Using Expert Judgments to Estimate the Multipliers of Performance Shaping Factors in Digital Main Control Rooms of Nuclear Power Plants

Peng Liu[a*], Yongping Qiu[b], Juntao Hu[b], Jiejuan Tong[c], Zhizhong Li[d]
[a]Department of Industrial Engineering, Tianjin University, Tianjin, China
[b]Shanghai Nuclear Engineering Research and Design Institute, Shanghai, China
[c]Institute of Nuclear and New Energy Technology, Tsinghua University, Beijing, China
[d]Department of Industrial Engineering, Tsinghua University, Beijing, China

Abstract: The lack of human performance data in human reliability analysis (HRA) has been lamented in the literature for a long time. It can increase the variability, uncertainty, unreliability of HRA outcomes to probabilistic risk analysis (PRA). This situation is even worse for PRA/HRA in digital main control rooms (MCRs) of nuclear power plants (NPPs). Expert judgment is used when data is unavailable. This study applied two expert judgment techniques, absolute probability judgment (APJ) and ratio magnitude estimation (RME), to generate licensed operators' judgments on the multipliers of performance shaping factors (PSFs) in digital MCRs. It is found that the PSF multipliers by APJ and RME are highly convergent with each other and have similar numerical values. Taking Time Pressure PSF for example, its multipliers were found to be also consistent with those from other data sources. This study may show the usefulness of expert judgments for producing PSF multipliers in digital MCRs.

Keywords: Human Reliability Analysis, Expert Judgment, Absolute Probability Judgment, Ratio Magnitude Estimation.

1. INTRODUCTION

1.1. Human Reliability Analysis (HRA)

Human reliability analysis (HRA) is a means to identify, model, quantify, and reduce human contributions to system risk. It is a vital component of probabilistic risk analysis (PRA). Since the birth of the first HRA method, Technique for Human Error Rate Prediction (THERP) (Swain and Guttmann, 1983), more than 50 HRA tools have been developed in the nuclear, aviation, space, oil and gas, and other complex, safety-critical domains (Bell and Holroyd, 2009). In the last three years, several HRA methods have been proposed, including the Integrated Decision-Tree Human Event Analysis System (IDHEAS) (Xing et al., in press, Liao, 2015), Phoenix (Ekanem et al., 2016), and Petro-HRA (Laumann and Rasmussen, 2016).

The HRA realm has made progress over these years on several important issues, such as error of commission, psychological foundation, causal model of operator errors, and outcome validation. However, it has less progress on human performance data which is the base of HRA (Kirwan et al., 2008). The lack of data in HRA has been criticized for a long time. Although some human performance databases are developed or under development, such as computerized operator reliability and error database (CORE-DATA) (Gibson and Megaw, 1999) and scenario authoring, characterization, and debriefing application (SACADA) (Chang et al., 2014), this fundamental problem remains the same. It can increase the variability, uncertainty, unreliability of HRA outcomes to PRA. In digital main control rooms (MCRs) of NPPs, the scarcity of data is even more serious. The nuclear industry is at the digitalization age and operator activities are changed by digital techniques in MCRs (Liu and Li, 2016). However, HRA does not change with the trend of MCR digitalization. Its model and data for assessing human reliability are needed to be updated for digital MCRs (Boring, 2014).

1.2. Expert Judgments in HRA

Expert judgment is widely used in PRA and also in HRA (Mosleh et al., 1988, Acharya et al., 1985). Although expert judgment has been criticized for a long time (Mosleh et al., 1988), we have to turn to it when data is unavailable. It is critically important to the use of HRA. Without it, it would be difficult or even

* pengliu@tju.edu.cn

impossible to conduct a PSA/HRA (Spurgin, 2010). Several HRA methods strongly rely on expert judgments, such as the Success Likelihood Index Methodology (SLIM) (Embrey et al., 1984), A Technique for Human Error Analysis (ATHEANA) (Cooper et al., 1996, Forester et al., 2004), and IDHEAS (Xing et al., in press, Liao, 2015). The rationality of expert judgments may be the "wisdom of crowds" (Surowiecki, 2005) that under the right circumstances, a group of people are remarkably intelligent and are able to make right assessments and predictions.

Psychological scaling techniques for expert judgment in HRA include paired comparison, ranking/rating, absolute probability judgment (APJ) (also called as direct numerical estimation), and ratio magnitude estimation (RME) (also called as indirect numerical estimation). The latter two are of interest to this study. APJ and RME techniques have been suggested to estimate HEP in HRA (Seaver and Stillwell, 1983, Comer et al., 1984, Cooper et al., 1996, Basra and Kirwan, 1998). Seaver & Stillwell (1983) argued that they are promising to produce HEP data. In comparison with other techniques, they have relative higher empirical supports and lower data processing requirements, but, lower acceptability to experts (Seaver and Stillwell, 1983). Comer et al. (1984) reported that APJ has high convergent validity with the paired comparison technique and also with the data from the THERP Handbook (Swain and Guttmann, 1983).

APJ and RME have not been used for deriving the PSF multipliers in HRA. The multiplier of a PSF is used to modify the nominal HEP to reflect the negative or positive effects of this PSF on human reliability. In this study, we applied APJ and RME to update the PSF multipliers in the Standardized Plant Analysis of Risk-Human Reliability Analysis (SPAR-H) method (Gertman et al., 2005).

1.3. SPAR-H

SPAR-H is a HRA quantification method, simple and easy to use. A human failure event (HFE) involves diagnosis, action, or both. The nominal HEP for diagnosis is 0.01 and for action is 0.001. The nominal HEPs are modified by the eight PSFs (Available Time, Stress/Stressors, Complexity, Experience/Training, Procedures, Ergonomics/HMI, Fitness for Duty, and Work Processes), in order to quantify the contextual effects on human reliability. A PSF may have two or more levels and corresponding multipliers. Take Time Pressure PSF for example. Table 1 illustrates its levels and multipliers in diagnosis. The multiplier of the nominal level is 1.0. Other levels and multipliers are defined compared with this level, except the "inadequate time" level which directly specifies HEP as 1.0.

Table 1. Levels and Multipliers for Time Pressure Level.

PSF Level	**Level Description**	**PSF Multiplier**
Inadequate time	If the operator cannot diagnose the problem in the amount of time available, no matter what s/he does, then failure is certain	P(failure) = 1.0
Barely adequate time	2/3 the average time required to diagnose the problem is available	10
Nominal time	On average, there is sufficient time to diagnose the problem	1
Extra time	Time available is between one to two times greater than the nominal time required, and is also greater than 30 min	0.1
Expansive time	Time available is greater than two times the nominal time required and is also greater than a minimum time of 30 min; there is an inordinate amount of time (a day or more) to diagnose the problem	0.01

The original SPAR-H was developed for conventional MCRs in NPPs. Its appropriateness in digital MCRs is not warranted. This study collected and aggregated the licensed operators' judgments on subjective probabilities and multipliers by APJ and RME. It is expected to update the PSF multipliers of SPAR-H for informing HRA in digital MCRs. The remaining is organized as follows. Section 2 describes the methodological issues, including experimental design, participants, psychological scale design, and data analysis method. Section 3 gives the results of inter-rater reliability and convergent validity of APJ and RME, and the multipliers of Time Pressure PSF. Section 4 discusses and compares the multipliers of Time Pressure PSF with other data sources as an illustration and gives suggestions on the multipliers of Time Pressure PSF. Section 5 concludes this study.

2. METHODOLOGY

2.1. Experimental Design

The study implemented a within-subjects design. Participants filled APJ and RME with a random order.

2.2. Participants

Licensed operators in a digital MCR of a pressurized water reactor (PWR) NPP participated in this study. They were performing re-training in the class and full-scope simulator following an annual plan. The experimenters (including human factors, HRA, and PRA researchers) visited the MCR for four consecutive weeks, each for the survey of one shift. A total of 44 questionnaires were distributed in the re-training class. The operators were informed the purpose of this study, the definitions of PSFs including their levels, and the way to fill the scales. They completed the scales during the break of the re-training or after work. They delivered the scales to the experimenters in another distribution of the scales or the designated operator instructors.

Thirty-seven operators completed the questionnaires. Among them, data of nine operators were obviously abnormal. For example, they assigned a higher HEP to the nominal level of a PSF and a lower HEP to the negative levels of this PSF, missed one or more PSFs, assigned a ratio value lower than 1.0 to the levels of a PSF in RME, or assigned a HEP higher than 1.0 to the levels of a PSF in APJ. In addition, data of 11 operators had minor flaws. For example, they assigned a ratio value lower than 1.0 to the negative levels of a PSF, which is expected to be higher than 1.0, compared with the nominal levels, or they assigned a HEP higher than 0.1 to the nominal levels of PSF, which is expected to be 0.01.

Finally, the data of 17 operators (female = 2 and male = 15) were kept in the following data analysis. Their age ranged from 30 to 37 (Mean = 32.7, SD = 1.7). The mean years for being licensed was 4.3 (SD = 2.4). Among them, nine were senior reactor operators including instructors, seven were reactor or turbine operators, and the left one operator did not present this information. The relative low percentage (17/37 = 46%) of efficient operators was due to that it was difficult to monitor how the operators fill the questionnaire and to timely respond their possible questions.

2.3. Scale Design

2.3.1. PSF Design

Several deficiencies to the definition of PSFs, their levels and multipliers in SPAR-H have been suggested in the literature (Laumann and Rasmussen, 2016, Forester et al., 2014). It was found that the PSF definitions may be unclear, too broad, and overlapping with each other, that the definitions and indicators of PSF levels may be unclear and ambiguous, and that the multiplier values of PSF levels may be non-transparent and impropriate and do not accommodate the state-of-art knowledge of human performance. We are re-conceptualizing PSFs including their levels in SPAR-H for digital MCRs. Nine PSFs and 44 PSF levels including nine nominal levels were involved. Due to the limited space, only Time Pressure levels/indicators are presented for an illustration (see Table 2).

We use the concept of Time Pressure in place of Available Time and define it as the difference (ratio) between available time and required time (or nominal time) for completing a task. Required/nominal time for a task is defined as the time during which most operators would succeed. If many data points of operation time of the task in NPP simulators are available, required time is operationalized as the 95th or 99th percentile operation time when operators successfully complete the task (Liu and Li, 2014). The five levels for Time Pressure PSF are suggested in Table 2.

Table 2. Levels of Time Pressure PSF in the Current Study

PSF Level	Level Description
Extremely high time pressure	Ratio between available time and required time is less than 0.5
Very high time pressure	Ratio between available time and required time is between 0.5 and 1.0
High time pressure	Ratio between available time and required time is between 1 and 1.5
Nominal level	Ratio between available time and required time is 1.5
Extra time	Ratio between available time and required time is between 1.5 and 2, and available time > 30 min
Expansive time	Ratio between available time and required time is greater than 2, and available time > 30 min

2.3.2. APJ Scale Design

For each PSF, the APJ scale firstly introduced its definition and its levels including level description. Then, it required the participants to assess the probability that operator crews cannot complete a task for each level (assuming the levels of other PSFs are nominal). The instruction for the "extremely high time pressure" level of Time Pressure PSF, for example, was written as follows:

> *Under the "**extremely high time pressure**" level, the probability that crews cannot complete a task is about_______%*

2.3.3. RME Scale Design

In the RME scale, we followed the classical Stevens' method (Stevens, 1957) for magnitude estimation. The RME scale firstly gave the definition of a PSF including its level design. Then it required operators to assess the relative likelihood that operator crews cannot complete a task in different levels of a PSF comparing the nominal level of the PSF (assuming the levels of other PSFs are nominal), given that the multiplier of a nominal level is set to 1. The instruction for the "extremely high time pressure" level of Time Pressure PSF, for example, was written as follows:

> *Under the "**extremely high time pressure**" level, the probability that crews cannot complete a task is about _____ times as high as that under the "**nominal**" level.*

2.4. Data Analysis

APJ obtained the estimated HEP values in various PSF levels. An equally weighted aggregation method was used to determine the single subjective HEP of PSF levels. The estimated multiplier of a PSF level was the subjective HEP of this PSF level divided by that of the nominal levels. RME directly obtained the estimated multipliers of PSF levels.

Expert opinions should be aggregated to generate a point or distribution estimate. The most common aggregation rules are linear (arithmetic) and geometric means. In general, arithmetic mean is higher than geometric mean. Few empirical studies were done to compare those two aggregation rules. Several analytic or simulation studies have mixed results. One analytic study found that arithmetic aggregation is a safer choice than geometric aggregation in reality when experts are not independent and not well-calibrated (Hora, 2010). We are open to the choice of arithmetic or geometric methods. At the current stage of our study, arithmetic aggregation is chosen.

3. RESULTS

3.1. Inter-Rater Reliability

Intra-class correlation (ICC) was used to examine the inter-rater reliability of the two scales. For the present study, of interest is the overall inter-rater consistency of the operators' judgments on PSF levels in each scale. The ICC (C, k) value was calculated (McGraw and Wong, 1996) and independent raters were treated as a random factor to ensure the generality of the ratings. The ICC indices for APJ and RME scales were 0.964 and 0.716, respectively, indicating the acceptable inter-rater reliability across the operators.

3.2. Convergent Validity

The mean HEP in APJ strongly correlated with the mean multiplier in RME (R = 0.930, p < 0.001). According to Stevens' power law (Stevens, 1957), a magnitude of a sensation (e.g., multiplier in RME) may be a power function of the stimulus (e.g., subjective probability in APJ). We checked the logarithmic relationship between HEPs in APJ and multipliers in RME. It was highly significant (R = 0.953, p < 0.001). It indicated a high convergent validity between the two scales.

3.3. PSF Multipliers

The mean HEP in APJ across all PSF nominal levels was 3.7E-2. For APJ, the multiplier of a PSF level was its mean HEP divided by that of all nominal levels. For RME, the multiplier of a PSF level was its mean ratio judgment relative to its nominal level. Table 2 illustrates the mean HEPs and PSF multipliers in APJ scale and the PSF multipliers in RME for Time Pressure PSF. Only the data of Time Pressure PSF are illustrated here, due to the limited space. Figure 1 depicts the relationship between the multipliers obtained by APJ and RME for all levels of the nine PSF. It shows that the PSF multipliers in APJ scale increased with those in RME ($R = 0.930$, $p < 0.001$). On the whole, the multipliers obtained by APJ were somewhat larger than those by RME. The mean difference between them was about 2.0.

Table 2. Mean HEPs and Multipliers in APJ and RME for Time Pressure PSF

PSF	PSF levels	Subjective HEP in APJ	Multiplier in APJ	Multiplier in RME
Time Pressure	Extremely high time pressure	5.5E-01	14.8	7.9
	Very high time pressure	3.9E-01	10.6	4.2
	High time pressure	2.0E-01	5.5	2.2
	Nominal	4.7E-02	1.3	1.0
	Extra time	2.6E-02	0.7	0.7
	Expansive time	2.1E-02	0.6	0.6

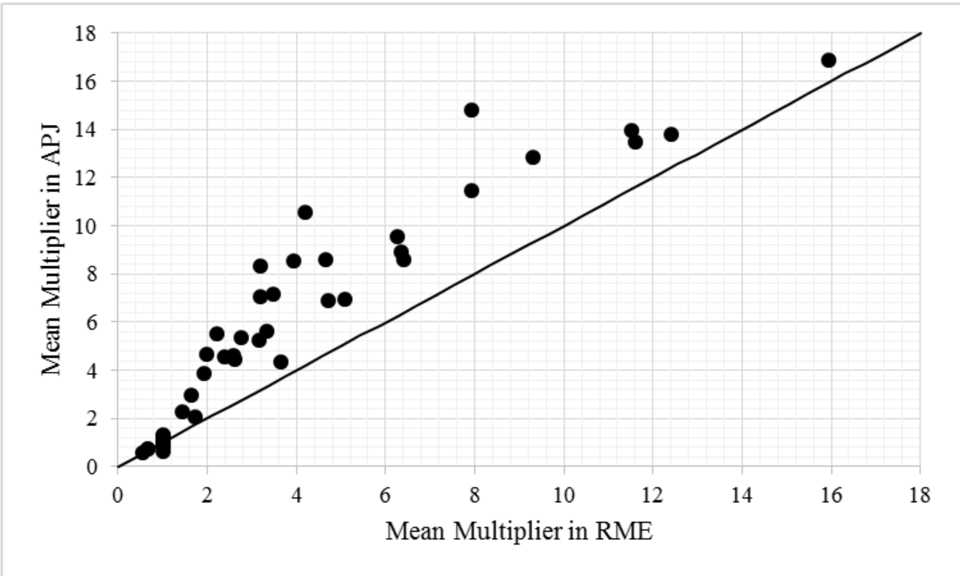

Figure 1. Mean Multipliers in APJ and RME Scales

4. DISCUSSION AND COMPARISON

Expert judgments are widely used in risk analysis. HRA relies much on expert judgments. Few empirical studies have examined the relationship between APJ and RME scales and compared their validity. Our study applied them to obtain expert judgments' on PSF multipliers in digital NPP MCRs for informing HRA. In our application, it is found that the subjective HEPs/multipliers by APJ has high convergent validity with the multipliers by RME. Surprisingly, it is found that the multipliers by the two scales were comparable, although those indirectly obtained by APJ were larger than those directly obtained by RME (see Figure 1).

At present, we neither can make sure that APJ outperforms RME, or vise visa, and nor can make sure that both techniques are feasible for assessing PSF multipliers. Thus, the results by both techniques have to be heavily compared with other data sources. Note that the comparison is difficult due to the lacking of data or the conflicts of data in HRA and human performance literature. For the sake of saving place, only the comparison results for Time Pressure are shown below.

We would like to bring forward a term called as "relative effect". Relative effect between Level 1 and Level 2 is the ratio of multipliers or HEPs at those levels for a specific PSF. If Level 2 is the nominal one, then the relative effect is the multiplier for Level 1. The relative effect of two PSF levels may be assumed to be the same for tasks in different environments (Hallbert et al., 2004, Liu and Li, 2014). Only with this assumption, we can compare the expert judgments with existing experimental studies and other data sources. In addition, it is very difficult to extract the data of PSF multipliers in the human performance literature. Less empirical studies have been done to examine the effect of PSFs in NPP by HRA, human factors, or psychology researchers.

Both the multiplier designs for Time Pressure/Available Time in SPAR-H inherited from the time-reliability curve (TRC) in THERP (Swain and Guttmann, 1983). Swain had a session with experts to draw the TRC and Swain TRC was based on expert judgments (Spurgin, 2010). That is, the multiplier design of Time Pressure/Available Time in SPAR-H was based on expert judgments. Few studies have examined the appropriateness of the multiplier design for Time Pressure/Available Time in SPAR-H. The following comparison shows that SPAR-H's multiplier design may be divergent from our expert judgments and other empirical studies.

4.1. HEP of the Most Negative Level

For the most negative level, SPAR-H describes that "if the operator cannot diagnose the problem in the amount of time available, no matter what s/he does, then failure is certain" (Gertman et al., 2005, p. 20) and assigns the HEP of 1.0 to this level. If operators do not have time to perform, then definitely, they will fail 100%. At the first glance, that this level with a HEP of 1.0 in SPAR-H is appropriate. However, SPAR-H does not define the case in which operators cannot diagnose in the available time or the case in which operators do not have enough time to perform the task. It does not give a clear operationalization for this most negative level. Thus, although this level seems to be true literally, it is useless. Someone may argue that when available time is less than that defined in the "barely adequate time" level in SPAR-H (also see Table 1), it can be assigned to this level. It also seems to be true. However, the predicted HEP of 1.0 for this level of Time Pressure PSF may be overestimated and too pessimistic. Swain TRC (Swain and Guttmann, 1983) and its successor SPAR-H assume that operators have a very high probability to fail in a short time. They may not realize the compensatory control ability (Hockey, 1997) and maximal adaptability (Hancock and Warm, 1989) of humans to deal with stressful conditions, and high creativity of humans to mitigate accidents in such conditions (Apostolakis, 2004). As Strater (2005) and Spurgin (2010) argued, even when available time is very short, humans can perform efficiently. Strater even pointed out that time "plays a minor role for reliability" (Strater, 2005, p. 190). In our study, the estimated HEP of the most negative level (available time is less than half required time) in APJ was 5.5E-1. This level can be assigned as the "inadequate time" level in SPAR-H. And then the HEP that is predicted to be 1.0 in SPAR-H is larger than the estimated HEP for this level in our APJ. The estimated HEP in APJ may be more appropriate. The required time is operationalized as the 95^{th} or 99^{th} operation time when operators successfully complete a task in this study. If operators do not perceive the shortage of time window, then the HEP of the most negative level of Time Pressure PSF would be near 5.0E-1. The estimated HEP of this level in APJ happened to approach this value.

4.2. Multiplier of the Most Negative Level

The multiplier of the most negative level was estimated to be 14.8 in APJ and 7.9 in RME, respectively. For two data-based HRA methods, Human Error Assessment and Reduction Technique (HEART) (Williams, 1988) and Nuclear Action Reliability Assessment (NARA) (Kirwan et al., 2004), the maximum multiplier for Time Pressure is 11, which is closer to the highest multiplier estimated by APJ. For another HRA method, Cognitive Reliability and Error Analysis Method (CREAM) (Hollnagel, 1998), the multiplier for Available Time is five. The value in CREAM would be the average one rather than the highest one. If so, the averaged multiplier by APJ is closer to the multiplier in CREAM. Despite the support of other HRA methods for APJ, one study (Kubicek, 2009) tends to support RME. Kubicek analyzed the data from Task Complexity Experiment in the OECD Halden Reactor Project (Laumann et al., 2005), coded PSFs using a 3-level scale (poor, nominal, and good), and estimated that the relative effect between the poor and nominal levels of Available Time PSF is 2.5. Kubicek did not explicitly describe the three levels for available time. Thus, for Time Pressure (or Available Time) PSF, its nominal level in our study may be not equivalent to that in Kubicek's study.

4.3. Relative Effect between the "Very High Time Pressure" and "High Time Pressure" Levels

We would like to discuss the relative effect between the "very high time pressure" and "high time pressure" levels. It is 1.9 in both APJ (=10.6/5.5) and RME (=4.2/2.2). Liu and Li (2014) investigated the effect of time availability (five levels), experience (unskilled vs. skilled), and task complexity on HEP in a simulated emergency operating procedure (EOP) and found that the relative effect between the lowest available time design (available time is 0.8 times required time, corresponding the "very high time pressure" level) and the highest available time design (available time is 1.2 times required time, corresponding the "high time pressure" level) was 2.2 in the unskilled phase and 2.5 in the skilled phase (Liu and Li, 2014). Another study (Zhao et al., 2012) investigated the effect of time availability on the execution of simulated EOPs and manipulated five levels for which the available time is 0.8, 1, 1.2, 1.4, and 1.6 times the standard operation time. This standard operation time was defined as the average step operation time when the participants were skilled, which is lower than the required time defined in this study. Considering the difference between the standard operation time and the required time, we assume that the level with 0.8 times the standard operation time is classified as the "very high time pressure" level and other four levels as the "high time pressure" level. The available time design in the five levels had two types: absolute (i.e., same available time for the same level for all of the participants) and relative (i.e., available time was dependent on the individual's standard operation time). Of interest is the absolute time design in this study. The trial error rates for the five available time levels were 0.59, 0.43, 0.31, 0.18, and 0.19, respectively. Then, the relative effects between the "very high time pressure" level and other four "high time pressure" levels are 1.4 (= 0.59/0.43), 1.9 (= 0.59/0.31), 3.3 (= 0.59/0.18), and 3.0 (= 0.59/0.19). If the two levels with the most available time (1.4 and 1.6 times the standard operation time) can be classified as the "nominal" level, then the relative effects between the "very high time pressure" level and the two "nominal" levels are 3.3 and 3.0 and the relative effects between the two "high time pressure" levels and the two "nominal" levels range from 1.6 (=0.31/0.19) to 2.4 (=0.43/0.18). For the relative effect between the "very high time pressure" and "high time pressure" levels, the two experimental studies and the two judgment techniques have similar results. It can be assumed to be 2.

However, one experimental study on diagnosis performance (Chen and Li, 2015) showed that the difference of diagnosis accuracy in its high time pressure treatment (available time is 0.8 times required time, mean diagnosis accuracy = 0.69, corresponding the "very high time pressure" level) and low time pressure treatment (available time is 1.2 times required time, mean diagnosis accuracy = 0.64, corresponding the "high time pressure" level) was not significant, showing that the relative effect between these two levels is approaching 1.0 in Chen and Li's study.

4.4. Relative Effect between the "High Time Pressure" and "Nominal" Levels

The relative effect between the "high time pressure" level and "nominal" level is 4.2 (=5.5/1.3) in APJ and 2.2 in RME. The two techniques produced inconsistent results. If the two levels with the most available time (1.4 and 1.6 times the standard operation time) (Zhao et al., 2012) are classified as the "nominal" level, then the relative effects between the "high time pressure" level and the "nominal" level range from 1.6 to 2.4. For the nominal level, human would not perceive time pressure. If so, the empirical studies involving no time pressure and time pressure can be referred to. In Chen and Li's study (Chen and Li, 2015), the mean diagnosis accuracy without time pressure in its pilot study (which was not reported in this study) is 0.86. The relative effect between low time pressure (available time is 1.2 times required time, corresponding the "high time pressure" level) and no time pressure is: (1-0.64)/(1-0.86)=2.6. The relative effect obtained from another study (Lin and Su, 1998) between time pressure and no time pressure is 1.6 and 1.9 in two different treatments (with expert system vs. without expert systems). The relative effect between the "high time pressure" and nominal level may be assumed to be 3.0.

4.5. Multipliers of Positive Levels

The multipliers of the two positive levels ("extra time" and "expansive time") in SPAR-H (Gertman et al., 2005) are 0.1 and .01, respectively. SPAR-H highlights the positive effect of time adequacy to reduce HEP. From the nominal level to the "expansive time" level, the predicted HEP reduces by 99%. As aforementioned, time is an important PSF but not a dominant one. Thus, the highly positive effect of increasing time availability on reducing HEP may be doubtful. For the multipliers of two positive levels ("extra time" and "expansive time"), APJ and RME obtained the almost same results, 0.7 and 0.6, respectively. CREAM (Hollnagel, 1998) also assumes the weight of adequate level as 0.5, which seems to be consistent with our finding. We may not assign a very small multiplier for the positive levels of Time Pressure PSF. It seems that two positive levels for Time Pressure PSF may be not necessary. The current two positive levels can be collapsed into one positive level. The multiplier could then be assumed to be 0.5.

4.6. Suggestions for Time Pressure Multipliers

Finally, we would like to suggest the multipliers for Time Pressure PSF, as following:

- Extremely high time pressure: multiplier = 12;
- Very high time pressure: multiplier = 6;
- High time pressure: multiplier = 3
- Nominal level: multiplier = 1;
- Expansive time: multiplier = 0.5.

Note that this suggestion is not the final one. As Moray (Moray, 1990) suggested, we do not make sure that the multiplier obtained from, say, aviation or laboratory experiments, can be transferred to a superficially similar task in MCRs. The comparison results should be used cautiously. Next we will review more empirical studies and even design specific EOP experiments to collect more data about the multipliers of Time Pressure PSF and other PSFs.

5. CONCLUSIONS

This study applied different expert judgment techniques to estimate the PSF multipliers in digital MCRs. To our best knowledge, it is the first time. Licensed operators in digital MCRs were surveyed to make their judgments on PSF multipliers. It was found that the multipliers obtained APJ and RME had a very high convergent validity and were close in terms of numerical values. Take Time Pressure PSF for example, the estimated multipliers by APJ and RME were compared with other data sources and certain consistencies were found. It may imply that expert judgment is feasible to generate and update the PSF multipliers in digital MCRs.

6. ACKNOWLEDGEMENTS

This study was supported by the National Natural Science Foundation of China (Project no. 71371104 and 71601139), and Advanced Pressurized Water Reactor NPP Key Project (Project no. 2011ZX06002).

7. REFERENCES

Acharya, S., Glynn, J., Agrawal, B., Johnson, J., Cunningham, M., Niyogi, P., Denning, R. & VanderMole, H. (1985). *Severe Accident Risks: An Assessment for Five U.S. Nuclear Power Plants*. NUREG-1150 Vol. 1. Washington, D.C.: U.S. Nuclear Regulatory Commission.

Apostolakis, G. E. (2004). How useful is quantitative risk assessment? *Risk Analysis,* 24**,** 515-520.

Basra, G. & Kirwan, B. (1998). Collection of offshore human error probability data. *Reliability Engineering and System Safety,* 61**,** 77-93.

Bell, J. & Holroyd, J. (2009). *Review of Human Reliability Assessment Methods*. Derbyshire: Health and Safety Laboratory.

Boring, R. L. (2014). Human reliability analysis for digital human-machine interfaces: a wish list for future research. In*: Proceedings of the 12th Probabilistic Safety Assessment and Analysis (PSAM12)*, June 22-27, 2014, Honolulu, Hawaii.

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**,** 117-133.

Chen, K. & Li, Z. (2015). How does information congruence influence diagnosis performance? *Ergonomics,* 58**,** 924-934.

Comer, M. K., Seaver, D. A., Stillwell, W. G. & Gaddy, C. D. (1984). *Generating Human Reliability Estimates Using Expert Judgment Volume 1. Main Report*. NUREG/ CR-3688/ 1 of 2. Washington, D.C.: U.S. Nuclear Regulatory Commission.

Cooper, S. E., Ramey-Smith, A. M., Wreathall, J., Parry, G. W., Bley, D. C., Luckas, W. J., Taylor, J. H. & Barriere, M. T. (1996). *A Technique for Human Error Analysis (ATHEANA): Technical Basis and Methodology Description*. NUREG/CR-6350. Washington, D.C.: U.S. Nuclear Regulatory Commission.

Ekanem, N. J., Mosleh, A. & Shen, S.-H. (2016). Phoenix–A model-based Human Reliability Analysis methodology: Qualitative analysis procedure. *Reliability Engineering and System Safety,* 145**,** 301-315.

Embrey, D. E., Humphreys, P., Rosa, E. A., Kirwan, B. & Rea, K. (1984). *An Approach to Assessing Human Error Probabilities Using Structured Expert Judgment*. NUREG/CR-3518. Washington, D.C.: U.S. Nuclear Regulatory Commission.

Forester, J., Bley, D., Cooper, S., Lois, E., Siu, N., Kolaczkowski, A. & Wreathall, J. (2004). Expert elicitation approach for performing ATHEANA quantification. *Reliability Engineering and System Safety,* 83**,** 207-220.

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

Gertman, D. I., Blackman, H., Marble, J., Byers, J. & Smith, C. (2005). *The SPAR-H Human Reliability Analysis Method*. NUREG/CR-6883. Washington, D.C.: U.S. Nuclear Regulatory Commission.

Gibson, W. H. & Megaw, E. D. (1999). *The Implementation of CORE-DATA, a computerised human error probability database*. HSE contract research report 245/1999. Suffolk: Health and Safety Executive.

Hallbert, B., Gertman, D., Lois, E., Marble, J., Blackman, H. & Byers, J. (2004). The use of empirical data sources in HRA. *Reliability Engineering and System Safety,* 83**,** 139-143.

Hancock, P. A. & Warm, J. S. (1989). A dynamic model of stress and sustained attention. *Human Factors,* 31**,** 519-537.

Hockey, G. R. J. (1997). Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. *Biological Psychology,* 45**,** 73-93.

Hollnagel, E. (1998). *Cognitive Reliability and Error Analysis Method.* Oxford, UK: Elsevier Science Ltd.

Hora, S. C. (2010). An analytic method for evaluating the performance of aggregation rules for probability densities. *Operations Research,* 58**,** 1440-1449.

Kirwan, B., Gibson, H., Kennedy, R., Edmunds, J., Cooksley, G. & Umbers, I. (2004). Nuclear action reliability assessment (NARA): a data-based HRA tool. In*: PSAM7&ESREL2004*, 14-18 June, 2004, Berlin, Germany.

Kirwan, B., Gibson, W. H. & Hickling, B. (2008). Human error data collection as a precursor to the development of a human reliability assessment capability in air traffic management. *Reliability Engineering and System Safety,* 93**,** 217-233.

Kubicek, J. (2009). Using human event data to validate PSF multipliers: a proof-of-concept study. In*: Workshop Proceedings of Simulator Studies for HRA Purposes*, Nov. 4-6, 2009, Budapest, Hungary.

Laumann, K., Braarud, P. Ø. & Svengren, H. (2005). *The Task Complexity Experiment 2003/2004*. HWR-758. Halden, Norway: OECD Halden Reactor Project.

Laumann, K. & Rasmussen, M. (2016). Suggested improvements to the definitions of Standardized Plant Analysis of Risk-Human Reliability Analysis (SPAR-H) performance shaping factors, their levels and multipliers and the nominal tasks. *Reliability Engineering and System Safety,* 145**,** 287-300.

Liao, H. (2015). Insights from pilot testing of the IDHEAS HRA method. *Procedia Manufacturing,* 3**,** 1350-1357.

Lin, D.-Y. M. & Su, Y.-L. (1998). The effect of time pressure on expert system based training for emergency management. *Behaviour & Information Technology,* 17**,** 195-202.

Liu, P. & Li, Z. (2014). Human error data collection and comparison with predictions by SPAR-H. *Risk Analysis,* 34**,** 1706-1719.

Liu, P. & Li, Z. (2016). Comparison between conventional and digital nuclear power plant main control rooms: A task complexity perspective, Part I: Overall results and analysis. *International Journal of Industrial Ergonomics,* 51**,** 2-9.

McGraw, K. O. & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods,* 1**,** 30-46.

Moray, N. (1990). Dougherty's dilemma and the one-sidedness of human reliability analysis (HRA). *Reliability Engineering and System Safety,* 29**,** 337-344.

Mosleh, A., Bier, V. M. & Apostolakis, G. (1988). A critique of current practice for the use of expert opinions in probabilistic risk assessment. *Reliability Engineering and System Safety,* 20**,** 63-85.

Seaver, D. A. & Stillwell, W. G. (1983). *Procedures for Using Expert Judgment to Estimate Human Error Probabilities in Nuclear Power Plant Operations*. NUREG/CR-2743. Washtington, D.C.: U.S. Nuclear Regulatory Commission.

Spurgin, A. J. (2010). *Human Reliability Assessment Theory and Practice.* London: CRC Press.

Stevens, S. S. (1957). On the psychophysical law. *The Psychological Review*, 64**,** 153-181.

Strater, O. (2005). *Cognition and Safety: An Integrated Approach to Systems Design and Assessment.* Burlington, VT: Ashgate.

Surowiecki, J. (2005). *The Wisdom of Crowds.* New York: A Division of Random House, Inc.

Swain, A. D. & Guttmann, H. E. (1983). *Handbook of Human Reliability Analysis with Emphasis on Nuclear Power Plant Applications*. NUREG/CR-1278. Washington, D.C.: U.S. Nuclear Regulatory Commission.

Williams, J. C. (1988). A data-based method for assessing and reducing human error to improve operational performance. In*: Proceedings of the IEEE 4th Conference on Human Factors in Power Plants*, June 5-9, 1988, Monterey, CA.

Xing, J., Presley, M., Parry, G., Forester, J., Hendrickson, S. & Dang, V. (in press). *An Integrated Decision-Tree Human Event Analysis System (IDHEAS) Method for NPP Internal At-Power Operation*. Draft Report. Washington, D.C.: U.S. Nuclear Regulatory Commission.

Zhao, F., Dong, X. & Li, Z. (2012). How does time availability influence the execution of computerized emergency operating procedures. In*: Proceedings of the 11th International Probabilistic Safety Assessment and Management Conference (PSAM11) and the Annual European Safety and Reliability Conference (ESREL 2012)*, 25-29 June, 2012, Helsinki, Finland.