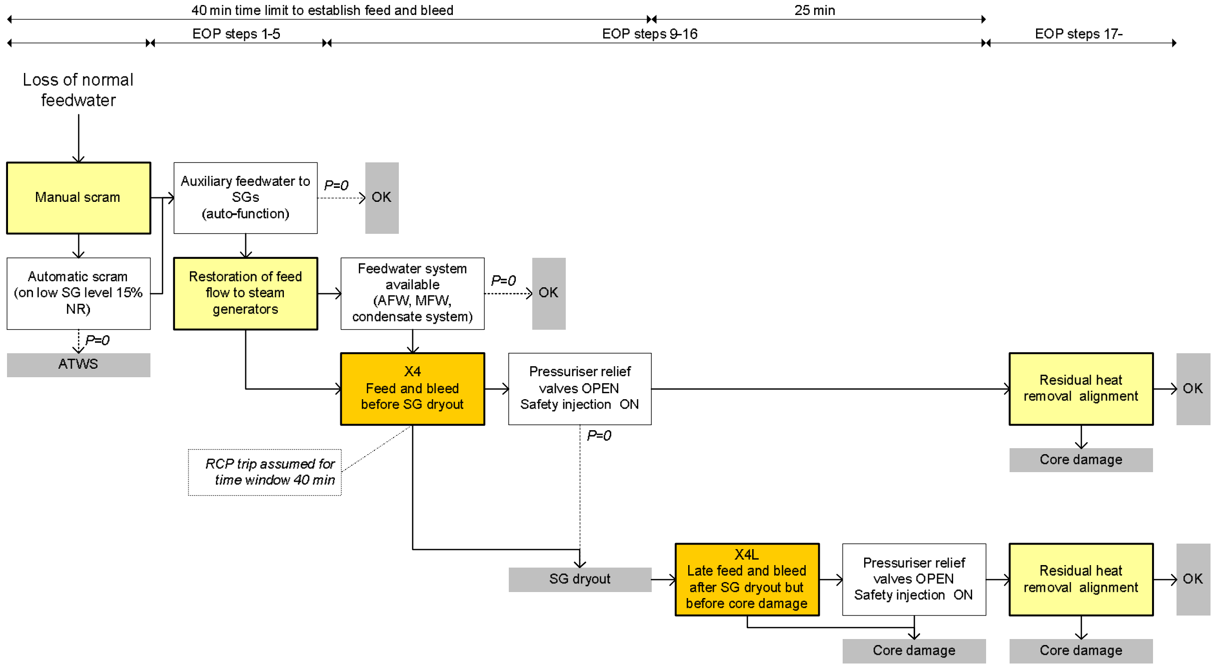

Figure 1: Block diagram of the LOFW scenario.

3.2. Quantitative analysis

The human error probability is derived using the time-dependent human error probability model as follows,

$$p(t) = \min\left\{1,\ p_0(t)\prod_{i=1}^{5} K_i\right\}, \qquad (1)$$

where $p_0(t)$ is the basic human error probability taken from [3], see Figure 2, t is the time available for identification and decision making, and K_1, …, K_5 are the performance shaping factors. The min-function ensures that the final probability stays within the range 0 to 1.

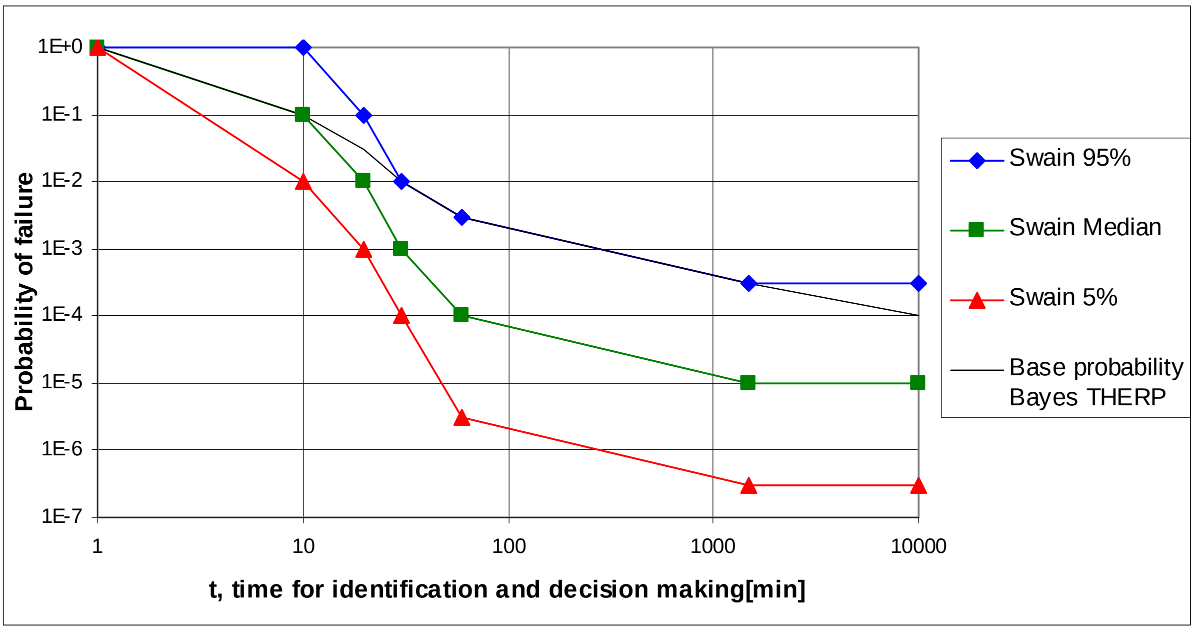

Figure 2: Time-dependent human error probability curve [4].

The time available for identification and decision making is shorter than the total time available for the operator action t_{tot} which is assumed to be composed of three parts as follows

$$t_{tot} = t_{ind} + t + t_{act}, \qquad (2)$$

where t_{ind} is time for first indication, t time for identification and decision making and t_{act} time for action. The following performance shaping factors are used [4]:

K_1: Quality and relevance of procedures
K_2: Quality and relevance of training
K_3: Quality and relevance of feedback from process (MMI)
K_4: Mental load (stress) in the situation
K_5: Need for coordination and communication.

Each performance shaping factor can receive a value 1/5, 1/2, 1, 2 or 5. Value 1/5 or 1/2 means that the action has a complicating character compared to a "nominal" situation. Consequently, value 2 or 5 means that the action is easier than the nominal case. Level "1" means that the factor plays no major role or that this factor is in a nominal level. The meaning of each value for each PSF is explained qualitatively in the method, see Table 1.

Table 1: Explanation of the scales for the performance shaping factors [4].

	Performance shaping factors				
	Quality and importance of procedures	**Quality and importance of training**	**Feedback from process, quality of MMI**	**Mental load**	**Communication and coordination**
Scale	*Are there procedures? Are they needed? Do they give support?*	*Has the situation been trained? What kind of training? How often it has been trained? Is the action well known?*	*Is critical information available for the operators? How easily, understandably, rapidly? Are there redundant information? Can there be misleading signals?*	*Is the situation or action unusual? Are there any special uncertainties in the situation?*	*Is it a scattered disturbance? Are actions needed inside/outside of control room (CR)? Does communication work? Is coordination of actions needed?*
5	No instructions or misleading instructions, instructions would be needed	No training or misleading training, training would be needed	No feedback from process or misleading information or too late feedback	Mental load is so high that it nearly hinders to make a rational decision, situation is chaotic, an extreme decision needs to made	Information must be obtained from inside and outside of CR, coordination of many activities, poor conditions for communication
2	Instructions are important but they are imperfect	Some training has been given but it is not fully applicable for the situation	Feedback is obtained but there are defects in the presentation of the critical information	Mental load is considerable, situation is serious, a serious decision needs to be made	Information must be obtained from inside and outside of CR, coordination of many activities, good conditions for communication
1	Instructions play no major role in the situation	Training plays no major role in the situation	Feedback plays no major role in the situation	Mental load plays no major role in the situation	Communication and coordination play no major role in the situation
1/2	Good instructions, applicable for the situation and they support well the selection of correct actions	Situation has been trained (in a simulator), an important theme in the training	Symptoms can be easily observed and identified	NA	Operator(s) can act directly based on available information without further communication
1/5	Very good instructions, operators should not make any mistake	Situation is often trained in a simulator, a very important theme	It is practically impossible to miss the symptoms, several redundant indications	NA	NA

The performance shaping factors will be given independently by a number of experts, and these judgments are consolidated with a Bayesian approach. In this approach, the performance shaping factors are assumed to be random variables following a multinomial probability distribution,

$$P(K_i = j \mid q_j) =, q_j \;, j = 1/5, 1/2, 1, 2, 5, \quad q_{1/5} + \ldots + q_5 = 1. \tag{3}$$

The prior distribution for the parameters of the multinomial distribution is assumed to be a Dirichlet distribution. The convenient feature of Dirichlet distribution is that if we assume the expert judgments as independent observations from a multinomial distribution, the posterior distribution is also Dirichlet and can be easily derived. The prior distribution is chosen by maximizing the entropy function. This distribution has an interpretation to represent maximal uncertainty. The mathematical procedure is presented in [5].

Four experts have participated in this exercise, and made their assessments independently of each other based on material obtained from Halden and processed by VTT. It should be noted that experts normally include also members from the operation crews at the actual plant, which was not possible during this experiment.

4. COMPARISON OF EMPIRICAL AND PREDICTED RESULTS

4.1. Steam generator tube rupture

Comparison of the empirical failure rates [1, 2] to the predicted mean values (i.e. posterior mean values) are shown in Table 1. Time available for identification and decision making was taken from the information package submitted by Halden. The prior human error probability is from the Swain's curve, see Figure 2. Four experts assessed independently the performance shaping factors and the assessments were aggregated using the Bayesian procedure. Figure 3 gives a graphical illustration. Error ranges (0,05–0,95 fractiles) are shown for the predicted human error probabilities.

Table 2: Comparison of empirical failure rates to predicted mean values in the SGTR scenario.

Human failure events		**Empirical (k failure, n trials)**				**Bayes THERP**	
ID	Description	k	n-k	p = (k+0,5)/(n+1)	Rank	p	Rank
HFE-1A	SG isolation in base scenario	1	13	0,1	6	0,03	9
HFE 2A	Cooldown in base scenario	1	13	0,1	6	0,08	5
HFE-3A	Depressurization in base scenario	0	14	0,03	4	0,05	7
HFE-4A	Stop SI to stop primary to secondary leakage (base scenario)	0	14	0,03	9	0,04	8
HFE-1B	SG isolation in complex scenario	7	7	0,5	2	0,17	2
HFE-2B	Cooldown in complex scenario	0	14	0,03	6	0,08	4
HFE-3B	Depressurization in complex scenario	1	13	0,1	3	0,12	3
HFE-5B1	"PORV indicating closed" (complex version)	7	0	0,94	1	0,52	1
HFE-5B2	"PORV indicating open" (base version)	0	7	0,06	8	0,06	6

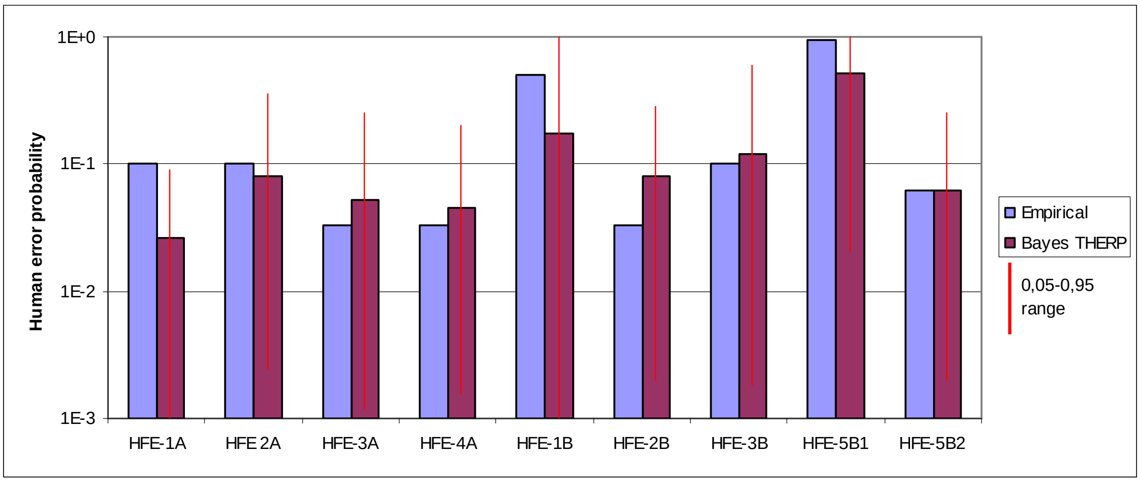

Figure 3: Comparison of empirical failure rates to predicted mean values in SGTR scenario.

Tables 3 and 4 show the empirical [1, 2] vs. predicted performance shaping factors. The scales used in the empirical rating are explained in chapter 2.3. Positive factors or drivers are indicated by a pale blue background color, while negative factors or drivers have light yellow background color. Strong factors or drivers have a bolded font. Nominal factors or drivers have white background color.

Table 3: Comparison of empirical PSFs to predicted PSFs in the SGTR base scenario.

Empirical judgements (observational/HRA)							
PSF	**1A**	**2A**		**3A**		**4A**	
Adequacy of time	good						
Time pressure		0	0	0	0	0	0
Stress		0	0	-0,5	ND	0	0
Scenario complexity		0 (-1)	**MND**	0	N/P	0	N/P
Indication of conditions	**very good**	0	N/P	0	N/P	0	N/P
Execution complexity	somewhat high	0 (-1)	ND	-1	ND	**2**	N/P
Training	**good to very good**	1	N/P	1	N/P	1	N/P
Experience	**good to very good**	0	0	0	0	0	0
Procedural guidance	good	-1	ND	1	N/P	**2**	N/P
HMI	**very good**	0	N/P	0	N/P	0	N/P
Work processes		1	N/P	0	0	0,5	N/P
Communication		1	N/P	0,5	N/P	0,5	N/P
Team dynamics		1 (-1)	ND	1 (-2)	ND	0,5	N/P
Bayes THERP							
PSF	**1A**	**2A**		**3A**		**4A**	
Stress	1,8	1,3		1,3		1,0	
Training	0,5	0,5		0,5		0,6	
Coordination, communication	1,4	1,1		1,1		1,1	
Procedural guidance	0,6	**0,4**		**0,4**		**0,4**	
HMI	**0,4**	**0,4**		0,5		0,5	

Table 4: Comparison of empirical PSFs to predicted PSFs in the SGTR complex scenario.

Empirical judgements (observational/HRA)									
PSF	**1B**	**2B**		**3B**		**5B1**		**5B2**	
Adequacy of time	somewhat poor								
Time pressure		0	0	0	0	0	0	0	0
Stress		0 (-0,5)	ND	0 (-1)	ND	0	0	0	0
Scenario complexity	high	0 (-1)	ND	0 (-1,5)	ND	**-2**	**MND**	**2**	N/P
Indication of conditions	somewhat poor to poor	0	N/P	0	N/P	**-2**	**MND**	**2**	N/P
Execution complexity	somewhat high	0 (-1)	ND	-1	ND	0	N/P	**2**	N/P
Training	somewhat poor	1	N/P	1	N/P	0	0	**2**	N/P
Experience		0	0	0	0	0	0	0	0
Procedural guidance	poor	1	N/P	1	N/P	0	0	1	N/P
HMI		0	N/P	0	N/P	0	0	0	N/P
Work processes	high	1 (-0,5)	N/P	0	0	-1	ND	0	N/P
Communication		1 (-0.5)	N/P	1	N/P	0	N/P	0	N/P
Team dynamics		0,5 (-2)	**MND**	1 (-1,5)	ND	0	N/P	0	N/P
Bayes THERP									
PSF	**1B**	**2B**		**3B**		**5B1**		**5B2**	
Stress	**2,7**	1,5		1,5		**2,7**		1,3	
Training	1,4	0,6		0,8		1,1		0,6	
Coordination, communication	1,2	0,8		0,8		0,8		0,8	
Procedural guidance	0,7	**0,3**		0,5		0,6		**0,4**	
HMI	**2,4**	0,5		0,6		**3,0**		0,6	

4.2. Loss of feedwater

Comparison of the empirical failure rates [6] to the predicted mean values (i.e. posterior mean values) are shown in Table 5. Three experts assessed independently the performance shaping factors and the assessments were aggregated using the Bayesian procedure. Figure 4 illustrates same graphically.

Table 5: Comparison of empirical failure rates to predicted mean values in LOFW scenario.

Human failure events		**Empirical (k failure, n trials)**				**Bayes THERP**	
ID	Description	k	n-k	p = (k+0,5)/(n+1)	Rank	p	Rank
X4-A	Feed and bleed in base scenario	0	10	0,045	3	0,01	4
X4L-A	Late recovery of feed and bleed in base scenario					0,07	2
X4-B	Feed and bleed in complex scenario	7	3	0,42	1	0,06	3
X4L-B	Late recovery of feed and bleed in complex scenario	0	7	0,06	2	0,3	1

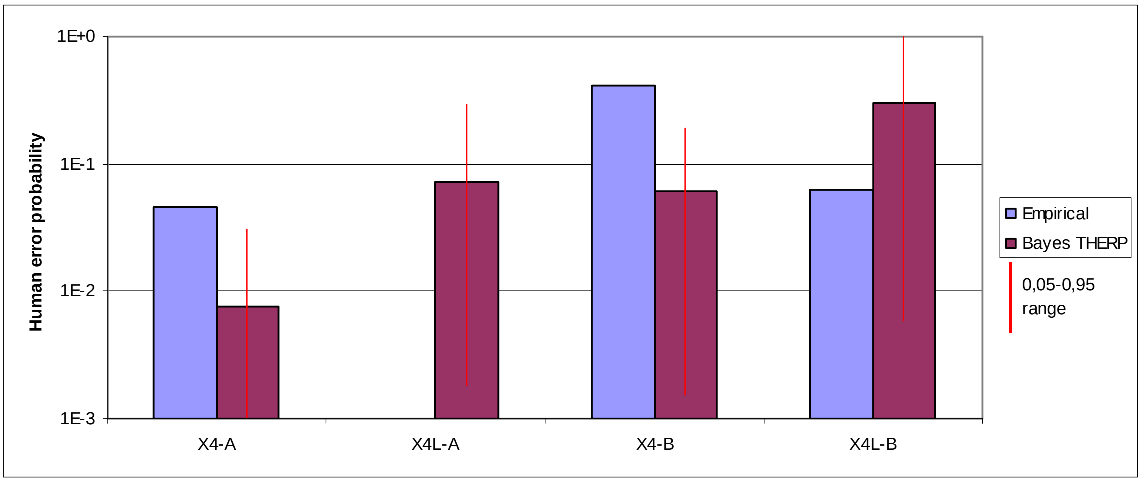

Figure 4: Comparison of empirical failure rates to predicted mean values in the LOFW scenario.

Table 6 shows the empirical [6] vs. predicted performance shaping factors.

Table 6: Comparison of empirical PSFs to predicted PSFs in the LOFW scenario.

Empirical judgements (observational/HRA)								
PSF	**X4-A**		**X4L-A**		**X4-B**		**X4L-B**	
Adequacy of time	1	N/P				N/P		N/P
Time pressure	0	0			0	0	0	0
Stress	0	0			0	0	0	0
Scenario complexity	0	N/P			**-2**	**MND**	**-2**	**MND**
Indication of conditions	0	N/P			**-2**	**MND**	-1	ND
Execution complexity	0	N/P			0	N/P	0	N/P
Training	1	N/P			1	N/P	1	N/P
Experience	0	0			0	0	0	0
Procedural guidance	1	N/P			-1	ND	-1	ND
HMI	0	N/P			0	N/P	0	N/P
Work processes	1 (-1)	ND			-1	ND	-1	ND
Communication	0	0			0	0	0	0
Team dynamics	0,5	N/P			-1,5	ND	-1	ND

Bayes THERP				
PSF	**X4-A**	**X4L-A**	**X4-B**	**X4L-B**
Stress	1,0	2,0	1,3	2,0
Training	1,5	2,0	2,0	2,0
Coordination, communication	1,7	2,0	1,7	2,0
Procedural guidance	0,5	1,0	0,7	1,2
HMI	0,7	0,7	**4,0**	**4,0**

5. EVALUATION FINDINGS

5.1 Human error probability predictions

In the SGTR scenario, the predictions and the empirical outcomes are well in balance (table 2, figure 3). A possible reason for this result is that the time dependent human error probability model seems to work well when the time window for the operator action is relatively short (< 1 hour). In the SGTR scenario, there are two difficult actions while others are easy or at least relatively easy for well trained operators. From the quantification point of view, the main issue is to properly distinguish between difficult and easy tasks. The scale used in the enhanced Bayesian THERP method is capable for that.

In the LOFW scenario, there is a clear mismatch between predictions and the empirical outcomes in the complex scenario (X4-B and X4L-B). Regarding the feed-and-bleed action before SGTR dryout (X4-B), the complexity of the scenario was underestimated. Regarding the feed-and-bleed action after SGTR dryout (X4L-B), importance of the new indication of conditions as positive driver was missed. When the SGs are empty, the failures in the SG level measurements were easy to detect for the crew, as the WR level trend showed flat lines at different levels.

The LOFW complex scenario is a good example demonstrating how human error probability estimates are sensitive to judgements. In this study, misjudgements were made compared to empirical outcomes. Using knowledge from the simulator runs, different values would have been used for certain PSFs.

Another lesson is that error probabilities in complex tasks can be tricky to estimate. Even if the time window is relatively long, like in X4-B, the complexity of the situation made it difficult for the operators to succeed. On the other hand, one crucial piece of information, like in X4L-B, can turn out to be a success driver in the situation.

5.2. Performance shaping factor predictions

Comparison of empirically rated and predicted performance shaping factors cannot be done straightforwardly, e.g., due to different sets of PSFs. In the empirical rating, a large number of PSFs were used, in order to cover all PSFs used by different methods. In the Bayesian THERP, the number of PSFs is small in order to avoid having overlapping PSFs. A mapping between PSFs used in the empirical rates and in the HRA assessment is shown Table 7. It should be also noted that PSFs are not really predictable entities from HRA point of view, but rather means to predict outcomes of operators' performance.

Table 7: Mapping between PSFs used in the empirical and in the HRA assessment.

PSF used in the empirical rating	PSF used in the Bayesian THERP
Adequacy of time	NA, parameter of the time reliability curve
Time pressure	Stress
Stress	Stress
Scenario complexity	Coordination, communication
Indication of conditions	HMI
Execution complexity	Coordination, communication
Training	Training
Experience	Training
Procedural guidance	Procedural guidance
HMI	HMI
Work processes	—
Communication	Coordination, communication
Team dynamics	Coordination, communication

In the SGTR and LOFW base scenarios, the conditions for success are good and therefore no major negative factor or driver should be present. Some small differences can be noticed between the empirical and predicted assessments, but the differences are not radical or systematic. In the SGTR and LOFW complex scenarios, presence of negative factors or drivers is obvious. The question is just which of the factors are assumed to be more negative than others. In the SGTR action 1B, the support from procedures was assessed differently (empirically negative, HRA prediction positive). This can be explained by lack of plant-specific knowledge of the HRA team to evaluate the quality of the procedure. This mismatch is rather an expert opinion issue than as an HRA method issue.

Based on experience from the empirical study, potential for method improvements are seen in the definitions for PSFs. "Stress" is a vague PSF, and should be perhaps replaced by a more unambiguous PSF. A significant empirical observation was the variability between crews with regard to affecting performance shaping factors which means that PSFs are not only action dependent but also crew dependent. This variability is not explicitly accounted in the enhanced Bayesian THERP method, even though the method produces a probability distribution for each PSF. These probability distributions, however, reflect variability of expert judgements not the variability of crews.

Also the expert judgement method could be improved. Critical point is the interpretation of the performance shaping factors and their numerical rating. It is obvious that different experts will always interpret differently the explanations given for the scaling. As long as an expert is consistent in his/her judgments, values given for different operator actions can be compared. From the absolute level point of view, some calibration may be needed. So far, results from the empirical study did not clarify the actual need for calibration.

Experts should be urged to justify the rates. This is an essential way to collect insights, e.g., for improvements of the human factors. One finding was that the method could be complemented with a discussion phase after the expert judgements where experts could jointly comment the results and draw conclusions from the assessments. This would facilitate the interpretation of the results which is now based on pure interpretation of the numbers.

6. CONCLUSIONS

The empirical comparison gives confidence that the time reliability curve is a feasible and cost effective method to estimate human error probabilities when the time window is well defined and relatively short. Since the enhanced Bayesian THERP method includes performance shaping factors to modify the probabilities upwards or downwards, there are also reasonable possibilities to differentiate between more and less difficult operator actions. The enhanced Bayesian THERP method is not intended to predict PSFs. It however gives framework to discuss them and thus it can support to identify weaknesses and to find improvements in the operators' working environment.

The comparison of empirical observations with predictions was found as a useful exercise to identify areas of improvements in the HRA method. An important fact missed by the method is the variability between the crews with regard to importance of different PSFs. Also explanations for numerical scales for PSFs could be improved to harmonize the way experts interpret the scales. In this way, empirical tests are necessary to validate an HRA method.

Acknowledgements

The study has been made in the context of the project "International HRA Empirical Study", also called "Empirical Study of HRA Methods." The authors gratefully acknowledge the background material provided by the international project team and the OECD Halden Reactor Project for the work. The comments received from the project team have been valuable. This work is a subproject of the Challenges in Risk-Informed Safety Management (CHARISMA) project, which belongs to the SAFIR2010 Research Programme http://virtual.vtt.fi/virtual/safir2010/. VTT's work has been financed jointly by VYR, OECD Halden Reactor Project (MTO in-kind) and VTT.

References

[1] E. Lois, 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, International HRA Empirical Study – Phase 1 Report. Description of Overall Approach and Pilot Phase Results from Comparing HRA Methods to Simulator Performance Data, NUREG/IA-0216, Vol. 1, U.S. Nuclear Regulatory Commission, 2009, Washington, DC.

[2] E. Lois, V.N. Dang, G. Parry, J. Forester, S. Massaiu, R. Boring, P.Ø. Braarud, H. Broberg, J. Julius, I. Männistö and A. Bye, The international HRA empirical study — results from comparing HRA methods predictions to HAMMLAB simulator data on SGTR scenarios, HWR-915, OECD Halden Reactor Project, August 2009, Halden, draft (limited distribution).

[3] A.D. Swain and H.E. Guttmann. Handbook of Human Reliability Analysis with Emphasis on Nuclear Power Plant Applications. NUREG/CR-1278, Sandia National Laboratories, 1983, Albuquerque.

[4] P. Pyy, and R. Himanen. A Praxis Oriented Approach for Plant Specific Human Reliability Analysis — Finnish Experience from Olkiluoto NPP. In: Cacciabue, P.C., and Papazoglou, I.A. (eds.), Proc. of the Probabilistic Safety Assessment and Management '96 ESREL'96 — PSAMIII Conference, Crete. Springer Verlag, 1996, London, pp. 882–887.

[5] J. Holmberg and P. Pyy. An expert judgement based method for human reliability analysis of Forsmark 1 and 2 probabilistic safety assessment. In: Kondo, S. & Furuta, K. (eds.), Proc. of the 5th International Conference on Probabilistic Safety Assessment and Management (PSAM 5), Osaka, JP. Vol. 2/4. Universal Academy Press, 2000 Tokyo, pp. 797–802.

[6] Halden data for the International HRA empirical project: Aggregated operational descriptions, PSFs ratings and HFEs difficulty ranking for the LOFW-scenarios, Draft 4, OECD Halden Reactor Project, 27 October 2009, Halden (limited distribution).