Quantitative Assessment of Human Error Probabilities and PSF Multipliers for Railway: An Expert Elicitation Study

Taewon Yang[a], Seong Keun Kang[b], and Jonghyun Kim[a*]
[a] Department of Nuclear and Quantum Engineering, Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea, {taewon.yang, jonghyun.kim}@kaist.ac.kr
[b] Department of System Safety Research, Korea Railroad Research Institute, 176, Cheoldobangmulgwan-ro, Uiwang-si, Gyeonggi-do, Republic of Korea, skkang@krri.re.kr

---

Abstract: Railway accidents occurring during operation, facility management, or infrastructure maintenance can lead to catastrophic casualties, as well as significant social and economic losses. According to statistics from the Ministry of Land, Infrastructure and Transport, South Korea experienced 86 major train accidents (e.g., collisions, derailments, and fires) between 2020 and 2024, resulting in 178 casualties. Notably, various research and reports indicate that approximately 50% of these incidents are attributed to human errors by railway employees, including drivers and controllers. Consequently, Human Reliability Analysis (HRA) methods are needed within the railway industry to quantitatively evaluate Human Error Probability (HEP), reduce error occurrence, and enhance task performance. However, performing reliable HEP assessments remains a challenge due to the scarcity of empirical data required to quantify nominal HEP and Performance Shaping Factors (PSFs). The present study aims to generate highly reliable HRA data by applying a systematic approach to use experts' knowledge. This study employs Cooke's Classical Model to quantitatively estimate the nominal HEPs and PSF multipliers required for HRA by systematically aggregating expert judgments. Expert opinions were elicited via a structured questionnaire administered to a panel of 16 experts with extensive experience and specialized knowledge in the South Korean railway industry. Regarding the target questions, nominal HEPs for diagnosis and execution errors are elicited, along with multiplier values for each level of eight predefined PSFs. For the seed variables, seed questions are formulated based on nominal HEPs and PSF multipliers drawn from existing HRA research in railway and other industrial domains. Based on the experts' judgments regarding the seed variables, informativeness and statistical accuracy are evaluated to calculate information and calibration scores. Utilizing these metrics, specific weights are assigned to each expert, and an aggregated assessment of the target questions is generated using the "expert" package in R.

---

1. INTRODUCTION

Railway accidents encompass both railway traffic accidents associated with the operation of rolling stock and railway safety accidents related to the management of railway operations and facilities. Such accidents, occurring during operation, facility management, or infrastructure maintenance, can lead to catastrophic casualties as well as significant social and economic losses. According to statistics from the Ministry of Land, Infrastructure and Transport, South Korea experienced 86 major train accidents (e.g., collisions, derailments, and fires) between 2020 and 2024, resulting in 178 casualties among railway employee [1].

Among the principal causes of such accidents, human factors account for a substantial proportion. Various studies and reports indicate that approximately 50% of these incidents are attributed to human error by railway employees, including train operators and traffic controllers [2-5]. Moreover, South Korea's railway accident casualty rate is reported to be considerably higher than that of major advanced nations such as Germany and France [6]. These figures indicate that, despite continued safety investment heavily concentrated on facility and rolling-stock measures, the contribution of human error to railway accidents remains significant and insufficiently addressed.

Given this context, systematic measures are required to reduce human errors committed by railway employees. In particular, human error data should be collected systematically in accordance with the work environments and task characteristics of railway employees, and the underlying factors that induce such errors should be identified to enable accidents to be prevented in advance [7]. At present, however, foundational research on the factors influencing human error of railway employees remains insufficient, which constrains efforts to fundamentally reduce human error of railway employees in key roles such as train operators, traffic controllers, and maintenance workers [8, 9].

One promising approach to analyzing and mitigating human error of railway employees are the application of Human Reliability Analysis (HRA). HRA is a technique that identifies the potential for human error by operators or workers and quantitatively estimates its probability of occurrence within the probabilistic safety assessment of high-hazard facilities [10-13]. Because human error arises in a limited manner under specific task and environmental conditions, HRA-based qualitative and quantitative analyses utilizing expert knowledge and limited data are widely employed in safety-critical domains such as nuclear power, railways, aviation, and healthcare.

A typical HRA methodology estimates the final human error probability (HEP) by applying multipliers, which reflect the influence of Performance Shaping Factors (PSFs), to a baseline nominal HEP. The nominal HEP denotes the basic probability of error that may occur when a worker performs a given task, whereas PSFs represent the various factors influencing worker performance, acting to increase or decrease the HEP according to their respective weights. Accordingly, the quantification of human error necessitates three essential components: the estimation of the task-specific nominal HEP, the derivation of relevant PSFs, and the quantification of the influence of each PSF. Nevertheless, performing reliable HEP assessments remains challenging owing to the scarcity of the empirical field data required to quantify the nominal HEP and PSF multipliers, particularly within the railway domain.

To address this limitation, the present study aims to quantify the nominal HEP and PSF multipliers necessary for the development of a railway-specific HRA methodology, thereby establishing foundational data for the quantitative assessment of HEP. To this end, this study adopts Cooke's classical model, a structured expert-judgment framework that derives rational consensus estimates by weighting and aggregating the assessments of multiple experts according to their demonstrated statistical accuracy and informativeness. Expert opinions were elicited through a structured questionnaire administered to a panel of sixteen experts with extensive experience and specialized knowledge in the South Korean railway industry.

The remainder of this paper is organized as follows: Section 2 introduces the fundamental concepts of the classical model; Section 3 describes the expert elicitation methodology, including the design of seed variables and target questions; Section 4 presents and discusses the elicitation results; and the final section concludes the study.

2. CLASSICAL MODEL

The classical model, as formulated by Cooke [14, 15], offers a systematic framework for deriving rational consensus estimates in situations where sufficient empirical field data are lacking. In this approach, assessments are solicited from subject-matter experts, thereby constructing probability distributions that encapsulate their collective knowledge and experience. Consensus is established by evaluating both the statistical accuracy and the informativeness of each expert's uncertainty quantification, and subsequently assigning differential weights to individual judgments. The underlying premise of the classical model is that an expert's proficiency in quantifying uncertainty for unknown quantities can be inferred from their demonstrated accuracy when assessing variables with known outcomes. The fundamental concepts of the classical model are introduced in this section.

2.1. Seed Variable and Target Question

The classical model employs two distinct categories of questionnaires for expert elicitation: seed variables and target questions. For each question type, a panel of selected experts provides responses that reflect their subjective uncertainty. Seed variables correspond to quantities whose true values are known to the analyst, yet remain undisclosed to the experts throughout the elicitation process. Seed variables serve three principal functions [15]:

1) To quantify each expert's performance as a subjective probability assessor,
2) To enable performance-optimized aggregation of expert probability distributions, and
3) To evaluate and, where possible, validate the resulting combination of expert judgments.

Target questions, by contrast, pertain to variables that necessitate novel quantitative evaluation by the analyst. In the context of this study, the nominal HEP and the quantitative multiplier of the PSF applicable to railway area employees constitute the subject of the target questions.

For both question types, uncertainty is characterized using predefined 5th, 50th, and 95th percentile estimates. A well-calibrated uncertainty judgment is indicated when an expert's elicited distribution closely approximates the theoretical distribution defined by these percentile bounds. The relative contribution of each expert is quantified through a weight, computed as the product of their calibration score *C(e)* and information score *I(e)*. A specific aggregation of expert assessments yields a 'virtual' expert whose combined calibration and information score can be maximized through the exclusion of poorly calibrated experts [16]. This 'virtual' expert is referred to as the Decision Maker (DM), who integrates the target question responses from all participating experts by applying weights derived from their respective seed variable performance. In this manner, the DM functions as a statistically optimal composite expert that consolidates individual judgments into a single, performance-weighted probability distribution.

2.2. Calibration Score

The calibration score *C(e)* quantifies the statistical accuracy of an expert's uncertainty assessments with respect to the seed variables. This score is derived from the relative information *I(s(e),p)*, which measures the divergence between an expert's elicited probability distribution *s(e)* and the theoretical reference distribution *p*. In this formulation, *s(e)* denotes the empirical probability distribution function for expert *e*, whereas *p* represents the theoretical probability distribution function corresponding to the j-th quantile interval. When employing the 5th, 50th, and 95th percentiles as quantile boundaries, the theoretical distribution *p* takes the form (0.05, 0.45, 0.45, 0.05). The relative information is computed via Eq. (1), which serves as a quantitative measure of the degree to which an expert's elicited distribution deviates from the theoretically expected distribution.

$$I(s(e),p) = \sum_{j=1}^{4} s_j(e)\, ln\frac{s_j(e)}{p_j} \tag{1}$$

The calibration score C(e) is subsequently obtained by transforming the relative information into a formal statistical measure, as expressed in Eq. (2), where *x* follows a chi-squared distribution with 3 degrees of freedom, and *N* denotes the total number of seed variables employed in the elicitation.

$$C(e) = 1 - \chi_3^2(2N \cdot I(s(e),p)) \tag{2}$$

2.3. Information Score

The information score *I(e)* quantifies the concentration of an expert's probability assessments, reflecting the degree to which the expert's elicited ranges are narrower or broader relative to the intrinsic range. This intrinsic range is bounded by the lowest value *L* and the highest value *U* observed across all experts' responses for a given seed variable. The corresponding formulation is presented in Eq. (3), where the overshoot parameter k is conventionally set to 0.1.

$$[L^*, U^*] = [L - k(U - L), U + k(U - L)] \tag{3}$$

Once the intrinsic range has been established, the information score is computed using Eq. (4), where N denotes the number of items assessed by the expert, and q_{5i}, q_{50i}, and q_{95i} correspond to the 5th, 50th, and 95th percentile estimates for the i-th item, respectively. A higher information score indicates a greater degree of concentration in the expert's assessments, thereby reflecting a higher level of informativeness.

$$\mathrm{I(e)} = \frac{1}{N}\sum_{i=1}^{N}\left[\ln(U^* - L^*) + 0.05\ln\frac{0.05}{q_{5i}-L^*} + 0.45\frac{0.45}{q_{50i}-q_i^5} + 0.45\frac{0.45}{q_{95i}-q_{50i}} + p_4\ln\frac{0.05}{U^*-q_{95i}}\right] \tag{4}$$

2.4. Derivation of Decision-Makers

A DM, serving as a virtual expert, is constructed by aggregating individual expert opinions through the integration of calibration and information scores. The initial DM is derived based on performance-based weights assigned to each expert, where each weight is determined by the combined score — defined as the product of the calibration score and the information score — as formulated in Eq. (5).

$$w_i = \frac{CS(e_i)}{\sum_{j=1}^{N} CS(e_j)} \tag{5}$$

The DM can be iteratively re-evaluated based on performance across the seed variables, enabling a reassessment of the DM's overall statistical accuracy. This performance-based weighting scheme optimizes the DM's weight by applying a significance threshold to individual calibration scores, thereby systematically excluding experts whose calibration performance falls below the prescribed level.

In other words, the significance level represents the criterion that ensures the retained subset of experts collectively maximizes the DM's combined performance, defined as the product of the calibration and information scores. Ultimately, the optimized DM weight is computed using Eq. (6). Here, α denotes the significance level, and $1\{\, Cal(e_j) \geq \alpha \,\}$ is an indicator function that returns a value of 1 when the specified condition is satisfied, and 0 otherwise.

$$w_\alpha(e_j) = \frac{Cal(e_j)\cdot Inf(e_j)\cdot 1\{Cal(e_j)\geq\alpha\}}{\sum_{j=1}^{N} Cal(e_j)\cdot Inf(e_j)\cdot 1\{Cal(e_j)\geq\alpha\}} \tag{6}$$

3. ELICITATION OF EXPERT OPINION

This section outlines the methodology employed for expert elicitation based on the classical model. The primary objective of this study is to quantify the nominal HEP and PSF multipliers for HEP assessment of railway area employees. To this end, distinct questionnaires were developed separately for the elicitation of nominal HEP estimates and PSF multipliers. The elicitation process was carried out through a structured approach encompassing the development of both target questions and seed variables. Detailed guidelines from prior studies were referenced to ensure methodological rigor [17, 18].

3.1. Seed Variables

Both the nominal HEP questionnaire and the PSF multiplier questionnaire comprised nine seed variables designed to assess the reliability of expert judgments. Of the nine seed variables pertaining to the nominal HEP, three focused on error probabilities for execution and omission errors commonly referenced in THERP (Technique for Human Error Rate Prediction), a widely recognized HRA method [10]. The remaining six were formulated based on situation-specific HEP values for train operators, as presented in RARA (Railway Action Reliability Assessment), an HRA methodology developed for the railway domain [13, 19].

The nine seed variables associated with the PSF multiplier were developed with reference to RARA and data from prior studies grounded in the RARA framework. From the 16 PSFs proposed in the existing literature, nine were selected for inclusion by excluding those deemed less relevant to the PSFs evaluated in the target questions of the present study. For the seed variables pertaining to the PSF multiplier, experts were asked to estimate the factor by which the HEP would increase when a given PSF is in effect. Table 1 shows the questions from the two types of seed variable surveys.

Table 1: Seed Variable's Questions in the Nominal HEP and PSF multiplier Surveys

Questionnaire	No	Questions
Nominal HEP	NHEP 1	Probability of human error by a worker when performing a task that requires following system instructions as given
	NHEP 2	Probability of human error by a worker when performing a familiar task that is repeated on a regular basis
	NHEP 3	Probability of human error by a worker in a task that requires simply responding to a specific action-required alert and executing the prescribed procedure
	NHEP 4	Probability of human error by a worker when performing a communication task that has potential for confusion
	NHEP 5	Probability of human error by a worker when performing a task that involves restoring a system to its previous state or transitioning it to a new state according to established procedures
	NHEP 6	Probability of human error by a train operator when identifying an alarm or abnormal indication and assessing and responding to the situation
	NHEP 7	Probability of human error occurring when a worker incorrectly performs a task specified in procedural documents
	NHEP 8	Probability of human error by a worker in omitting a task related to equipment status or work items not displayed on control room monitors or instrument panels
	NHEP 9	Probability of human error by a worker in skipping or omitting an intermediate step when executing a multi-step work procedure
PSF Multiplier	PSF 1	Cases where the environmental conditions of the workspace — both personal/systemic factors and in-transit climate conditions — are unfavorable compared to comfortable conditions, negatively affecting cognitive and perceptual performance of the train operator
	PSF 2	Cases where available time margin for the train operator to perform a task becomes insufficient due to delays, breakdowns, or construction, compared to when adequate time margin is available
	PSF 3	Cases where the train operator experiences high workload due to simultaneously performing multiple tasks or taking actions in response to abnormal situations
	PSF 4	Cases where the train operator experiences a loss of concentration due to boredom, simple repetitive tasks, or similar factors while performing a task
	PSF 5	Cases where the train operator is required to perform an unfamiliar task, compared to performing a familiar task
	PSF 6	Cases where the train operator underestimates or overestimates the risk associated with the task being performed
	PSF 7	Cases where feedback information received by the train operator from the system or fellow workers is inadequate or insufficient
	PSF 8	Cases where the train operator must rely solely on information from a single source
	PSF 9	Cases where the train operator is experiencing emotional stress or health-related issues

Collectively, these seed variables were employed to assess the statistical accuracy and informativeness of each expert's responses, to derive performance-based weights, and to ultimately integrate the elicited judgments into the final aggregated results.

3.2. Target Questions

The target questions of the nominal HEP questionnaire ask experts to evaluate the diagnosis error probability and action error probability of railway area employees, respectively. Specifically, a total of eight questions were developed to assess both the diagnosis error probability and action error probability for each of the four representative employee types in the railway domain: train operators, traffic controllers, maintenance workers, and track workers. Additionally, illustrative examples were provided alongside each question to assist the responding experts.

The target questions of the PSF multiplier questionnaire, by contrast, consist of fourteen items in total. The eight PSFs evaluated in this study are Workload, Equipment/HMI, Procedures/Guidelines, Experience/Training, Work Process, Complexity, Communication, and Environment. Among these, Workload, Equipment/HMI, Procedures/Guidelines, Experience/Training, Work Process, and Complexity each encompass three qualitative levels including the nominal level, whereas Communication and Environment each comprise two qualitative levels including the nominal level. For the PSF multiplier questionnaire, the multiplier at the nominal level was assumed to be 1, and experts were asked to assess the magnitude of change in the multiplier as the qualitative level varies. Table 2 shows the questions from the two types of target question surveys.

Table 2: Target Question's items in the Nominal HEP and PSF multiplier Surveys

Questionnaire	No	Questions
Nominal HEP	NHEP 1	Error probability of train operators when performing diagnosis-related tasks
	NHEP 2	Error probability of train operators when performing action-related tasks
	NHEP 3	Error probability of traffic controllers when performing diagnosis-related tasks
	NHEP 4	Error probability of traffic controllers when performing action-related tasks
	NHEP 5	Error probability of maintenance workers when performing diagnosis-related tasks
	NHEP 6	Error probability of maintenance workers when performing action-related tasks
	NHEP 7	Error probability of track workers when performing diagnosis-related tasks
	NHEP 8	Error probability of track workers when performing action-related tasks
PSF Multiplier	PSF 1	Effect on human error probability when the employee's workload level is poor
	PSF 2	Effect on human error probability when the employee's workload level is very poor
	PSF 3	Effect on human error probability when the equipment/HMI used by the employee is at a poor level
	PSF 4	Effect on human error probability when the equipment/HMI used by the employee is at a good level
	PSF 5	Effect on human error probability when the procedures/guidelines used by the employees are at a poor level
	PSF 6	Effect on human error probability when the procedures/guidelines used by the employees are at a good level
	PSF 7	Effect on human error probability when the employee's training or experience is at a poor level
	PSF 8	Effect on human error probability when the employee's training or experience is at a good level

	PSF 9	Effect on human error probability when the work process of the employee's task environment is at a poor level
	PSF 10	Effect on human error probability when the work process of the employee's task environment is at a good level
	PSF 11	Effect on human error probability when the complexity of the task performed by the employee is at a high level
	PSF 12	Effect on human error probability when the complexity of the task performed by the employee is at a very high level
	PSF 13	Effect on human error probability when the quality of communication among employee is at a poor level
	PSF 14	Effect on HEP when the environment conditions of the workspace in which the employee performs tasks are at a poor level

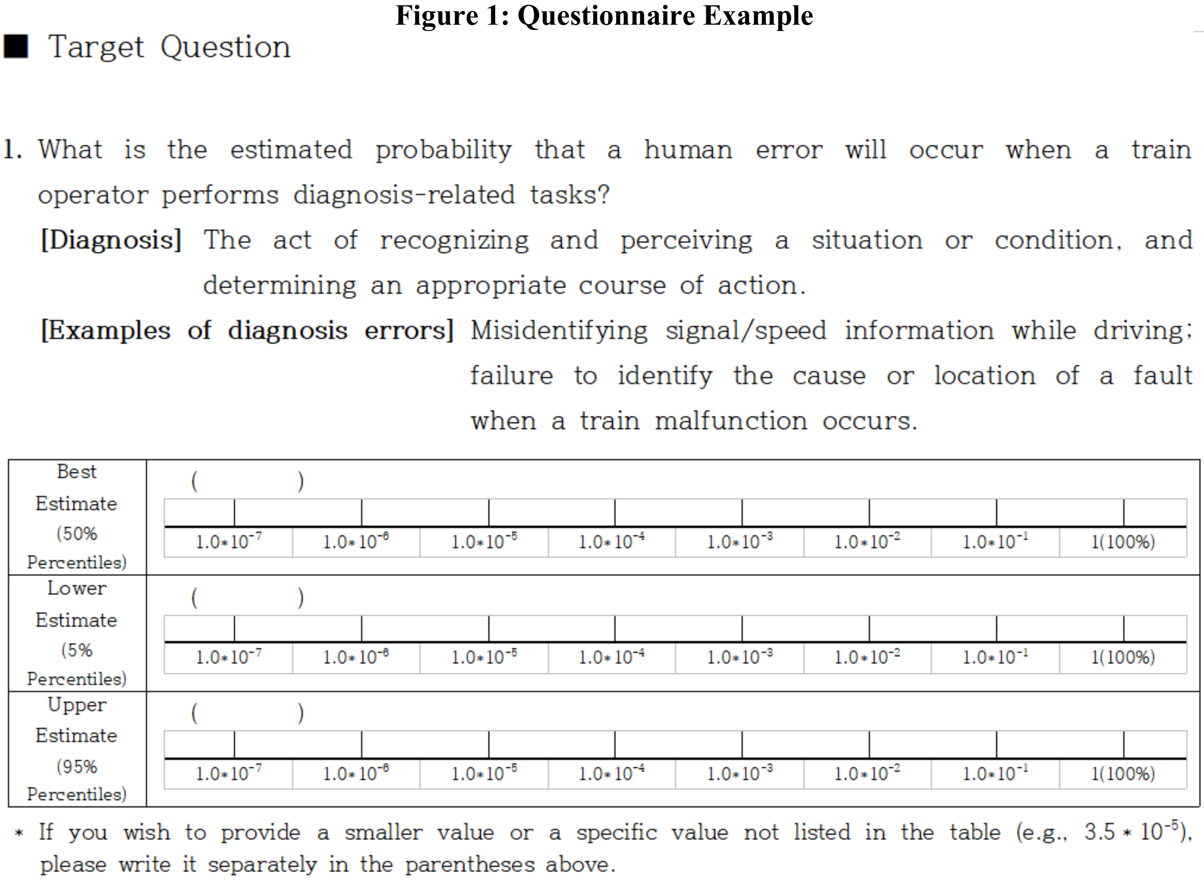
3.3. Survey Respondents

In this study, a panel of 16 domain experts in the railway field was recruited to elicit expert opinions. According to Cooke, an ideal group size ranges between six and twelve experts [18]. It has been reported that with fewer than six experts, the robustness of results may be compromised, while beyond twelve, the marginal benefit of incorporating additional experts diminishes considerably. However, given that the present study targets four distinct occupational types within the railway domain (i.e., train operators, traffic controllers, maintenance workers, and track workers), a panel of twelve experts was deemed insufficient for an accurate assessment of the nominal HEP for each worker type. Therefore, sixteen experts were recruited to ensure sufficient estimation accuracy. To incorporate diverse perspectives, experts were selected from a broad range of professional backgrounds, including academia, research institutions, training agencies, and field operations across the four target occupational types, with experience ranging from a minimum of three years to a maximum of forty-three years, averaging twenty-nine years.

3.4. Conducting the Survey

Based on the developed questions, a structured questionnaire was constructed to systematically consolidate expert opinions, as illustrated in Figure 1. To ensure the reliability and accountability of the elicited responses, the survey was administered through in-person interviews, in which an analyst met individually with each respondent. This face-to-face approach promoted a heightened sense of responsibility among participants and facilitated immediate clarification of any uncertainties encountered during the elicitation process.

Figure 1: Questionnaire Example



■ Target Question

1. What is the estimated probability that a human error will occur when a train operator performs diagnosis-related tasks?

[Diagnosis] The act of recognizing and perceiving a situation or condition, and determining an appropriate course of action.

[Examples of diagnosis errors] Misidentifying signal/speed information while driving; failure to identify the cause or location of a fault when a train malfunctions occurs.

Best Estimate (50% Percentiles)	( )							
	1.0∗10⁻⁷	1.0∗10⁻⁶	1.0∗10⁻⁵	1.0∗10⁻⁴	1.0∗10⁻³	1.0∗10⁻²	1.0∗10⁻¹	1(100%)
Lower Estimate (5% Percentiles)	( )							
	1.0∗10⁻⁷	1.0∗10⁻⁶	1.0∗10⁻⁵	1.0∗10⁻⁴	1.0∗10⁻³	1.0∗10⁻²	1.0∗10⁻¹	1(100%)
Upper Estimate (95% Percentiles)	( )							
	1.0∗10⁻⁷	1.0∗10⁻⁶	1.0∗10⁻⁵	1.0∗10⁻⁴	1.0∗10⁻³	1.0∗10⁻²	1.0∗10⁻¹	1(100%)

∗ If you wish to provide a smaller value or a specific value not listed in the table (e.g., $3.5 * 10^{-5}$), please write it separately in the parentheses above.

Prior to addressing the seed variables and target questions, respondents were provided with a comprehensive overview of the study, an explanation of relevant research ethics, and a concise introduction to the classical model. Guidance was also given on the response methodology — specifically, how to express uncertainty in terms of the 5th, 50th, and 95th percentile estimates — along with an outline of cognitive biases that may potentially influence their responses. Following the provision of background information, including institutional affiliation and years of domain experience, respondents proceeded through the questionnaire in a prescribed sequence, beginning with the seed variables and subsequently addressing the target questions.

4. RESULTS AND DISCUSSION

Each expert's responses were evaluated and aggregated using the classical model upon completion of the survey. To accurately reflect the conceptual and quantitative distinctions between nominal HEP and PSF multiplier, the calculation procedure was divided into two separate processes. Detailed computations were subsequently performed using the "expert package" in R [20]. Table 3 presents the estimated nominal HEP values derived from the expert elicitation. The mean value reported in the table represents a probability-weighted average produced by the DM, generated by applying the performance-based weights derived from the seed variables to the individual expert responses. The 5th, 50th, and 95th percentile values correspond to the respective quantiles of the overall distribution derived by the DM.

Table 3: Results of Expert Elicitation—Nominal HEP

Question	Mean	5%	50%	95%
Train operator - Diagnosis	1.23×10^{-2}	1.00×10^{-3}	1.00×10^{-2}	2.00×10^{-2}
Train operator - action	6.15×10^{-3}	1.00×10^{-3}	2.00×10^{-3}	1.00×10^{-2}
Traffic controller - Diagnosis	5.50×10^{-3}	2.00×10^{-4}	1.00×10^{-3}	1.00×10^{-2}

Traffic controller - action	3.52×10^{-3}	3.00×10^{-4}	1.00×10^{-3}	2.00×10^{-3}
Maintenance worker - Diagnosis	2.79×10^{-3}	1.00×10^{-5}	1.00×10^{-4}	1.00×10^{-3}
Maintenance worker - action	3.47×10^{-3}	1.00×10^{-4}	1.00×10^{-3}	2.00×10^{-3}
Track worker - Diagnosis	1.23×10^{-2}	1.00×10^{-3}	1.00×10^{-2}	2.00×10^{-2}
Track worker - action	3.50×10^{-3}	2.00×10^{-4}	1.00×10^{-3}	2.00×10^{-3}

The nominal HEP values for railway area employees ranged from a minimum of 2.79×10^{-3} (Maintenance worker - Diagnosis) to a maximum of 1.23×10^{-2} (Train operator – Diagnosis and Track worker - Diagnosis). When comparing nominal HEP across diagnosis and action tasks, train operators, traffic controllers, and track workers were assessed to have higher nominal HEP for diagnosis-related tasks than for action-related tasks. Maintenance workers, by contrast, were evaluated to have higher nominal HEP for action-related tasks than for diagnosis-related tasks. With respect to the mean nominal HEP across both diagnosis and action tasks, maintenance workers exhibited the lowest values, followed in ascending order by traffic controllers, track workers, and train operators.

Similarly, Table 4 presents the estimated PSF multiplier values derived from the expert elicitation. The multiplier represents the factor by which the HEP changes at a given qualitative level, assuming a value of 1 at the nominal level.

Table 4: Results of Expert Elicitation—PSF Multiplier

Question	Mean	5%	50%	95%
Workload - Poor	14.450	2	6.441	36.105
Workload – Very Poor	23.424	4	14.068	56.273
Equipment/HMI - Poor	5.654	1.593	3.186	9.186
Equipment/HMI - Good	0.492	0.186	0.444	0.881
Procedures/Guidelines - Poor	8.559	2	5.593	14.068
Procedures/Guidelines - Good	0.292	0.036	0.280	0.525
Experience/Training - Poor	7.229	2	4.407	12.882
Experience/Training - Good	0.421	0.048	0.363	0.881
Work Process - Poor	12.110	2	6.780	28.137
Work Process - Good	0.322	0.0319	0.263	0.681
Complexity - High	4.654	1.593	3.186	9.186
Complexity – Very High	11.873	3.186	5.593	28.137
Communication – Poor	9.559	2	4.407	22.205
Environment - Poor	11.576	2	5.593	28.137

Among the negative levels, Workload exhibited the largest multipliers, with values of 23.424 at the very poor level and 14.450 at the poor level. This was followed, in descending order, by Work Process (Poor, 12.110), Complexity (Very High, 11.873), Environmental (Poor, 11.576), Communication (Poor, 9.559), Procedures/Guidelines (Poor, 8.559), Experience/Training (Poor, 7.229), and Equipment/HMI (Poor, 5.654). At the positive levels, Procedures/Guidelines – Good yielded the greatest reduction in error probability with a multiplier of 0.292, followed by Work Process (Good, 0.322), Experience/Training (Good, 0.421), and Equipment/HMI (Good, 0.492).

5. CONCLUSION

This study quantified the nominal HEP and PSF multipliers for railway employees through a structured expert elicitation. By applying Cooke's classical model, judgments from sixteen experts spanning academia, research institutions, training agencies, and field operations were systematically weighted and aggregated, yielding performance-weighted estimates for both target quantities.

The elicited nominal HEP values exhibited distinct variation across employee type and task type, ranging from 2.79×10^{-3} (maintenance worker – diagnosis) to 1.23×10^{-2} (train operator – diagnosis and track worker – diagnosis). The mean nominal HEP across diagnosis and action tasks was lowest for maintenance workers, followed in ascending order by traffic controllers, track workers, and train operators. For train operators, traffic controllers, and track workers, diagnosis tasks were assessed as more error-prone than action tasks, whereas the reverse tendency was observed for maintenance workers — a pattern that reflects the distinct cognitive and physical demands across employee types.

The derived PSF multipliers consistently increased the HEP at negative levels and decreased it at positive levels, and they varied monotonically with the qualitative level of each PSF, indicating coherent expert recognition of the relative hierarchy among PSF levels. Workload exhibited the largest multiplier among the negative levels, with the HEP increasing by approximately 23-fold at the Very Poor level, whereas Equipment/HMI yielded the smallest negative-level multiplier. The magnitude of variation at the positive levels was substantially smaller than that at the negative levels, suggesting an asymmetric influence in which the deterioration of a PSF exerts a far greater impact on the HEP than its enhancement.

The findings of this study provide foundational quantitative data for the development of a railway-specific HRA methodology. Future work will proceed along two complementary directions. First, the employee classification will be refined into four occupational categories — train operators, traffic controllers, train attendants, and field workers, the last of which encompasses track workers, shunting workers, and electrical facility workers — and a panel of ten experts per category will be recruited to derive more detailed nominal HEPs and PSF multipliers specific to each occupation. Second, building upon these refined estimates, a dedicated HRA methodology for the quantitative assessment of HEP among railway employees will be developed.

Acknowledgements

This work is supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (Grant RS-0023-00239464 and RS-2025-02222955).

References

[1] Ministry of Land, Infrastructure and Transport (MOLIT). *Railway Accident Statistics, 2020–2024.* Republic of Korea: MOLIT, 2025. Available at: https://stat.molit.go.kr
[2] Kyriakidis, M., Pak, K. T., & Majumdar, A. *Railway accidents caused by human error: historic analysis of UK railways, 1945 to 2012*. Transportation Research Record, 2476(1), 126-136, (2015).
[3] Kyriakidis, M., Majumdar, A., & Ochieng, W. Y., *The human performance railway operational index—a novel approach to assess human performance for railway operations*. Reliability Engineering & System Safety, 170, 226-243, (2018).
[4] Byeoung-Soo, Y. U. M., Tae-Yoon, K. I. M., Jong-Uk, W. O. N., Chi-Nyon, K. I. M., & Won-Mo, G. A. L., *A study on the relational analysis of human errors in railway accidents*. Journal of Wellbeing Management and Applied Psychology, 7(1), 35-41, (2024).
[5] Park, M. S., Eom, J. K., Choi, J., & Heo, T. Y., *Analysis of the railway accident-related damages in South Korea*. Applied Sciences, 10(24), 8769, (2020).
[6] Kang, S., & Park, C. W., *Development of an Advanced Framework for Preventing Railway Accidents Caused by Human Error*. Journal of the Korean Society of Safety, 39(4), 93-100, (2024).
[7] Baysari, M. T., McIntosh, A. S., & Wilson, J. R., *Understanding the human factors contribution to railway accidents and incidents in Australia*. Accident Analysis & Prevention, 40(5), 1750-1757, (2008).
[8] Kim, D. S., & Yoon, W. C., *An accident causation model for the railway industry: Application of the model to 80 rail accident investigation reports from the UK*. Safety Science, 60, 57-68, (2013).
[9] Ciani, L., Guidi, G., Patrizi, G., & Galar, D., *Improving human reliability analysis for railway systems using fuzzy logic*. IEEE Access, 9, 128648-128662, (2021).

[10] Swain, A. D., & Guttmann, H. E., *Handbook of human-reliability analysis with emphasis on nuclear power plant applications. Final report* (No. NUREG/CR-1278; SAND-80-0200). Sandia National Labs., Albuquerque, NM. (1983).
[11] Gertman, D., Blackman, H., Marble, J., Byers, J., & Smith, C., NUREG/CR-6883: *The SPAR-H Human Reliability Analysis Method*: Tech. Rep. (2005).
[12] Williams, J. C., *HEART—a proposed method for achieving high reliability in process operation by means of human factors engineering technology*. In Safety and Reliability (Vol. 35, No. 3, pp. 5-25). Taylor & Francis, (2015).
[13] Gibson, W. H., *Railway Action Reliability Assessment User Manual: A technique for the quantification of human error in the rail industry*. RSSB, (2012).
[14] Cooke, R. M., *Validating expert judgment with the classical model*. In Experts and consensus in social science (pp. 191-212). Cham: Springer International Publishing, 2014.
[15] Cooke, R. M., & Goossens, L. L., *TU Delft expert judgment data base*. Reliability Engineering & System Safety, 93(5), 657-674, (2008).
[16] Ter Berg, C. J. A., Leontaris, G., van den Boomen, M., Spaan, M. T. J., & Wolfert, A. R. M. *Expert judgement based maintenance decision support method for structures with a long service-life*. Structure and Infrastructure Engineering, 15(4), 492–503, (2019).
[17] Knol, A. B., Slottje, P., Sluijs, J. P., & Lebret, E. *The use of expert elicitation in environmental health impact assessment: a seven-step procedure*. Environmental Health, 9, pp. 1–19, (2010).
[18] Cooke, R. M., & Probst, K. N. *Highlights of the Expert Judgment Policy Symposium and Technical Workshop*, Resources for the Future, (2006).
[19] Wang, J., B., et al., *Establishment of Risk Analysis and Assessment System for Railway Accidents and Hazards*. Ministry of Land, Infrastructure and Transport, Republic of Korea. (2011).
[20] Goulet, V., Jacques, M., & Pigeon, M. *Expert: Modeling Without Data Using Expert Opinion*. The R Journal, 1(1), 31-37, (2009).