Measuring the impacts of human and organizational factors on human errors in the Dutch construction industry using structured expert judgement

Ren, Xin; Nane, Gabriela F.; Terwel, Karel C.; van Gelder, Pieter H.A.J.M.

DOI
10.1016/j.ress.2024.109959

Publication date
2024

Document Version
Final published version

Published in
Reliability Engineering and System Safety

Reliability Engineering and System Safety 244 (2024) 109959

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress

Measuring the impacts of human and organizational factors on human errors in the Dutch construction industry using structured expert judgement

Xin Ren [a,*], Gabriela F. Nane [b], Karel C. Terwel [c], Pieter H.A.J.M. van Gelder [a]

[a] *Safety and Security Science Group, Faculty of Technology, Policy and Management, Delft University of Technology, Jaffalaan 5, Delft, 2628 BX, The Netherlands*

[b] *Delft Institute of Applied Mathematics, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Mekelweg 4, Delft, 2628 CD, The Netherlands*

[c] *Structural Design and Building Engineering, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Stevinweg 1, Delft, 2628 CN, The Netherlands*

ARTICLE INFO

Keywords:
Structural safety
Human error probability
Human and organizational factors
Classical model
Structured expert judgement

ABSTRACT

This study focuses on measuring the influence of critical Human and Organizational Factors (HOFs) on human error occurrence in structural design and construction tasks within the context of the Dutch construction industry. The primary research question addressed in this paper concerns the extent of HOFs' contribution to human error occurrence. To answer this question, the Classical Model for Structured Expert Judgement (SEJ) is employed, enabling experts to provide their judgments on task Human Error Probability (HEP) influenced by different HOFs, which are subsequently aggregated mathematically. SEJ is chosen as a suitable approach due to the limited availability of applicable data in the construction sector. As a result, the impacts of HOFs are quantified as multipliers, representing the ratio between the observed or evaluated task HEP and its baseline value. These multipliers are then compared with corresponding multipliers from existing Human Reliability Analysis methods and studies. The findings reveal that *fitness-for-duty*, *organizational characteristics* and *fragmentation* exhibit the most pronounced negative effects, whereas *complexity*, *attitude* and *fitness-for-duty* demonstrate the most significant positive impacts on task performance. These results offer valuable insights that can be applied to enhance structural safety assurance practices.

1. Introduction

While structural safety has long been viewed and treated with great importance, it remains a fundamental and critical issue in the construction industry. This is attributed to the often severe consequences in the economic, environmental, and life losses given the occurrence of a structural collapse, even though the possibility of such an event is low. In this regard, continuous research has been performed to assist and guide the engineering practice in the construction industry to prevent structural failures and enhance structural safety. It is acknowledged by numerous studies that the leading cause for structural failures is unintended human error [1–4], rather than technical problems. Furthermore, researchers have pointed out that human errors occurred in the structural design and construction process are most critical and have thus contributed to the largest number of structural defects and failures [1,5,6]. Therefore, it is pivotal to gain a better understanding of how human errors in these two phases come to be so that effective quality assurance measures and safety barriers can be strategically placed to prevent and reduce the occurrence of these errors.

As the understanding of human error grows, a new system approach towards human error has been brought to light [7–9]. In this new view, human error is no longer considered the cause of the system failure; instead, as a symptom of improper system design, organization, or other troublesome issues embedded inside the system. Considering this, the system should be designed in such a way that human errors do not propagate. Moreover, the system approach views humans as an inseparable part of the socio-technical system. Given that, human error is the outcome that arises from the coherent system environment created by local factors like tools and workplace environment, as well as upstream factors such as organizational structure and task design. This system environment contains latent conditions that can turn into error-provoking conditions at a specific time and space, which will result in errors [7]. For example, inappropriate project planning may cause time stress and consequently trigger people to make errors when there is no sufficient time to finish the task with the requirements being fully met.

As pointed out by Elms [10], in order to handle the current structural safety issues, it is important to be aware of the factors that

[*] Correspondence to: PO Box 5015, 2600 GA, Delft, The Netherlands.
E-mail address: x.ren@tudelft.nl (X. Ren).

https://doi.org/10.1016/j.ress.2024.109959

Available online 24 January 2024

lead to increased error proneness. Besides, a pioneering insight of "a fundamental change in viewpoint from a narrower technical focus to a broader systemic approach is in need" was concluded. The system context and the underlying factors, which include the human performance related factors such as physical and mental capabilities and limitations of the personnel at the job, and organization related factors that concern the organizational process and management strategies, which can shape the performance of people at work and potentially lead to the occurrence of human errors and system failures, are defined as the Human and Organizational Factors (HOFs). HOFs are the latent conditions in the building project system that play an important role in structural safety [11]. Unlike other safety-critical industries such as nuclear [12,13], maritime [14], and chemical processing [15,16], which have well adopted the system human error perspective and researched the error-provoking HOFs, the construction industry remains underdeveloped in this regard. As has been suggested by Melchers [17], human error and human intervention have not been studied extensively in the structural reliability theory within the structural safety field.

The approach that has been widely applied in safety-critical industries for human performance assessment in complex systems or processes is Human Reliability Analysis (HRA). It is a set of methods to evaluate human contributions to system reliability and risk by identifying potential human errors, estimating the likelihood of error occurrence, and assessing system degradation as a consequence of human errors [18]. Embodying a combination of qualitative and quantitative methods, HRA aims to provide a better understanding of the latent conditions and context behind errors. This way, designated proactive strategies can be developed to mitigate errors. Additionally, safety barriers can be placed at the root cause to prevent accidents and failures. As a result, the reliability performance of the systems is enhanced.

An important component of HRA methods is the Performance Shaping Factors (PSFs), which represent the system's personal, situational, and environmental characteristics that can affect human performance in a positive or negative manner [19]. HRA methods qualitatively or quantitatively consider the contribution of PSFs to the human error potential and human influence on the system. In a quantitative HRA, PSFs are quantified to measure their impacts on human performance in tasks to provide Human Error Probability (HEP) estimation. HOFs and PSFs are similar constructs given that they both depict the task context for human performance. Thus in this study, they are treated exchangeably. Several existing studies have identified HOFs that influence structural safety in the construction industry, such as [4,20–22]. However, how likely a human error is to occur under the influence of HOFs remains absent knowledge. Therefore, a closer investigation into the effects of HOFs on the task's HEP, and furthermore on the safety of the constructed structures, is in demand for the construction industry.

Thus, this study aims to contribute to quantitatively measuring the impacts of the identified HOFs using collected expert data employing the Classical Model (CM) for Structured Expert Judgment (SEJ). CM [23] is a mathematical model that validates and aggregates individual uncertainty assessments. The research question to be answered by the current study is:

How much do HOFs contribute to the human error occurrence in structural design and construction tasks?

This is the second step towards the development of an HRA method that provides human performance assessment for structural safety in the construction industry. Following its development, this HRA method could be integrated into structural reliability analysis to provide a more comprehensive failure risk assessment by accounting for the human and organizational contributions in the structural design and construction activities to the reliability of the constructed structures. It will take the largely neglected "soft" personnel and managerial component's influence on structural reliability into account when addressing the current human error issue challenge within structural safety from a broader socio-technical systems perspective. As a first step, HOFs recognized to influence structural safety by existing studies are reviewed [24]. Subsequently, 14 HOFs from the review results have been identified as critical for the Dutch construction industry [25]. These critical HOFs (as shown in Fig. 2) will be adopted in this study.

In the following part of this paper, Section 2 elaborates on how the impacts of HOFs are quantified by applying the SEJ method. Consequently, the expert judgement elicitation results are presented in Section 3. Furthermore, the quantified impacts of HOFs are calculated and shown in Section 4. In addition, the quantification results are compared with corresponding PSFs from existing HRA methods, as shown in Section 5. Section 6 justifies the validity of using expert judgement data and SEJ for the purpose of this study, and discusses the choices made for the study design. In the end, Section 7 concludes this study.

2. Methodology

2.1. Measuring the impacts of HOFs on HEP

In a quantitative HRA method, the influence of a PSF on the HEP of a given task is manifested by the degree of alteration in the HEP value resulting from the presence of the PSF, in comparison to the HEP value observed in the absence of the PSF, while keeping all the other task-related variables constant. In the context of a given task, a PSF can exert a detrimental influence on human performance, leading to an elevation in the associated HEP. Such a negative impact can be observed when taking the PSF of *experience and training* accounted in the Standardized Plant Analysis Risk-Human Reliability Analysis (SPAR-H) method as an illustrative example. If the personnel at the job possess a lower than required experience and training level, this can raise the likelihood of erroneous actions to as high as 10 times [26]. However, it is noteworthy that a PSF may also elicit a positive effect on the HEP. For example, in the SPAR-H method, highly experienced and intensively trained personnel tend to lower the probability of erroneous performance by a factor of two [27].

The HEP is the probability that an error will occur in a given task [18]. It is calculated as the proportion of the number of times an error has occurred in the total number of opportunities for an error to occur [28]. Another important concept in the HRA method is the Nominal Human Error Probability (NHEP), which is the probability that human error will occur without the influence of PSFs [18]. It is the baseline probability of human error occurrence in a task and the benchmark for evaluating the potential impact of PSFs on human performance. Thus in this study, the negative or positive effect of a HOF is quantified as a multiplier that increases or decreases the HEP value of a given task based on its baseline NHEP.

Yet such task HEP and NHEP data are scarce [29,30]. There are a few human error databases available in safety-critical industries [31,32], especially in the nuclear industry [33–37]. However, such data are missing in the construction industry. Four data sources for evaluating the impacts of HOFs were discussed by Bea [38]. It was pointed out that expert judgment is an important quantitative information source. Besides, Bea [38] argued for expert judgement data having the "primary and rightful place in making quantitative evaluations", particularly when considering the deficiency of available data when evaluating a specific situation. Therefore, expert judgement data are collected using SEJ to quantitatively measure the impacts of HOFs on task HEP in the Dutch construction industry. In this study, the Absolute Probability Judgement (APJ) is performed to acquire a direct estimation of the HEP value of a given task under the influence of each individual HOF under consideration.

2.2. The classical model for structured expert judgement

Developed by Prof. Cooke [23,39], the Classical Model or Cooke's method for SEJ is a well-known method for aggregating professional judgement from multiple experts for uncertain quantity assessment in situations where objective data are unavailable or incomplete. CM provides a mathematically rigorous, performance-based approach for eliciting and combining subjective uncertainty judgements to reach rational consensus under empirical control. As a sensible and practical method to pool expert knowledge to inform decision-making, CM has experienced broad applications in various fields such as risk assessment [40] for infrastructures [41] and medical device design [42]; environmental science and climate change [43–45]; policy analysis [46]; and more recently, COVID-19 studies [47].

In CM, instead of describing the entire distribution by specifying parameters, experts are expected to estimate several percentiles (e.g., the 5th, 50th and 95th percentile) of the probability distribution for the variable under elicitation. Thus a minimal non-parametric distribution can be derived from their assessments [48]. For empirical validation, two types of questions are elicited: one is the *Calibration question*, or *Seed question*, whose true value is known, or will be known post hoc to the SEJ facilitator, but not known to the experts at the time of elicitation; the *Target question* queries the uncertainty quantification of the target variable. Experts' uncertainty assessments are evaluated by two metrics namely statistical accuracy and informativeness. Statistical accuracy indicates how well the true values are captured by experts' uncertainty assessments. Informativeness intuitively denotes how uncertain experts' assessments are. In the CM, the statistical accuracy is measured by the calibration score that is calculated by comparing the empirical to the theoretical probability vector of the true values relative to experts' percentile assessments, using the Kullback–Leibler (KL) divergence measure. The informativeness is assessed by an information score computed from the expert's percentile assessments relative to a uniform background measure, using the KL divergence measure. The product of these two scores results in a combined score, that, in turn, leads to normalized weights used to aggregate experts' distributions. The ideal expert is both statistically accurate and informative, which leads to a high combined score and normalized weight. It is worth noting that in the CM, the weight is dominated by the calibration score due to the fact that its variation across experts is more significant than that of the information score. Therefore, the input from a statistically highly accurate expert will heavily influence the aggregated result.

The outcome of the CM is a weighted average across the elicited probability distribution of the target variable from all contributing experts, called the Decision Maker (DM). Based on the way the expert judgements are aggregated (performance-based or equal-weighted), there are three main types of DM, namely the Global Weight DM (GL), the Item Weight DM (IT), and the Equal Weight DM (EQ). While the GL averages the expert's information scores of all calibration questions, the IT allows for different weights for different questions for one expert based on the information score of each individual question. The EQ equally involves every expert's contribution to the result. Moreover, there exists an optimized DM for GL and IT, termed GLopt and ITopt, which possesses the highest combined score among GL and IT at any possible significance level (α). The significance level is a cut-off threshold to exclude experts whose calibration scores are smaller than the value of α. In CM, the significance level is often set as 0.05 (i.e., GL0.05 and IT0.05), which coincides with the classical hypothesis testing p-value. For a comprehensive introduction to the CM, the readers are referred to [39,48,49].

There are two existing data analysis tools for the CM. The earlier software is Excalibur, which was developed by the Delft University of Technology in the 1990s [48]. The latest developed tool is Anduryl [50], an open-access Python application for data processing for the CM. In this study, the data analysis is performed using the updated Anduryl version 1.2.1 [51].

2.3. Designing and performing the SEJ

The SEJ was performed by inquiring experts in the Dutch construction industry for the estimated HEP of given tasks. Prior to carrying out SEJ, the human research ethics of this study have been reviewed and approved by the Human Research Ethics Committee (HREC) at the Delft University of Technology. Afterwards, official invitation letters were issued out to 24 experts and 15 responded positively to participation in this study. The questionnaire and the SEJ procedure were tested by two dry runs with two experts. The response from one expert contained wrong data and therefore was discarded. Correct data were filled in and accepted from the other dry run. In the end, the expert judgement data from 14 experts were adopted in the CM analysis. Background information of the 14 responding experts is shown in Fig. 1.

The SEJ was performed individually for each expert. There was no exchange among the experts. Due to the COVID-19 restrictions at the time of this study, all expert elicitation sessions were carried out via scheduled online meetings. Each SEJ session lasted for 1.5 hours, of which the first half an hour provided a project introduction and some background knowledge, as well as showed example questions with answers as training for experts. The SEJ was assisted with a designed online questionnaire for expert uncertainty elicitation. The questionnaire consisted of two sections. The first section contained 10 calibration questions whose answers are known to the researcher of this study from the literature, but not known to the experts. The second section presented the nine target questions that each query the impacts of one factor. An overview of these questions is presented in Table A.3 in the Appendix. The experts could not differentiate between calibration questions and target questions. In this way, the confidence level of an expert is expected to be maintained relatively consistently when providing judgments.

The 14 experts were separated evenly into two panels, with a similar distribution of expert backgrounds in each panel. Experts estimated the $5th$, $95th$, and $50th$ percentile of the HEP value of a specified task under the negative or positive influence of a given factor in this order. The target HOFs were distributed to the two panels, with four overlapping factors judged by both panels, as illustrated in Fig. 2. As a result, the number of questions answered by each expert was reduced from 24 to 19, which greatly eased the mental demand for the experts during the elicitation. In addition, the systematic difference in judgement between the two expert panels can be investigated through the four commonly elicited HOFs. Furthermore, this innovative design enables insights into the robustness of the overall method, which can be observed from the combined scores in Figs. A.9–A.11. Despite the variability in individual expert performance, as captured by the combined score, the DMs' performance remained stable throughout the two panels and for the commonly elicited HOFs.

3. SEJ results

The results of the SEJ study are presented in this section. The performance of experts' assessments is first shown in Section 3.1. Subsequently, the negative or positive impacts of HOFs on the HEP of a given task are elicited through the target questions and the results are presented and discussed in Section 3.2.

3.1. Expert performance revealed by the calibration questions

3.1.1. Experts' and DMs' performance scores

In this study, seven types of DMs have been synthesized based on performance-based weight and equal weight, namely: the GL, the optimized GL (GLopt), the GL with a significance level of 0.05 (GL0.05); the IT, the optimized IT (ITopt), the IT with a significance level of 0.05 (IT0.05); and the EQ. The calibration score, information score, combined score, and normalized weight for each expert under the different DMs are calculated for the Panel 1 experts, the Panel 2

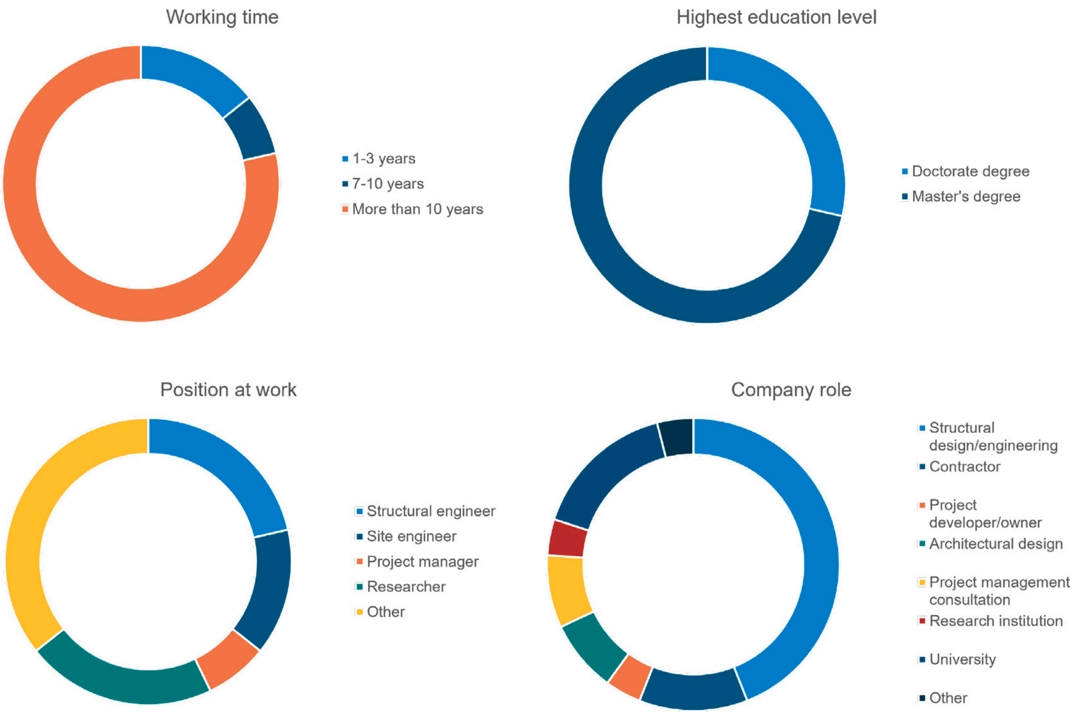


Fig. 1. Background information of the 14 experts.

experts, and all experts of both panels (Panel 1&2). These results are listed in Fig. A.9, Fig. A.10, and Fig. A.11 respectively in Appendix. It can be observed that while the calibration scores vary considerably across experts, the discrepancy in information scores among experts is limited to the same order of magnitude. Therefore, the synthesized DM is dominated by input from the expert with a high calibration score. In this case, since Expert 3 from Panel 1 (EXP1-3) holds the highest calibration score among all experts, the knowledge contributed by EXP1-3 greatly constitutes the DMs of Panel 1 and Panel 1&2. Similarly, the DMs of Panel 2 mainly comprise inputs from EXP2-2 and EXP2-7. Moreover, two types of information scores are calculated to monitor the expert's confidence level variation. One is computed based on the informativeness of the answers provided to the calibration questions, referred to as *Information score-realization*; the other type is calculated based on the answers to all questions, called *Information score-total*. Since the experts do not know whether they are responding to the calibration questions or the target questions, the difference between these two information scores is expected to be slim. Except for expert 2 from Panel 2 (EXP2-2), the values of these two types of information scores are similar, indicating a consistent confidence level of most experts in judging both the calibration variables and the target variables. EXP2-2 exhibits the lowest certainty in answers to the calibration questions, but a noticeably higher confidence level in providing judgements to target questions.

In addition, an overview of the calibration score and the information score of each expert and DM from each panel is shown in Fig. 3. It is evident that there is significant variation in the uncertainty quantification performance of the experts within each panel, both in terms of statistical accuracy and informativeness. Meanwhile, the calibration scores of the DMs exhibit a low level of variance and consistently surpass the scores of individual experts, with the exception of EXP1-3, who attains the highest calibration score among all the experts and non-optimized DMs.

An important observation from these scores is that the ITopt consistently emerges as the best performing DM across all panels. This is supported mathematically, as the DM with the highest combined score is considered the optimal DM. Consequently, the optimized weight DM consistently achieves the highest combined score and is therefore deemed the best DM. Alongside the ITopt, the IT0.05 and IT also exhibit strong performance according to the scores. Generally, the item weight DM tends to outperform the global weight DM and equal weight DM. This can be attributed to the item weight DM's feature to highlight the increased informativeness while preserving the same level of statistical accuracy. Therefore, in the following discourse, the IT, ITopt and IT0.05 will be presented and discussed as the DM for all variables.

3.1.2. Elicited HEP vs. Realization

The "true value", also referred to as the realization of a calibration question is obtained from the survey studies conducted in the Australian construction industry between 1982 to 1993 [52–54]. It is compared with the elicited HEP estimates from this SEJ study for the same question as an empirical measurement for expert performance. The HEP results of the three item weight DMs from the three panels for the 10 calibration questions are depicted in Fig. 4. The results indicated that except for the ITopt from Panel 1 and Panel 1&2 of questions *Q1-9* and *Q1-10*, all the other DMs of the three expert panels were able to capture the realization for every calibration question within the given 90% confidence interval. Except *Q1-4*, the medians of the DMs were found to be relatively close to the realization, indicating the high accuracy performance of the synthesized DMs. It is noteworthy that the experts in Panel 1 provided assessments which resulted in informative and low uncertainty intervals, whereas the experts in Panel 2 showed a lower level of agreement. Overall, when considering both statistical accuracy and informativeness, the best performance was observed in Panels 1&2, which includes responses from all 14 experts to the calibration questions. Furthermore, Fig. A.12 in Appendix shows the performance of each expert as well as the synthesized DMs in response to each calibration question.

3.2. HOFs' influence revealed by the target questions

This subsection presents the outcomes of the target questions that measure the impact of critical HOFs on human performance. The target question evaluates the HEP under the influence of a specified factor, using the NHEP value as a reference point. As an initial estimation,

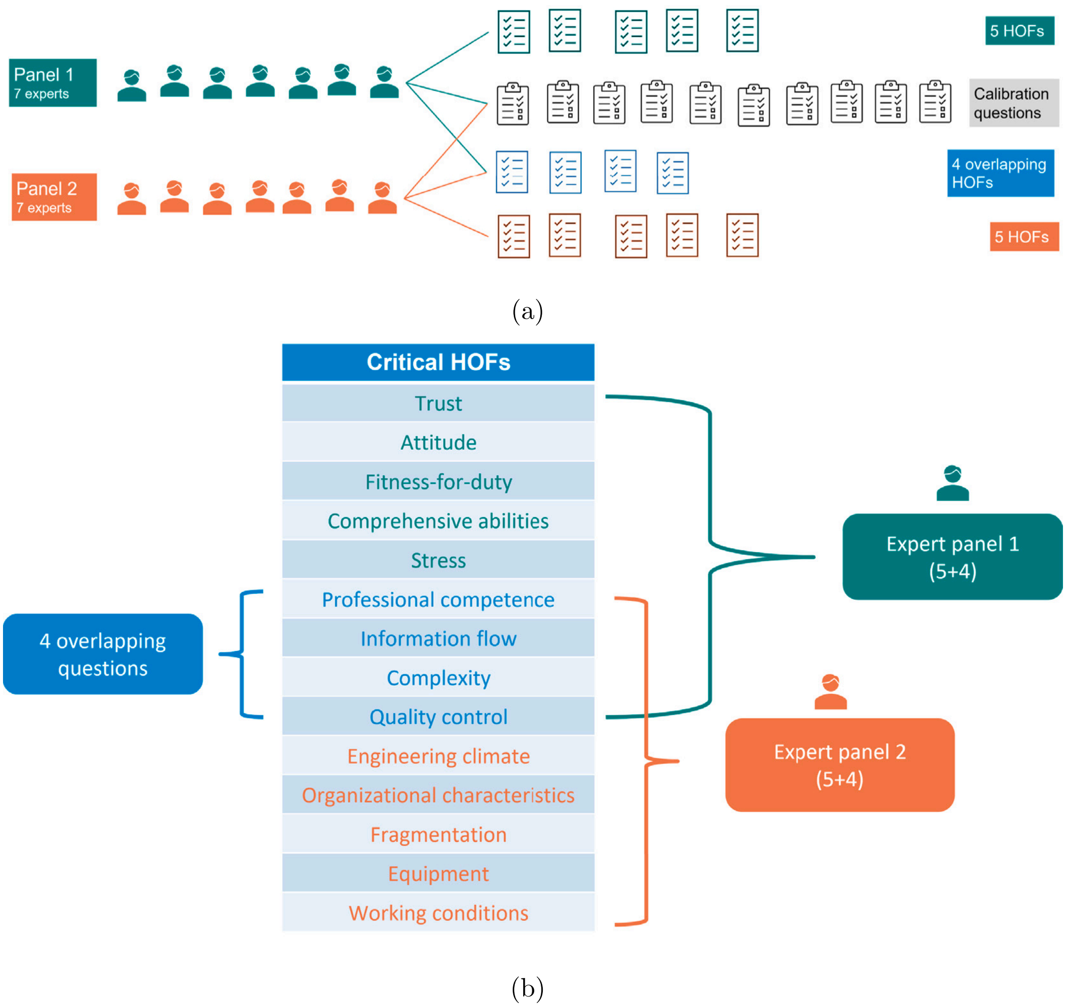


Fig. 2. SEJ design: questions and HOFs assigned to experts. (a) illustrates how the calibration questions and target questions are assigned to the expert panels. (b) shows how the 14 HOFs are distributed to experts for elicitation.

considering the feasibility of this SEJ study, a HOF's influence on various tasks in structural design and construction, such as defining load combinations and placing rebars according to design specifications, is assumed to be the same. Thus the impacts of each HOF are measured against a specified checking task whose NHEP is known from [55] as 1.1×10^{-3}. In addition, the HOFs are presumed to have both negative and positive effects on human performance, resulting in an increase or decrease in the HEP from the baseline NHEP, respectively. For instance, when poor communication, a lack of necessary information, or information overload is present in the task, the *information flow* factor has a negative impact on task performance. Conversely, when timely, effective communication and clear, high-quality information are available, the *information flow* factor poses a positive impact on task performance. The specific meaning of the negative or positive impact of each factor on task performance can be found in the descriptions of the surveyed target questions (TQ) as listed in Table A.3. The target question results have been obtained by aggregating expert judgments based on the CM using item-based weights.

3.2.1. Negative impacts

When a task is performed under the negative impact of a critical HOF, the probability of human error occurrence increases, compared to the NHEP. Fig. 5 illustrates the three item weight DMs of the estimated HEP under the negative effect of the 10 HOFs that were judged by Panel 1 and Panel 2 experts separately. As can be seen from the median values in Fig. 5, the HOFs that have a stronger negative impact on task HEP are *fitness-for-duty*, *organizational characteristics*, and *fragmentation*. In contrast, *working conditions* and *comprehensive abilities* are believed to have the least negative effect on task performance. It can be seen that except for the ITopt of factor *attitude*, the results are consistent among the three DMs in terms of median HEP estimates and the quantified uncertainty. The aggregated median HEP estimates of these 10 HOFs are all located within the range from 5×10^{-3} to 3×10^{-2}. The noticeably larger HEP for *attitude* under the ITopt is due to the inclusion of a single expert's input (EXP1-3) in the ITopt. In terms of the confidence intervals, these 10 HOFs vary considerably. In general, the HOFs assessed by Panel 2 experts exhibit wider ranges than HOFs assessed by Panel 1, except for the factor *attitude* which has the widest uncertainty span under IT and IT0.05. This may be attributed to that the experts find it challenging to judge people's attitudes or to relate an erroneous action to a bad working attitude. On the other hand, the lowest uncertainty can be observed from the factor *comprehensive abilities*, which the experts confidently consider posing a limited negative impact on task performance. While the $5th$ percentiles of the DMs for these 10 HOFs are comparable, the $95th$ percentile HEPs show greater variance.

The corresponding DMs for the negative effect of the four HOFs evaluated by all 14 experts are presented in Fig. 6. It appears that the factor *complexity* has the highest negative impact among these four HOFs on task HEP when taking all experts' judgements into

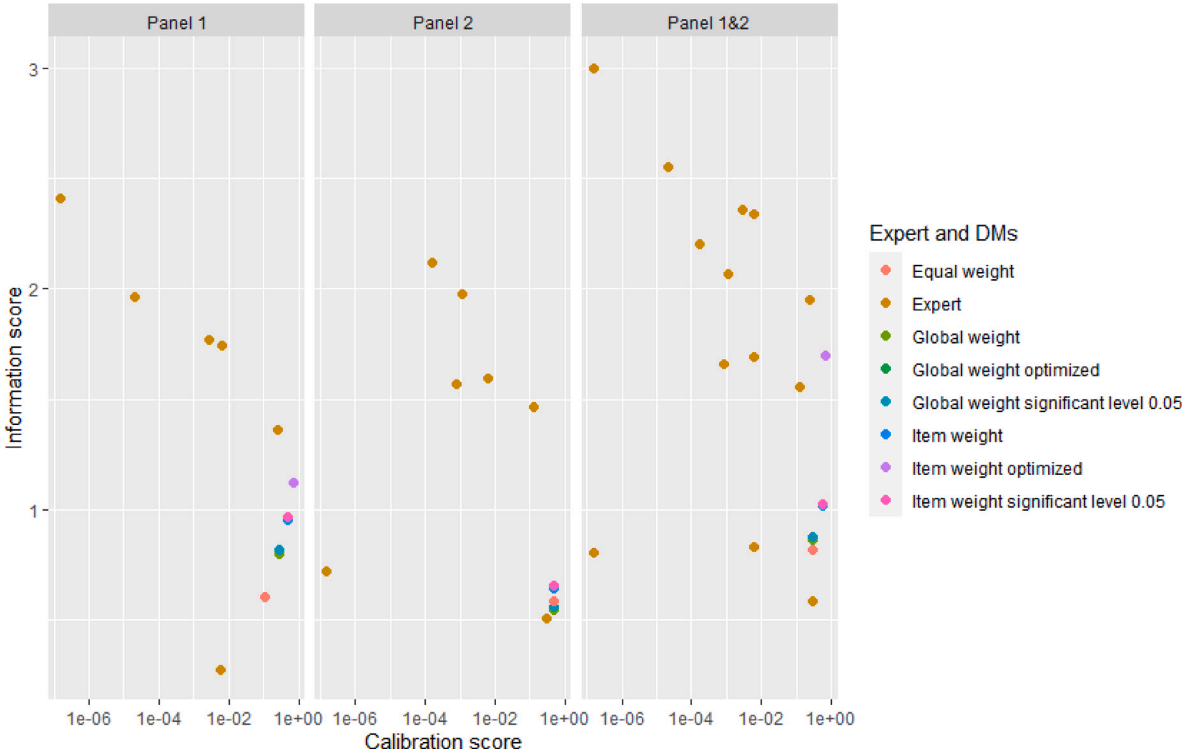


Fig. 3. The calibration score and information score of each expert and DM.

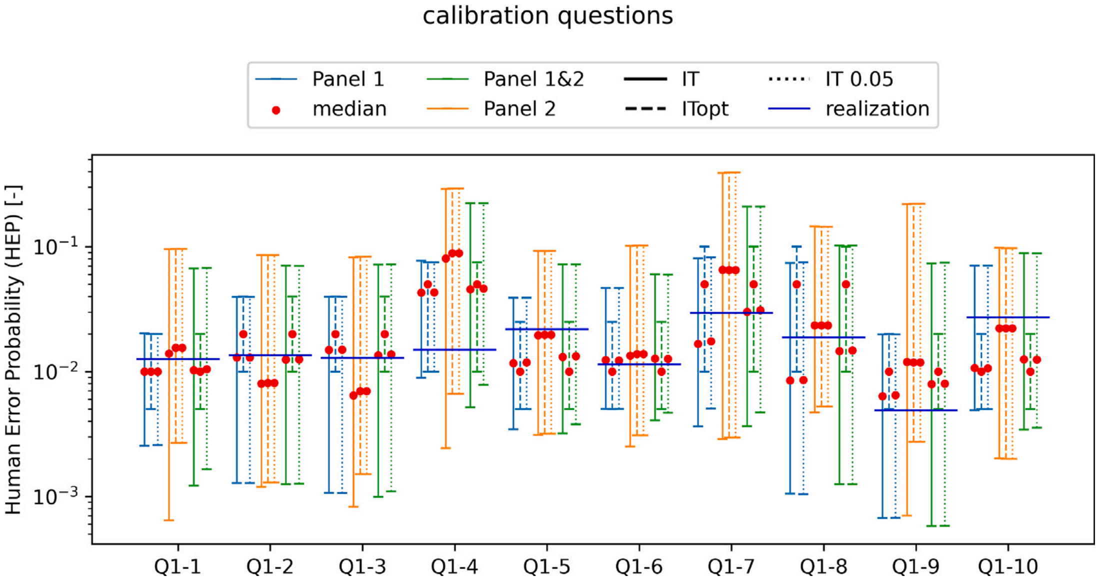


Fig. 4. Comparison between the elicited HEP and the realization of the calibration questions. The HEP estimates range from the 5th percentile (bottom) to the 95th percentile (top), with the median indicated as a red dot. The HEP estimates were sketched using the logarithmic scale.

consideration (DMs of Panel 1&2). Moreover, in the absence of minimum acceptable *professional competence* (95*th* percentile), errors are considered almost certainly will occur. Similar consistent results are noticeable among these four factors: the median HEP estimates for these HOFs are close to 1×10^{-2}, except for the ITopt of factor *professional competence* elicited from Panel 1 and Panel 1&2. This high median HEP is largely contributed by EXP1-3 who holds the highest weight in these two panels and provided a large HEP estimation for *professional competence*. Consequently, there is a noticeable disagreement regarding the negative effects of *professional competence* between Panel 1 experts and Panel 2 experts.

3.2.2. *Positive impacts*

The critical HOFs can also pose a positive impact on the task carrier to enhance performance, leading to a decreased task HEP from the NHEP. The positive effects of HOFs have been inquired through this SEJ study and the results are presented in Figs. 7 and 8 respectively.

In Fig. 7, the median HEP estimates indicate that *attitude* and *fitness-for-duty* can influence the task HEP positively to a large extent, whilst the aggregated assessments of the five HOFs by Panel 2 indicate a mildly positive impact on task performance. What is worth mentioning is the absence in value for the ITopt of factor *stress*. This is due to the fact that the ITopt in Panel 1 only includes the input from EXP1-3, who

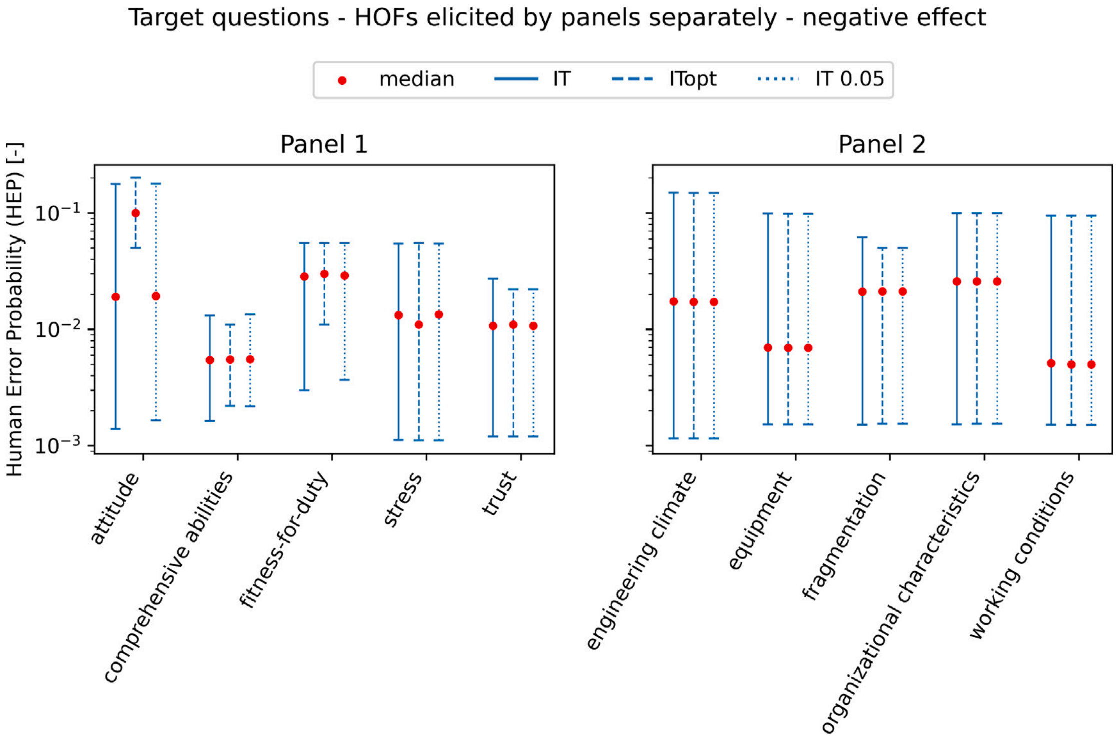


Fig. 5. Three DMs for the HEP estimation under the 10 HOFs' negative effects judged separately by expert Panel 1 and Panel 2. The 90% confidence interval of the IT, ITopt and IT0.05 are illustrated as vertical lines in log scale and the median value is denoted as a red dot.

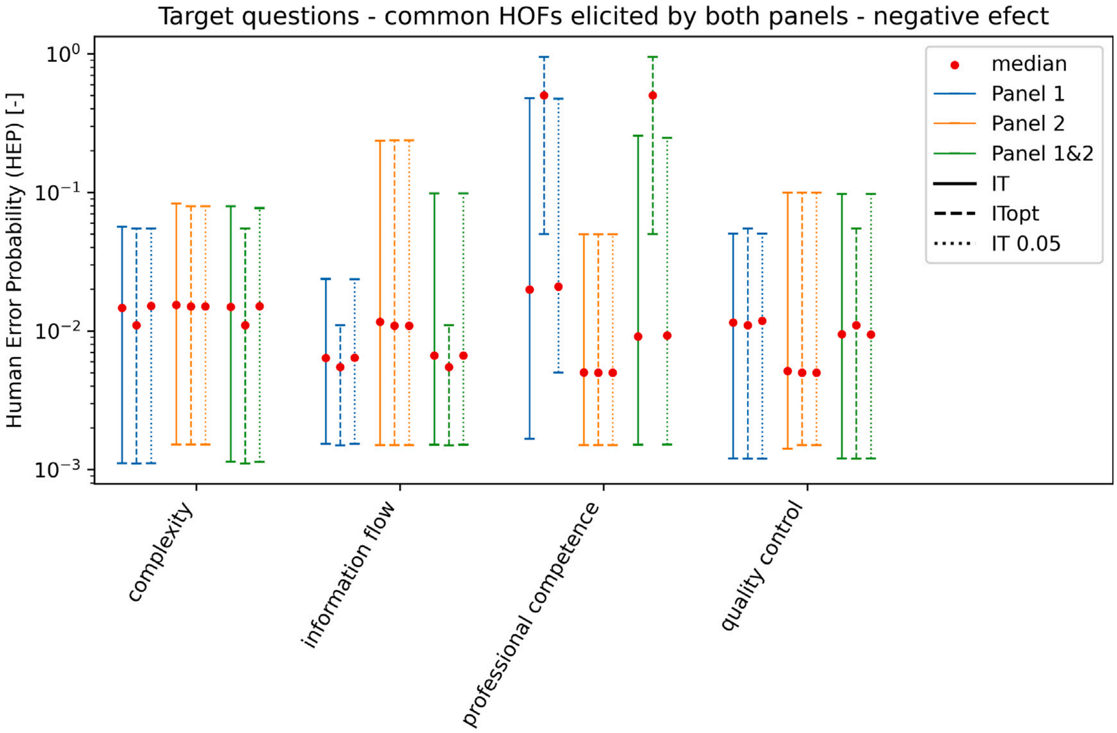


Fig. 6. Three DMs for the HEP estimation under the 4 common HOFs' negative effects judged by both expert panels. The 90% confidence interval of the IT, ITopt and IT0.05 are aggregated for each expert panel and illustrated in differently colored log scale lines. The median of these DMs is denoted as a red dot.

considers *stress* to have no positive influence on task performance and purposely leaves this question out. Additionally, note that while the DMs of the factor *equipment* exhibit the largest uncertainty, the HOFs in Panel 1 show higher uncertainty when compared with the other factors in Panel 2. This indicates that the Panel 2 experts believe that providing ideal tools and equipment for a given task can potentially lead to the largest reduction in HEP (to the $5th$ percentile) under extreme conditions. Finally, it is observed from Fig. 7 that there is less variation

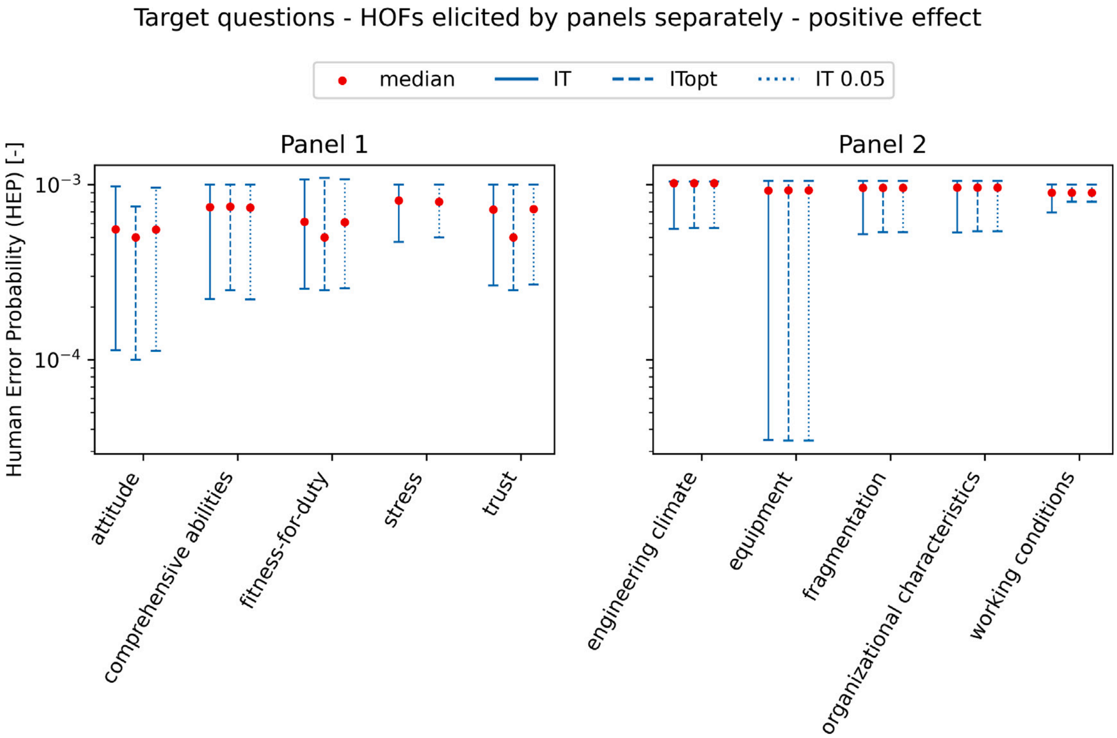


Fig. 7. Three DMs for the HEP estimation under the 10 HOFs' positive effects judged separately by expert Panel 1 and Panel 2. The DMs of these 10 HOFs are depicted in vertical log scale lines with the median value marked as a red dot. The intervals denote the 5th and 95th percentiles of DMs' distribution.

among the DMs' median HEP estimates for HOFs judged by Panel 2 than that of factors assessed by Panel 1. This results from the experts in Panel 2 receiving comparable weight allocation across various DMs.

The results for the positive effects of the four commonly elicited HOFs are demonstrated in Fig. 8. It can be seen from the DMs' median HEPs in Fig. 8 that *complexity* holds the highest positive impact on task performance, whilst the other three factors exhibit comparable positive effects. In addition, there appears to be a clear distinction in the median HEP estimates between the synthesized results from Panel 1 and Panel 2 for each factor. However, these medians all fall within the same order of magnitude at around 7×10^{-4}. Another observation is that the DMs' median estimates of Panel 1&2 are predominantly influenced by the inputs from the Panel 2 experts. Moreover, the 90% confidence interval of DMs from Panel 1 is evidently larger than that of Panel 2, showing less informativeness.

In general, the positive effects of critical HOFs on task HEP appear to be less significant compared to their negative effects. While under HOFs' negative impacts, the task HEP can rise up to 25 times the NHEP, it can only reduce to half of the NHEP under the HOFs' positive effects. Therefore, the results of this SEJ study suggest that great efforts are required to tackle the negative impacts of the HOFs on task performance.

Given the aforementioned findings and discussions, as well as the performance of different DMs, the outcomes obtained through the IT are accepted as the results for the CM. Consequently, the Cumulative Distribution Function (CDF) for the experts' estimates and the aggregated IT under the negative or positive impact of each HOF are presented in Figs. A.13–A.15 in the Appendix. These results are further employed to quantify the impacts of each HOF in the following section.

4. Quantified HOFs for the construction industry

Based on the elicited HEP under the negative and positive effects of the 14 critical HOFs through APJ in this SEJ study and the NHEP of the given task, the negative or positive impacts of HOFs on human error occurrence can be measured via a multiplier quantifying the change in the HEP value relative to the baseline NHEP. Therefore, the multiplier (denoted as M) for each factor can be obtained from the following calculation:

$$M = \frac{HEP_{task}}{NHEP_{task}} \tag{1}$$

In this SEJ study, all HOFs impacts are measured against the same checking task, whose NHEP is 1.1×10^{-3} [55]. Thus, the multipliers for both the negative and the positive effect of each factor, denoted as M_{neg} and M_{pos}, can be derived from the synthesized DM (IT). As a result, the best estimates for M_{neg} and M_{pos}, along with their uncertainties are summarized in Table 1. These multipliers reveal that there are greater variations in both the best estimates (M median) and the uncertainty (90% confidence interval) for the M_{neg} than for the M_{pos}. Moreover, it is interesting to note that *fitness-for-duty* is considered to have the highest negative effect with very low uncertainty, while the experts express strong confidence in the limited negative impact of *comprehensive abilities* on task performance. Furthermore, it is evident that *complexity* has the most substantial positive impact on task HEP, whereas *engineering climate* is regarded as providing minimal positive assistance in diminishing the task HEP.

In this way, the HOFs' impacts on human error occurrence are quantified for the construction industry based on the SEJ from Dutch experts. These multipliers provide essential parametric references for task HEP estimation considering the different influences of HOFs, which enables future human reliability assessment for the construction industry.

5. Comparison of the results

The elicited multipliers for HOFs were further compared with multipliers of corresponding PSFs from existing HRA methods and empirical studies. A total of six widely acknowledged HRA methods [18,56–60] and six PSFs' effects and multipliers studies, which are based on record or simulator data and expert judgment [61–66], have been reviewed. The multiplier value intervals, ranging from positive effects ($M < 1$) to

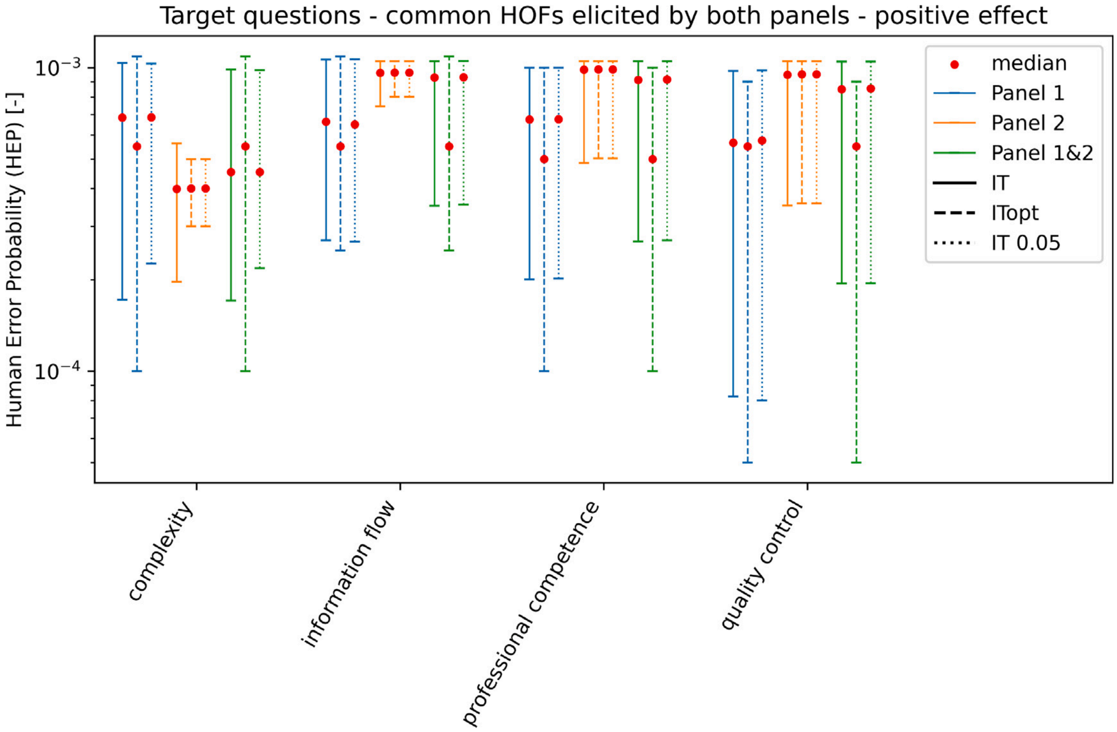


Fig. 8. Three DMs for the HEP estimation under the 4 common HOFs' positive effects judged by both expert panels. The 90% confidence interval of the IT, ITopt and IT0.05 are aggregated for each expert panel and illustrated in differently colored log scale lines. The median of these DMs is denoted as a red dot.

Table 1
The Multipliers of HOFs according to the item weight decision maker (IT). This table shows the $5th$, $50th$ and $95th$ percentiles of the negative effect multiplier's (M_{neg}) distribution and the positive effect multiplier's (M_{neg}) distribution for each HOF.

HOFs	Elicited by	M_{neg}			M_{pos}		
		5th	50th	95th	5th	50th	95th
Professional competence	Panel 1&2	1.37	8.29	232.55	0.24	0.83	0.95
Information flow	Panel 1&2	1.37	6.05	89.24	0.32	0.84	0.95
Complexity	Panel 1&2	1.04	13.55	72.12	0.16	0.41	0.90
Quality control	Panel 1&2	1.09	8.59	88.61	0.18	0.77	0.95
Stress	Panel 1	1.01	12.05	49.53	0.43	0.74	0.91
Fitness-for-duty	Panel 1	2.73	25.92	49.98	0.23	0.56	0.97
Attitude	Panel 1	1.27	17.32	161.73	0.10	0.51	0.89
Trust	Panel 1	1.09	9.76	24.77	0.24	0.66	0.91
Comprehensive abilities	Panel 1	1.48	4.96	12.03	0.20	0.68	0.91
Engineering climate	Panel 2	1.05	15.80	135.91	0.51	0.93	0.95
Fragmentation	Panel 2	1.38	19.16	56.27	0.47	0.87	0.95
Organizational characteristics	Panel 2	1.38	23.50	90.43	0.49	0.87	0.95
Equipment	Panel 2	1.38	6.36	89.81	0.03	0.84	0.95
Working conditions	Panel 2	1.38	4.65	86.09	0.63	0.82	0.91

negative effects ($M > 1$), are summarized in Table 2. When a reviewed method or study does not consider the positive effects of the PSFs, the multiplier range begins from the nominal condition ($M = 1$). In addition, the last column of Table 2 lists the medians of the elicited multipliers for critical HOFs in the construction industry, as derived from this SEJ study. The range is formed from the median values of M_{pos} to that of M_{neg} for each factor from Table 1.

One observation from this review is a lack of agreement among the multipliers of the PSFs used in the 12 existing HRA methods and studies. The main point of difference lies in the M_{neg} of each PSF. As a result, it appears that no consensus has been reached regarding the impacts of PSFs in HRA studies. There are several possible explanations for this variation. One reason is that some of these methods are related to one another or have evolved from one another. Consequently, the multipliers tend to be similar in these related studies. For example, *Improved SPAR-H* [62] and *Petro-HRA* [60] exhibit similar multipliers because they are related to each other. However, most of these HRA methods and studies are independent of each other and thus have varied multipliers for PSFs. Another reason for the difference in PSFs' multipliers is the distinct industrial background within which the method was developed. For instance, the multipliers differ largely between *HEART* [56] and *Marine-specific EPC* [61] due to the different industries for which these PSFs are measured and applied, even though the *Marine-specific EPC* was developed based on *HEART*. In addition, the same PSF might be interpreted differently [67] or be perceived with a distinct level of impact among different industries. Moreover, the contexts of applicability are different for these methods and studies. The impacts of the PSFs on human performance are measured against diverse types of tasks that involve various forms and levels of cognition, different system conditions (e.g., emergency operations), and distinct error modes (error of omission or error of commission) in different methods. For example, *INTENT* [57] was developed for decision-based HEP assessment, whilst most of the other methods and studies target operational errors in tasks. Finally, the data sources for obtaining the multipliers are different. While some studies derive the multiplier values from actual human performance records or simulator data, such

Table 2
Comparison of the HOFs' multipliers with PSFs' multipliers of existing HRA methods and studies.

HOFs	HEART	Marine-specific EPC	CREAM	THERP	SPAR-H	Improved SPAR-H	INTENT	Petro-HRA	Korean nuclear-I	Korean nuclear-II	Korean nuclear-III	China nuclear	HOFs for structural safety
References	[56]	[61]	[58]	[18]	[59]	[62]	[57]	[60]	[63]	[64]	[65]	[66]	This study
Professional competence	2–17	2.88–17	0.5–5	1–2	0.5–10	0.1–50[a]	7–12	0.1–50[a]	1–2.57	1–45.4		0.5–10	0.83–8.29
Trust													0.66–9.76
Attitude	1.2–1.4	2.56–3					6–9						0.51–17.32
Fitness-for-duty	1.02–1.8	1.64–10.3[b]			1–5[a]							1–4	0.56–25.92
Comprehensive abilities	1–6	1–5.29	0.5–5			0.5–50		0.5–50				1–5	0.68–4.96
Information flow	1.2–10	2.64–14.45			0.5–5		5–13			1–12.4		1–5	0.84–6.05
Organizational characteristics	1–1.6	1–1.22	0.8–2										0.87–23.5
Quality control	1.4–5	2.74–12.55	0.5–5	1–50	0.5–50	0.5–50	6–13	0.5–50	0.58–5.53	1–6.3	3–15	1–10	0.77–8.59
Engineering climate	2–2.5	2.15–3.62			0.5–5	1–50	5–23	0.5–50					0.93–15.8
Complexity	1.05–6	2.63–14.45		0.1–5	0.1–5	0.1–50		0.1–50		1–36.7	1.5–10	1–20	0.41–13.55
Stress	1.3–11	1.59–14.01	0.5–5	0.01–10[a]	0.01–10[a]	0.1–50[a]	6–13	0.1–50[a]	0.34–1.24	1–24	2–7.5	0.5–12	0.74–12.05
Fragmentation	1.03–1.06	3.85–4.14	1–5									1–5	0.87–19.16
Equipment	1.4–1.6	4.35–5.69	0.5–5	1–50	0.5–50	0.5–50[a]	6–14	0.5–50[a]	0.39–1			1–10	0.84–6.36
Working conditions	1–1.15	1–9.9	0.8–2			1–10[a]		1–10[a]				1–5	0.82–4.65

[a] The multiplier value is ∞ under extreme condition level (e.g., extremely high negative effect, unfit, inadequate time), leading to the HEP value equal to 1.

[b] The value 0.89 in this range is excluded according to [61].

as the three Korean nuclear studies [63–65], the multipliers from many studies are elicited from expert knowledge judgement, such as the *Improved SPAR-H* [62] and the Chinese nuclear study [68].

In conclusion, the multiplier assigned to the same PSF tends to vary between different HRA methods. The same is true for the present study, where the multipliers of HOFs differ from those found in the reviewed methods and studies. However, certain consistency can be observed in the multiplier ranges of some HOFs elicited in this study and those of the existing HRA methods and studies. These similarities are especially noticeable in factors such as *professional competence*, *comprehensive abilities*, *quality control*, *complexity*, *stress*, and *equipment*. On the other hand, there are notable differences in the multipliers assigned to the factors of *attitude*, *fitness-for-duty*, *organizational characteristics*, and *fragmentation* in construction, when compared to the reviewed methods and studies.

The dissimilar multipliers for *attitude* and *organizational characteristics* in task HEP estimations may stem from their abstract nature and lack of concrete reference points, making it difficult for experts to relate tangible experiences to these two factors. Similarly, the factor of *trust* lacks a multiplier reference in the reviewed methods and studies. It seems that the estimated large negative effect of *fragmentation* reflects the true belief of the experts from the construction industry. *Fragmentation*, though not commonly recognized as a PSF in existing HRA methods, is acknowledged as a crucial factor for causing structural failures in the construction industry [4]. The observed high M_{neg} in the factor of *fitness-for-duty* could be attributed to experts' belief that physical and mental health significantly impact the occurrence of errors. Alternatively, experts may have confused the intention of this factor with the "suitability for task" of personnel, which coincides with *professional competence* and has a relatively high negative effect.

The comparison between the multiplier ranges of HOFs in this study with those of the reviewed methods and studies shows reasonably consistent results on the measured effects of the HOFs on human performance in the construction industry with the multipliers of the PSFs from existing HRA methods and studies. As a side result, this finding provides empirical evidence for the viability of using HRA methods or data from a different field for human reliability assessment in the construction industry.

6. Discussion

6.1. The validity of expert judgement data as scientific data

The ideal human reliability data should be derived from valid experience, records, or robust experiments [69]. However, such data are, in most cases, not available. The reasons for the difficulty of collecting and generating human error data have been detailed in [70]. The scarcity of relevant data for human reliability quantification for the task of interest remains the most significant issue in the human reliability analysis field, as pointed out by Swain [71]. This is particularly the case for the construction industry as a result of the lack of attention and development of HRA in this industry.

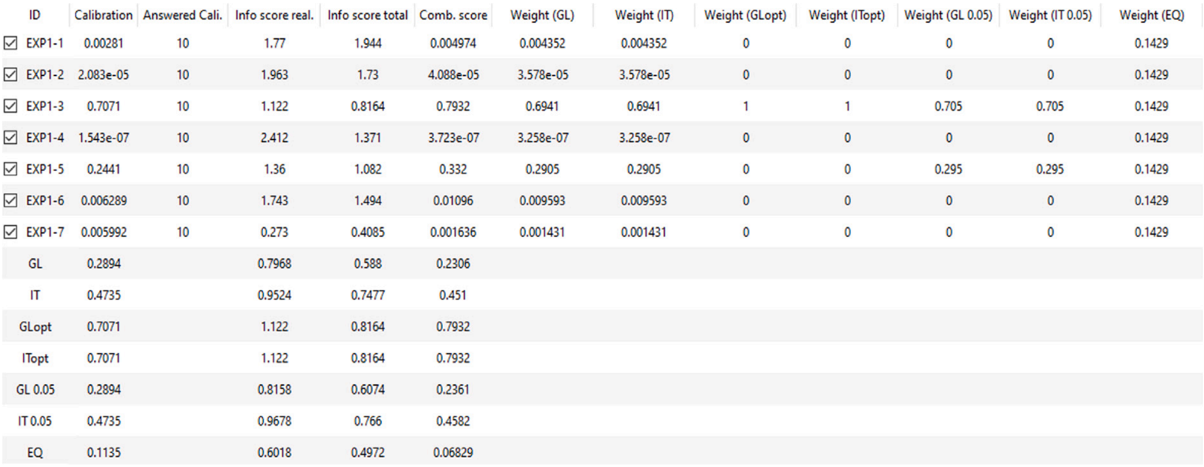
The primary source of uncertainty in all HRA methods is the less-than-adequate data, and it appears to be a challenge that cannot be readily overcome in the immediate future [72]. Even in situations where former data are available, it is questionable if and to what extent such data remain applicable in the context of the specific problem at hand [38,69,73]. Thus, in many circumstances, the only way to proceed with human reliability quantification lies in expert judgement, which is the data source for most, if not all, quantitative HRA methods [69].

Therefore, expert judgement data are used in this study due to the lack of proper HEP data in the construction industry. A structured protocol was chosen to elicit, objectively evaluate and aggregate expert data. Within the CM, expert knowledge is treated as "subjective but scientific" [74]. Cooke [23] proposed four principles (*Scrutability/accountability*, *Empirical control*, *Neutrality*, *Fairness*) to be satisfied by a structured elicitation method. The elicited expert judgement data from an SEJ meet all these requirements and thus can be treated as scientific data [75].

6.2. The concerns and proven benefits of the CM for SEJ

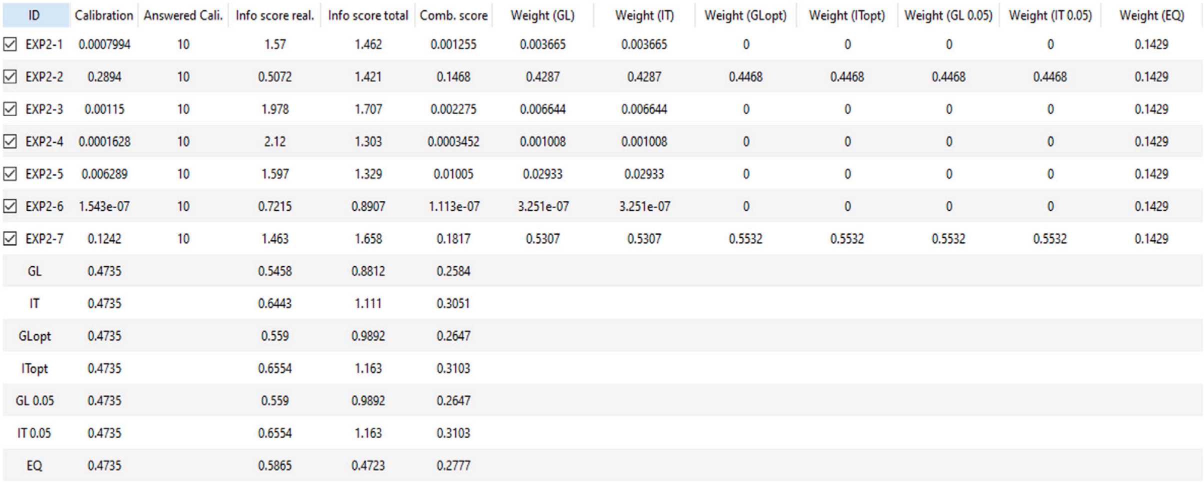
The CM for SEJ is employed to elicit and aggregate expert judgement with uncertainty using performance-based weight in the current study. The standout features of the CM include empirical control with calibration variables and performance-based weighting for combining expert opinions. Certain critiques and concerns have been raised related to these features of the CM. Regarding the effectiveness of calibration questions as empirical control, Hanea et al. [76] believed that it is necessary to assess expert performance in uncertainty quantification and the quality of their judgements in order to treat expert data as scientific data. However, critics question the consistency as a property in expert performance between judging the calibration variables and the target variables [75]. That is, can the expert's performance in answering the target questions be reflected by the performance in answering the calibration questions? Clemen [77] commented that "the decision makers should care about a method's performance on the seed variables only to the extent that it accurately reflects performance on the variables of interest". Hanea et al. [76] suggest from their observations that "prior performance predicts future performance". Furthermore, applying the *Random Expert Hypothesis* to simulate data from 49 SEJ studies, Cooke et al. [78] validated that the variations in expert performance reflect the expert's enduring characteristics rather than random influences during elicitation. This clears the concern that the expert's performance is purely arbitrary and is not a persistent property that propagates beyond the calibration questions.

Another important discussion concerns whether performance-based weighting is better than equal weighting in expert judgement aggregation. In fact, the performance-based weighting of the CM is found



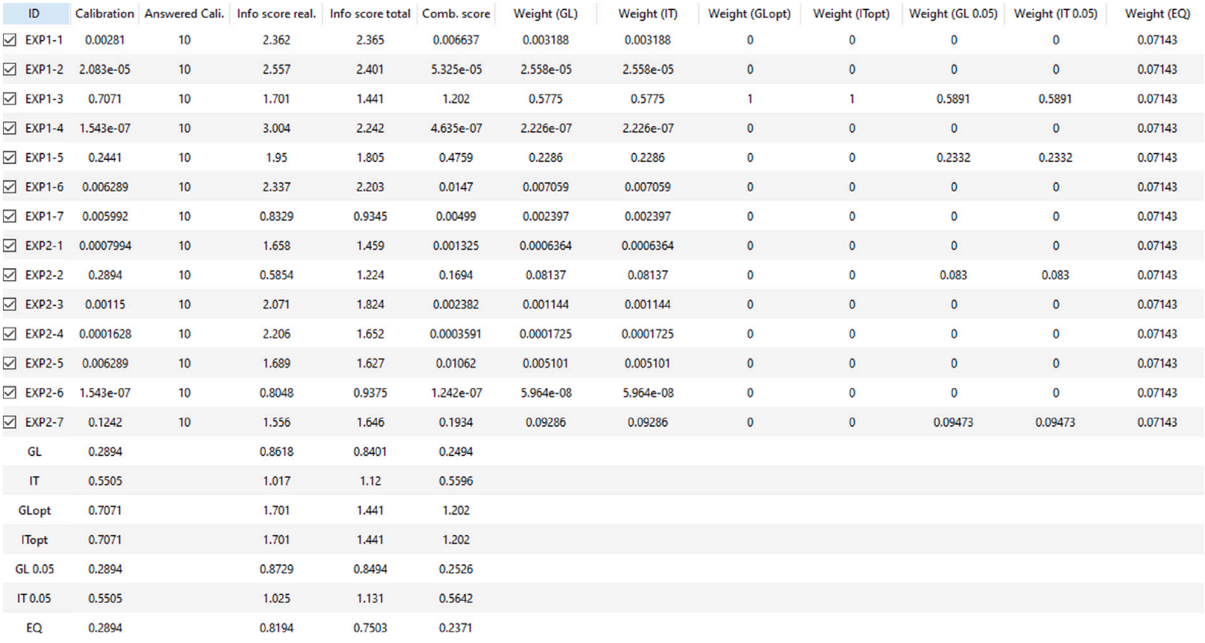
ID	Calibration	Answered Cali.	Info score real.	Info score total	Comb. score	Weight (GL)	Weight (IT)	Weight (GLopt)	Weight (ITopt)	Weight (GL 0.05)	Weight (IT 0.05)	Weight (EQ)
☑ EXP1-1	0.00281	10	1.77	1.944	0.004974	0.004352	0.004352	0	0	0	0	0.1429
☑ EXP1-2	2.083e-05	10	1.963	1.73	4.088e-05	3.578e-05	3.578e-05	0	0	0	0	0.1429
☑ EXP1-3	0.7071	10	1.122	0.8164	0.7932	0.6941	0.6941	1	1	0.705	0.705	0.1429
☑ EXP1-4	1.543e-07	10	2.412	1.371	3.723e-07	3.258e-07	3.258e-07	0	0	0	0	0.1429
☑ EXP1-5	0.2441	10	1.36	1.082	0.332	0.2905	0.2905	0	0	0.295	0.295	0.1429
☑ EXP1-6	0.006289	10	1.743	1.494	0.01096	0.009593	0.009593	0	0	0	0	0.1429
☑ EXP1-7	0.005992	10	0.273	0.4085	0.001636	0.001431	0.001431	0	0	0	0	0.1429
GL	0.2894		0.7968	0.588	0.2306							
IT	0.4735		0.9524	0.7477	0.451							
GLopt	0.7071		1.122	0.8164	0.7932							
ITopt	0.7071		1.122	0.8164	0.7932							
GL 0.05	0.2894		0.8158	0.6074	0.2361							
IT 0.05	0.4735		0.9678	0.766	0.4582							
EQ	0.1135		0.6018	0.4972	0.06829							

Fig. A.9. Scores and weights for Panel 1 experts.



ID	Calibration	Answered Cali.	Info score real.	Info score total	Comb. score	Weight (GL)	Weight (IT)	Weight (GLopt)	Weight (ITopt)	Weight (GL 0.05)	Weight (IT 0.05)	Weight (EQ)
☑ EXP2-1	0.0007994	10	1.57	1.462	0.001255	0.003665	0.003665	0	0	0	0	0.1429
☑ EXP2-2	0.2894	10	0.5072	1.421	0.1468	0.4287	0.4287	0.4468	0.4468	0.4468	0.4468	0.1429
☑ EXP2-3	0.00115	10	1.978	1.707	0.002275	0.006644	0.006644	0	0	0	0	0.1429
☑ EXP2-4	0.0001628	10	2.12	1.303	0.0003452	0.001008	0.001008	0	0	0	0	0.1429
☑ EXP2-5	0.006289	10	1.597	1.329	0.01005	0.02933	0.02933	0	0	0	0	0.1429
☑ EXP2-6	1.543e-07	10	0.7215	0.8907	1.113e-07	3.251e-07	3.251e-07	0	0	0	0	0.1429
☑ EXP2-7	0.1242	10	1.463	1.658	0.1817	0.5307	0.5307	0.5532	0.5532	0.5532	0.5532	0.1429
GL	0.4735		0.5458	0.8812	0.2584							
IT	0.4735		0.6443	1.111	0.3051							
GLopt	0.4735		0.559	0.9892	0.2647							
ITopt	0.4735		0.6554	1.163	0.3103							
GL 0.05	0.4735		0.559	0.9892	0.2647							
IT 0.05	0.4735		0.6554	1.163	0.3103							
EQ	0.4735		0.5865	0.4723	0.2777							

Fig. A.10. Scores and weights for Panel 2 experts.



ID	Calibration	Answered Cali.	Info score real.	Info score total	Comb. score	Weight (GL)	Weight (IT)	Weight (GLopt)	Weight (ITopt)	Weight (GL 0.05)	Weight (IT 0.05)	Weight (EQ)
☑ EXP1-1	0.00281	10	2.362	2.365	0.006637	0.003188	0.003188	0	0	0	0	0.07143
☑ EXP1-2	2.083e-05	10	2.557	2.401	5.325e-05	2.558e-05	2.558e-05	0	0	0	0	0.07143
☑ EXP1-3	0.7071	10	1.701	1.441	1.202	0.5775	0.5775	1	1	0.5891	0.5891	0.07143
☑ EXP1-4	1.543e-07	10	3.004	2.242	4.635e-07	2.226e-07	2.226e-07	0	0	0	0	0.07143
☑ EXP1-5	0.2441	10	1.95	1.805	0.4759	0.2286	0.2286	0	0	0.2332	0.2332	0.07143
☑ EXP1-6	0.006289	10	2.337	2.203	0.0147	0.007059	0.007059	0	0	0	0	0.07143
☑ EXP1-7	0.005992	10	0.8329	0.9345	0.00499	0.002397	0.002397	0	0	0	0	0.07143
☑ EXP2-1	0.0007994	10	1.658	1.459	0.001325	0.0006364	0.0006364	0	0	0	0	0.07143
☑ EXP2-2	0.2894	10	0.5854	1.224	0.1694	0.08137	0.08137	0	0	0.083	0.083	0.07143
☑ EXP2-3	0.00115	10	2.071	1.824	0.002382	0.001144	0.001144	0	0	0	0	0.07143
☑ EXP2-4	0.0001628	10	2.206	1.652	0.0003591	0.0001725	0.0001725	0	0	0	0	0.07143
☑ EXP2-5	0.006289	10	1.689	1.627	0.01062	0.005101	0.005101	0	0	0	0	0.07143
☑ EXP2-6	1.543e-07	10	0.8048	0.9375	1.242e-07	5.964e-08	5.964e-08	0	0	0	0	0.07143
☑ EXP2-7	0.1242	10	1.556	1.646	0.1934	0.09286	0.09286	0	0	0.09473	0.09473	0.07143
GL	0.2894		0.8618	0.8401	0.2494							
IT	0.5505		1.017	1.12	0.5596							
GLopt	0.7071		1.701	1.441	1.202							
ITopt	0.7071		1.701	1.441	1.202							
GL 0.05	0.2894		0.8729	0.8494	0.2526							
IT 0.05	0.5505		1.025	1.131	0.5642							
EQ	0.2894		0.8194	0.7503	0.2371							

Fig. A.11. Scores and weights for all experts.

Table A.3
SEJ questions.

ID	Type[a]	Panel	Question
Q1–1	CQ	Panel 1&2	When the task of deriving a value from a table is performed 100,000 times, how many times contain an error of deriving the wrong value?
Q1–2	CQ	Panel 1&2	When the task of comparing and ranking numbers is performed 100,000 times, how many times contain an error of the wrong order?
Q1–3	CQ	Panel 1&2	When the task of one-step calculation is performed 100,000 times, how many times contain an error of incorrect result?
Q1–4	CQ	Panel 1&2	When the task of interpreting code into design requirements is performed 100,000 times, how many times contain an error of wrong interpretation?
Q1–5	CQ	Panel 1&2	When the task of placing reinforcing bars is performed 100,000 times, how many times contain an error resulting in reduced tensile steel area?
Q1–6	CQ	Panel 1&2	When the task of placing reinforcing bars is performed 100,000 times, how many times contain an error resulting in increased tensile steel area?
Q1–7	CQ	Panel 1&2	When the task of placing reinforcing bars is performed 100,000 times, how many times contain an error resulting in decreased effective depth to tensile steel?
Q1–8	CQ	Panel 1&2	When the task of placing reinforcing bars is performed 100,000 times, how many times contain an error resulting in increased effective depth to tensile steel?
Q1–9	CQ	Panel 1&2	When the task of preparing (configuring, mixing) concrete mix is performed 100,000 times, how many times contain an inadequate mix resulting in reduced concrete compressive strength after 28 days?
Q1–10	CQ	Panel 1&2	When the task of removing framework or shoring is performed 100,000 times, how many times contain an error of premature removal?
Q2–1–2	TQ	Panel 1&2	When lack of or insufficient professional competence is present in this task (negative effect on task performance), how many times will contain an error if the considered task is performed 100,000 times?
Q2–1–3	TQ	Panel 1&2	When above average, excellent professional competence is present in this task (positive effect on task performance), how many times will contain an error if the considered task is performed 100,000 times?
Q2–2–2	TQ	Panel 1&2	When bad communication, necessary information being not available, information overload are present in this task (negative effect on task performance), how many times will contain an error if the considered task is performed 100,000 times?
Q2–2–3	TQ	Panel 1&2	When good and in-time communication, clear and good quality information being available are present in this task (positive effect on task performance), how many times will contain an error if the considered task is performed 100,000 times?
Q2–3–2	TQ	Panel 1&2	When high complexity is present in this task (negative effect on task performance), how many times will contain an error if the considered task is performed 100,000 times?
Q2–3–3	TQ	Panel 1&2	When low complexity is present in this task (positive effect on task performance), how many times will contain an error if the considered task is performed 100,000 times?
Q2–4–2	TQ	Panel 1&2	When lack of or insufficient checking, supervision and procedures are present in this task (negative effect on task performance), how many times will contain an error if the considered task is performed 100,000 times?
Q2–4–3	TQ	Panel 1&2	When checking, supervision and procedures present and in good order are present in this task (positive effect on task performance), how many times will contain an error if the considered task is performed 100,000 times?
Q2–5–2	TQ	Panel 1	When a high workload, tight or insufficient time and budget are present in this task (negative effect on task performance), how many times will contain an error if the considered task is performed 100,000 times?
Q2–5–3	TQ	Panel 1	When a low workload, more than sufficient time and budget are present in this task (positive effect on task performance), how many times will contain an error if the considered task is performed 100,000 times?
Q2–6–2	TQ	Panel 1	When fatigue, unfit, unstable mental/emotional condition are present in this task (negative effect on task performance), how many times will contain an error if the considered task is performed 100,000 times?
Q2–6–3	TQ	Panel 1	When a fit, energetic, clear mind staff is present in this task (positive effect on task performance), how many times will contain an error if the considered task is performed 100,000 times?
Q2–7–2	TQ	Panel 1	When a bad attitude, intentional violation of rules are present in this task (negative effect on task performance), how many times will contain an error if the considered task is performed 100,000 times?
Q2–7–3	TQ	Panel 1	When a very motivated and committed to the job and rules attitude is present in this task (positive effect on task performance), how many times will contain an error if the considered task is performed 100,000 times?

(*continued on next page*)

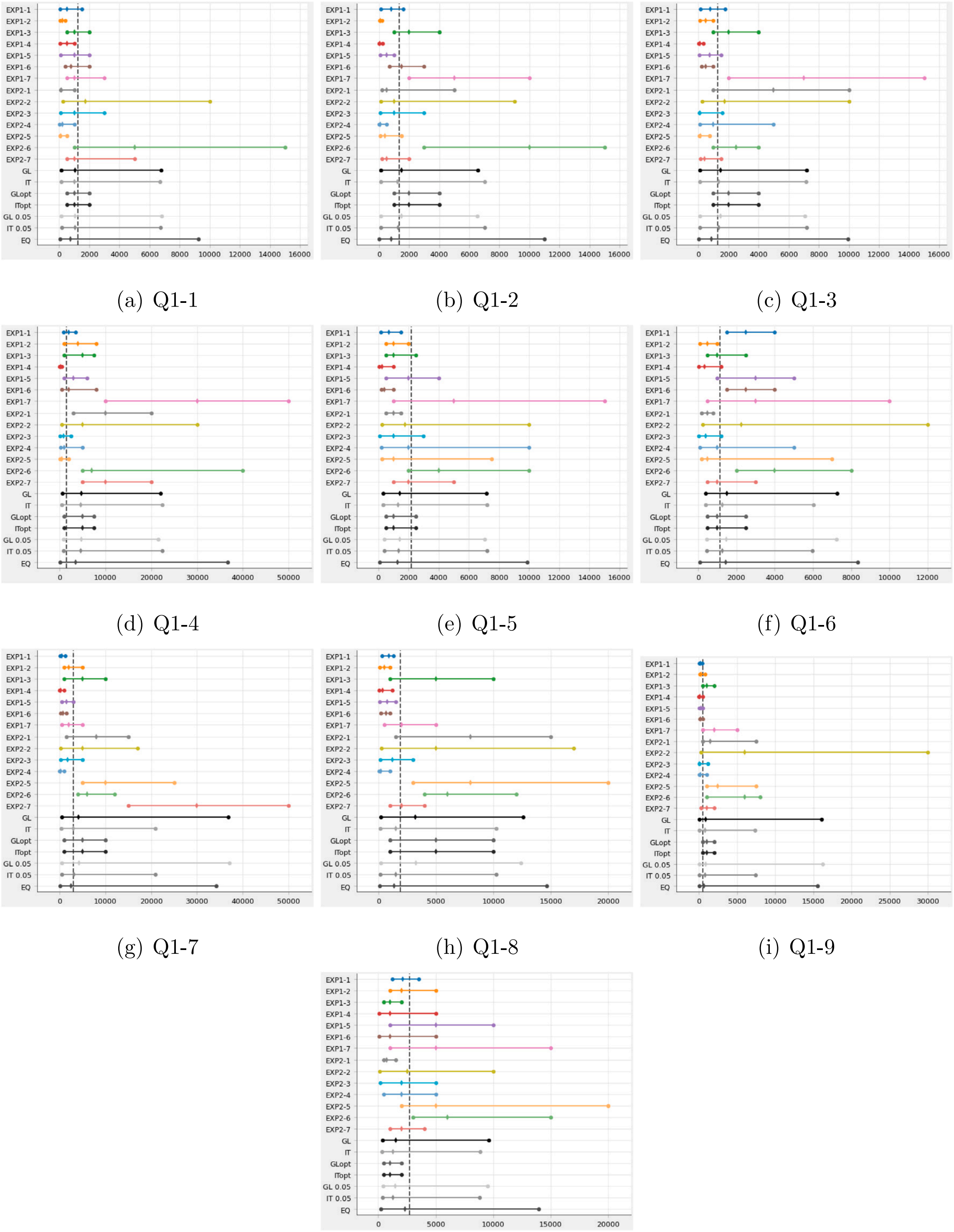
Table A.3 (*continued*).

ID	Type[a]	Panel	Question
Q2–8–2	TQ	Panel 1	When blind trust, overconfidence/over-reliance on others are present in this task (negative effect on task performance), how many times will contain an error if the considered task is performed 100,000 times?
Q2–8–3	TQ	Panel 1	When trusting while still adhering to procedure/verifying is present in this task (positive effect on task performance), how many times will contain an error if the considered task is performed 100,000 times?
Q2–9–2	TQ	Panel 1	When lack of or insufficient comprehensive abilities are present in this task (negative effect on task performance), how many times will contain an error if the considered task is performed 100,000 times?
Q2–9–3	TQ	Panel 1	When above-average or excellent comprehensive abilities are present in this task (positive effect on task performance), how many times will contain an error if the considered task is performed 100,000 times?
Q2–5–2	TQ	Panel 2	When unclear structural safety goals, structural safety goals not put to a prioritized position, underdeveloped safety culture, a safety engineering climate not integrated into daily practice are present in this task (negative effect on task performance), how many times will contain an error if the considered task is performed 100,000 times?
Q2–5–3	TQ	Panel 2	When clear and prioritized structural safety goals, mature safety culture well integrated into practice and keep improved are present in this task (positive effect on task performance), how many times will contain an error if the considered task is performed 100,000 times?
Q2–6–2	TQ	Panel 2	When high fragmentation, frequent personnel change, lack of project overview and network thinking, low planning and coordinating capability are present in this task (negative effect on task performance), how many times will contain an error if the considered task is performed 100,000 times?
Q2–6–3	TQ	Panel 2	When low fragmentation, seldom personnel change, possessing project overview and network thinking, high planning and coordinating capability are present in this task (positive effect on task performance), how many times will contain an error if the considered task is performed 100,000 times?
Q2–7–2	TQ	Panel 2	When chaotic and unstable organization, complex organizational structure, needed support from the parent company not available, redundant team size, confusing allocation of responsibilities are present in this task (negative effect on task performance), how many times will contain an error if the considered task is performed 100,000 times?
Q2–7–3	TQ	Panel 2	When clear, simple and stable organization, available support from the parent company, small and effective team size, clear responsibility allocation are present in this task (positive effect on task performance), how many times will contain an error if the considered task is performed 100,000 times?
Q2–8–2	TQ	Panel 2	When needed equipment being not available or in bad condition (cannot perform as designed), equipment with bad ergonomics or misleading Human–Machine-Interface are present in this task (negative effect on task performance), how many times will contain an error if the considered task is performed 100,000 times?
Q2–8–3	TQ	Panel 2	When the right equipment being available and in good condition, with good ergonomics are present in this task (positive effect on task performance), how many times will contain an error if the considered task is performed 100,000 times?
Q2–9–2	TQ	Panel 2	When bad or disrupting working conditions are present in this task (negative effect on task performance), how many times will contain an error if the considered task is performed 100,000 times?
Q2–9–3	TQ	Panel 2	When very good and promoting working conditions are present in this task (positive effect on task performance), how many times will contain an error if the considered task is performed 100,000 times?

[a] CQ denotes the Calibration question and TQ denotes the Target question.

to outperform equal weighting for most SEJ studies using in-sample data from the TU Delft database [75]. Bolger and Rowe [79] heated up this debate by arguing that when aggregating expert opinions, unequal weighting does not produce any obvious advantages over equal weighting. They reason that on the one hand, it is challenging to develop valid measurements for expert knowledge as the foundation for discriminated weights; on the other hand, the extra cost associated with CM outweighs the gained benefits, if any. In response to these comments, Cooke [80] justified the strength of CM by highlighting that while the mean tends to be of no significant difference, the performance-based weighting leads to improved informativeness in the aggregated result compared with equal weighting. Consequently, cross validation of the CM using data collected from continuously performed SEJ studies in various domains has been carried out using both in-sample and out-of-sample validation. These cross validations concluded the performance superiority of the performance-based weight over equal weight [81,82]. In terms of point prediction, it is found that the aggregated median using performance-based weight outperforms that using equal weight regarding forecast accuracy [78]. Moreover, Marti et al. [83] tested the *Random Expert Hypothesis* with data of 44 post-2006 SEJ studies and verified that the statistical accuracy of real experts is considerably better than simulated random experts. This finding supports the argument that expertise is a persistent property of an expert. Therefore, it is reasonable that experts exhibiting different performances in providing professional judgement should be discriminated against with unequal weights instead of being equally weighted regardless of their distinct performance.

Overall, the CM has been validated over the years in terms of various performance measures. Besides, data from numerous studies revealed an overall superior performance of this method. According to Aspinall [84], the CM for SEJ is “the most effective when data are sparse, unreliable or unobtainable”, which is the case for the current



(a) Q1-1 (b) Q1-2 (c) Q1-3

(d) Q1-4 (e) Q1-5 (f) Q1-6

(g) Q1-7 (h) Q1-8 (i) Q1-9

(j) Q1-10

Fig. A.12. Expert judgements and the aggregated DMs compared with the realizations for the calibration questions. In each sub-figure, the *x*-axis shows how many times contain an error out of the 100,000 repetition of the task; the *y*-axis displays the experts and DMs. The horizontal segment lines exhibit the elicited 90% confidence intervals and the dotes within the segments denote the best estimates. The vertical dash line shows the realization for each calibration question.

study. Additionally, the measured impacts of HOFs from this SEJ study have been compared with the PSF multipliers of existing HRA methods and studies, see Section 5. The results, in return, justified the soundness of the CM for SEJ as the chosen method for this research.

6.3. *The art of selecting experts and calibration variables*

There is no set definition of what constitutes an expert. However, the general expectation of an expert involves mastering abundant

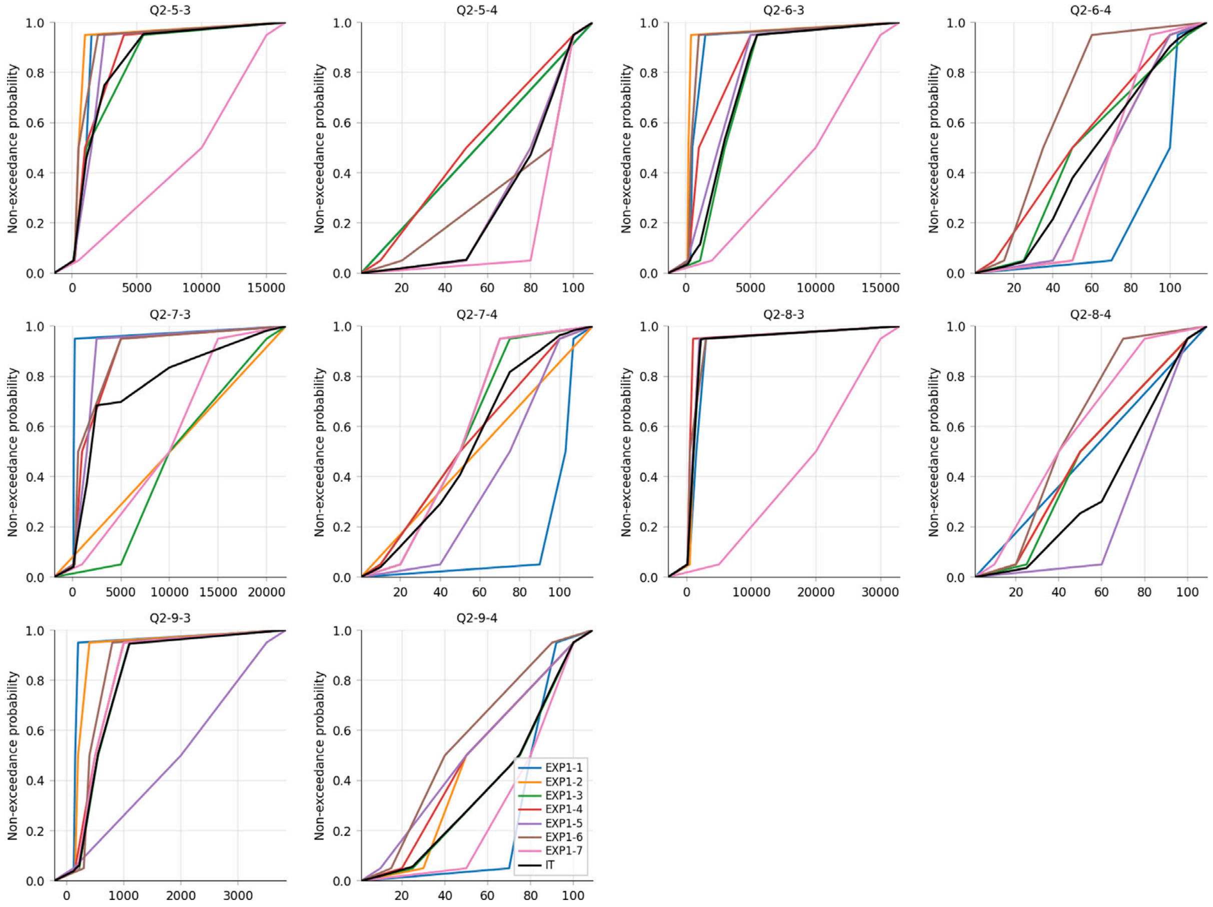


Fig. A.13. The elicited CDFs of the expert's estimates by Panel 1 experts and the aggregated IT.

knowledge and experience in one's domain expertise. In expert elicitation studies, an expert can simply be "the person whose knowledge we wish to elicit", or more sophisticated, "persons to whom society and/or his peers attribute special knowledge about the matters being elicited" [85]. In this SEJ study, 11 out of 14 respondents have more than 10 years of working experience in the Dutch construction industry, as shown in Fig. 1. The remaining three experts comprised a structural engineer holding a doctorate degree and two professionals whose master theses specifically researched the human error issue in the Dutch construction industry. Consequently, despite having less practical expertise, these three were regarded as experts for the purpose of this study due to their extensive expertise on this subject matter. In terms of the number of experts needed for an adequate answer to the target question, Aspinall [84] suggested 8-15 based on his experience, claiming that the results change in an insignificant way with an increased number of experts. Moreover, Quigley et al. [49] pointed out that the common practice with SEJ involves 5–20 experts. In this study, seven experts in each panel elicited five unique HOFs and together 14 experts elicited the four overlapping HOFs and calibration variables. Thus, the number of experts in this SEJ study meets the recommended practice.

The calibration variables are particularly critical to the CM as they form the basis for calibrating the model that is used to aggregate experts' uncertainty assessments. It is essential for the calibration variables to share sufficient similarity with and exhibit a direct link to the target variables, so as to activate similar judgment heuristics [49]. Quigley et al. [49] emphasized that "finding good seed variables is an art". The calibration questions in this study query the HEP of several commonly practised tasks in the structural design and construction process. Whilst the target questions inquire about the HEP of one specific task under the (negative or positive) influence of different HOFs. The true values of the calibration variables are obtained from the available studies [52–54], which makes the calibration variables least desirable since they are both "adjacent" and "retrodiction" [48]. However, the ideal "domain-prediction" type of calibration variable rarely exists in the CM practice [49]. Despite the potential doubts regarding the suitability of these data considering their age and region of origin, they still stand as the best possible calibration variables relevant to the current target questions the authors can find.

There is no definitive number of calibration variables for adequate application of the CM. While Quigley et al. [49] stated 8–20 is the common practice, Hanea and Nane [48] proposed at least 15 when the target variables are less than 35. However, Eggstaff et al. [81] imply that a maximum number of calibration variables may exist beyond which the CM no longer outperform the equal weight linear aggregation. There are 10 calibration variables and nine target variables for each expert panel in this study. Based on the studies used in the analysis by Eggstaff et al. [81], when 10 calibration variables are used in the CM, the performance measure ratio of the performance-based weighting scheme to the equal weighting scheme was assessed to be 1.06. In addition, the combined score of the performance-based weight reached 1.9 times that of the equal weight when there is one more calibration variable than the target variable. The significance in the performance of the performance-based weighting can also be observed

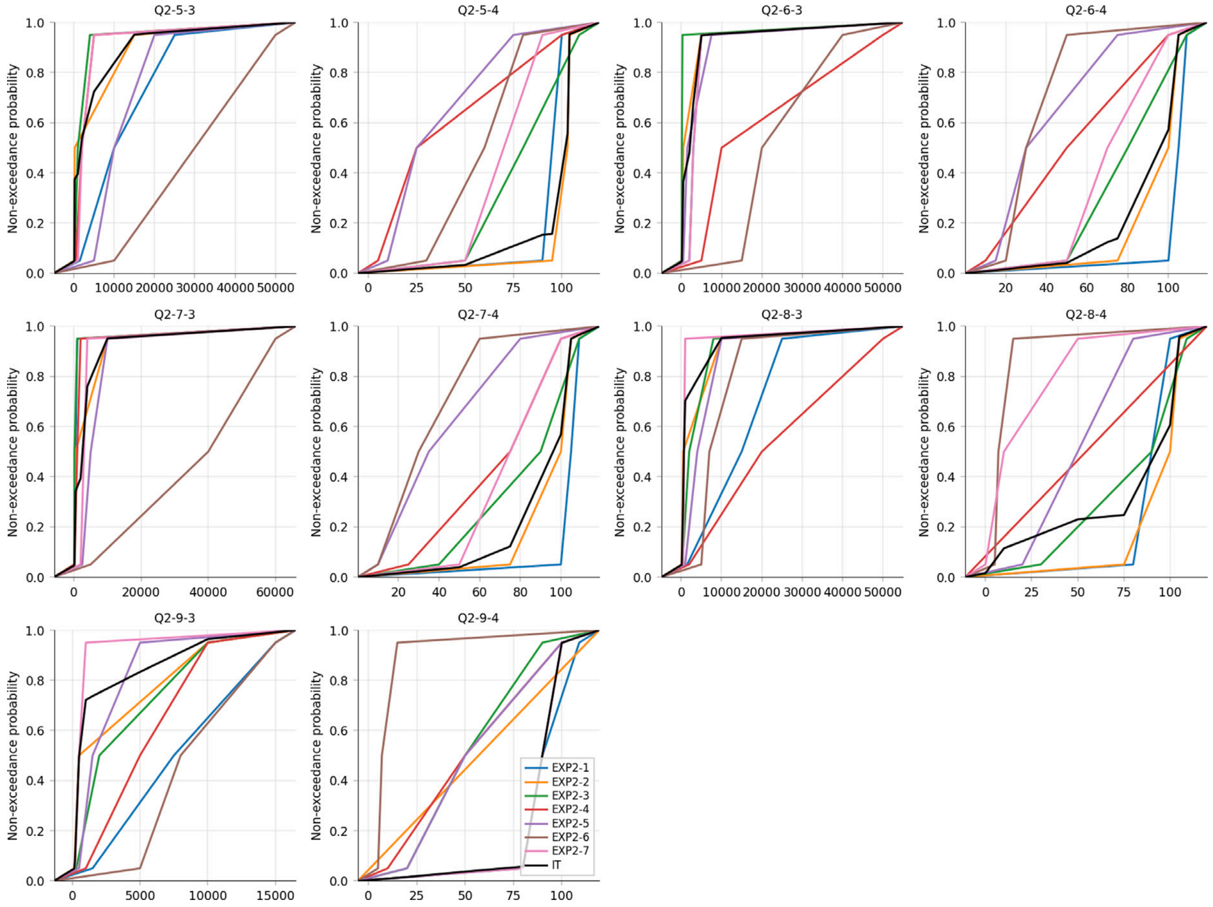


Fig. A.14. The elicited CDFs of the expert's estimates by Panel 2 experts and the aggregated IT.

in Panel 1 and Pane 1&2 in the current study, see Fig. 3. However, while the item weight DMs receive slightly higher combined scores, the performance of global weight DMs is inferior to that of EQ in Panel 2.

7. Conclusion

Human error in structural design and construction plays a major role in structural safety. Recent developments in safety science propose to adopt a socio-technical system view towards the human error issue and research into the task contextual HOFs behind human errors. Therefore, this study measures the impacts of the identified critical HOFs in the Dutch construction industry on human error occurrence, employing the CM for SEJ. Unlike other human reliability quantification studies that predominantly focused on the negative impacts of the HOFs, this study also assessed their positive effects, which has largely been overlooked.

The results of the CM reveal that *fitness-for-duty*, *organizational characteristics* and *fragmentation* are the primary factors associated with the highest negative effects on task performance. Conversely, the factors *complexity*, *attitude* and *fitness-for-duty* demonstrate considerable potential of positive influence to decrease the human error occurrence probability. These results offer valuable insights for industrial practice, highlighting the factors that demand extra attention and quality assurance resources for structural safety. Moreover, the quantified HOFs can serve as initial inputs for the future development of a quantitative HRA method tailored specifically for assessing human reliability for the construction industry. Due to the limitations of this SEJ study, the HOFs' impacts were measured based on a checking task, with an assumption of its relevance to broader structural design and construction tasks. Future research is required to validate this assumption. Moreover, the HOFs' influence ought to be assessed in a more complex setting, considering various task types and error modes. In addition, creating other forms of data sources than expert judgement, such as task record data and experiment data, for HEP estimation in the construction industry, is a worthwhile future endeavour to validate the results of this study.

CRediT authorship contribution statement

Xin Ren: Writing – original draft, Investigation, Funding acquisition, Formal analysis, Data curation. **Gabriela F. Nane:** Writing – review & editing, Supervision, Methodology. **Karel C. Terwel:** Writing – review & editing, Supervision, Conceptualization. **Pieter H.A.J.M. van Gelder:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study is supported by the China Scholarship Council under grant number 2016064340013. The authors would like to acknowledge the significant contributions of all the experts who voluntarily participated in this research.

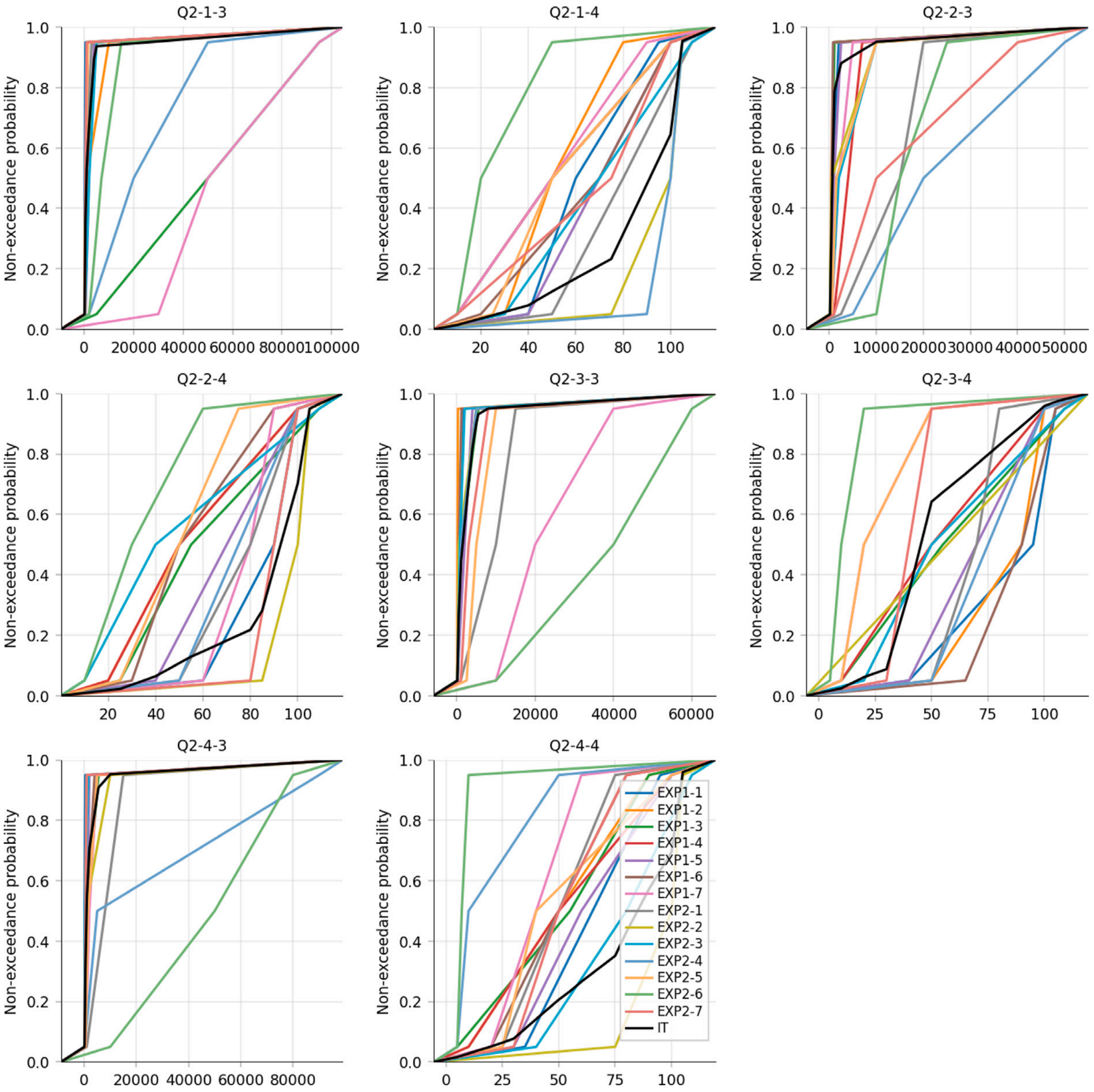


Fig. A.15. The elicited CDFs of the expert's estimates by Panel 1&2 experts and the aggregated IT.

Appendix

See Figs. A.9–A.15 and Table A.3.

References

[1] Ellingwood B. Design and construction error effects on structural reliability. J Struct Eng 1987;113(2):409–22.

[2] Melchers R. Structural reliability theory in the context of structural safety. Civ Eng Environ Syst 2007;24(1):55–69.

[3] Brown C, Elms D, Melchers R. Assessing and achieving structural safety. Proc Inst Civ Eng 2008;161(4):219–30.

[4] Terwel KC, Jansen SJ. Critical factors for structural safety in the design and construction phase. J Perform Constr Facil 2015;29(3):04014068.

[5] Eldukair ZA, Ayyub BM. Analysis of recent US structural and construction failures. J Perform Constr Facil 1991;5(1):57–73.

[6] Terwel K, Boot W, Nelisse M. Structural unsafety revealed by failure databases. Proc Inst Civ Eng 2014;167(1):16–26.

[7] Reason J. Human error: models and management. Bmj 2000;320(7237):768–70.

[8] Dekker SW. Reconstructing human contributions to accidents: the new view on error and performance. J Saf Res 2002;33(3):371–85.

[9] Hollnagel E. Understanding accidents-from root causes to performance variability. In: Proceedings of the IEEE 7th conference on human factors and power plants. IEEE; 2002, p. 1.

[10] Elms D. Structural safety–issues and progress. Prog Struct Eng Mater 2004;6(2):116–26.

[11] Terwel K. Should we focus on human or organizational factors? In: IABSE symposium report, vol. 107, no. 1. International Association for Bridge and Structural Engineering; 2017, p. 1–7.

[12] Liu J, Zou Y, Wang W, Zio E, Yuan C, Wang T, et al. A Bayesian belief network framework for nuclear power plant human reliability analysis accounting for dependencies among performance shaping factors. Reliab Eng Syst Saf 2022;228:108766.

[13] Park J. A framework to determine the holistic multiplier of performance shaping factors in human reliability analysis–An explanatory study. Reliab Eng Syst Saf 2024;242:109727.

[14] Fan S, Yang Z. Towards objective human performance measurement for maritime safety: A new psychophysiological data-driven machine learning method. Reliab Eng Syst Saf 2023;233:109103.

[15] Sezer SI, Akyuz E, Gardoni P. Prediction of human error probability under Evidential Reasoning extended SLIM approach: The case of tank cleaning in chemical tanker. Reliab Eng Syst Saf 2023;109414.

[16] Sezer SI, Camliyurt G, Aydin M, Akyuz E, Gardoni P. A bow-tie extended D-S evidence-HEART modelling for risk analysis of cargo tank cracks on oil/chemical tanker. Reliab Eng Syst Saf 2023;237:109346.

[17] Melchers R. Human intervention and the safety of complex structural systems. Civ Eng Environ Syst 2013;30(3–4):211–20.

[18] Swain AD, Guttmann HE. Handbook of human-reliability analysis with emphasis on nuclear power plant applications. Final report. Tech. rep. NUREG/CR-1278; SAND-80-0200, Albuquerque, NM (USA: Sandia National Labs; 1983.

[19] Di Pasquale V, Iannone R, Miranda S, Riemma S. An overview of human reliability analysis techniques in manufacturing operations. In: Operations management. 2013, p. 221–40.

[20] Atkinson A. Human error in the management of building projects. Const Manag Econ 1998;16(3):339–49.

[21] Bea RG. Human and organization factors: engineering operating safety into offshore structures. Reliab Eng Syst Saf 1998;61(1–2):109–26.

[22] Ren X, Terwel KC, Li J, van Gelder PHAJM. A science mapping review of human and organizational factors in structural reliability. In: Baraldi P, Di Maio F, Zio E, editors. Proceedings of the 30th European safety and reliability conference and the 15th probabilistic safety assessment and management conference. 2020, p. 4724–31.

[23] Cooke RM. Experts in uncertainty: Opinion and subjective probability in science. Oxford University Press on Demand; 1991.

[24] Ren X, Terwel KC, van Gelder PHAJM. Human and organizational factors influencing structural safety: A review. Struct Saf 2024;107:102407.

[25] Ren X, Terwel KC, Yang M, van Gelder PHAJM. 2024. Critical human and organizational factors for structural safety in the dutch construction industry, Unpublished manuscript.

[26] Blackman HS, Gertman DI, Boring RL. Human error quantification using performance shaping factors in the SPAR-H method. In: Proceedings of the human factors and ergonomics society annual meeting, vol. 52, no. 21. SAGE Publications Sage CA: Los Angeles, CA; 2008, p. 1733–7.

[27] Boring RL, Blackman HS. The origins of the SPAR-H method's performance shaping factor multipliers. In: 2007 IEEE 8th human factors and power plants and HPRCT 13th annual meeting. IEEE; 2007, p. 177–84.

[28] Massaiu S, Paltrinieri N. Human reliability analysis: from the nuclear to the petroleum sector. In: Dynamic risk analysis in the chemical and petroleum industry. Elsevier; 2016, p. 171–9.

[29] Park J, Boring RL, Ulrich TA, Lew R, Lee S, Park B, et al. A framework to collect human reliability analysis data for nuclear power plants using a simplified simulator and student operators. Reliab Eng Syst Saf 2022;221:108326.

[30] Podofillini L, Reer B, Dang VN. A traceable process to develop Bayesian networks from scarce data and expert judgment: A human reliability analysis application. Reliab Eng Syst Saf 2023;230:108903.

[31] Kirwan B, Basra G, Taylor-Adams S. CORE-DATA: a computerised human error database for human reliability support. In: Proceedings of the 1997 IEEE sixth conference on human factors and power plants, 1997.'global perspectives of human factors in power generation'. IEEE; 1997, p. 7–12.

[32] Emami KH. Human reliability data banks. Int J Occup Hyg 2019;11(3):232–46.

[33] Chang YJ, Bley D, Criscione L, Kirwan B, Mosleh A, Madary T, et al. The SACADA database for human reliability and human performance. Reliab Eng Syst Saf 2014;125:117–33.

[34] Preischl W, Hellmich M. Human error probabilities from operational experience of German nuclear power plants. Reliab Eng Syst Saf 2013;109:150–9.

[35] Preischl W, Hellmich M. Human error probabilities from operational experience of German nuclear power plants, Part II. Reliab Eng Syst Saf 2016;148:44–56.

[36] Jung W, Park J, Kim Y, Choi SY, Kim S. HuREX–a framework of HRA data collection from simulators in nuclear power plants. Reliab Eng Syst Saf 2020;194:106235.

[37] Yin Z, Li Z, Liu Z, Yang D, Zhang J, Long L, et al. Collection of IDHEAS-based human error probability data for nuclear power plant commissioning through expert elicitation. Ann Nucl Energy 2023;181:109544.

[38] Bea R. Evaluation of human and organization factors in design of marine structures: Approaches & applications. In: Proceedings of the international conference on offshore mechanics and arctic engineering. American Society of Mechanical Engineers; 1995, p. 523.

[39] Cooke RM, Goossens L. Procedures guide for structured expert judgment. Project report to the European Commission, EUR, European Commission; 2000.

[40] Hanea AM, Nane GF, Bedford T, French S. Expert judgement in risk and decision analysis. Cham: Springer International Publishing; 2021.

[41] Rongen G, Morales-Nápoles O, Kok M. Expert judgment-based reliability analysis of the Dutch flood defense system. Reliab Eng Syst Saf 2022;224:108535.

[42] Knisely BM, Levine C, Vaughn-Cooke M, Wagner L-A, Fink JC. Quantifying human performance for heterogeneous user populations using a structured expert elicitation. Saf Sci 2021;143:105435.

[43] Oppenheimer M, Little CM, Cooke RM. Expert judgement and uncertainty quantification for climate change. Nat Clim Chang 2016;6(5):445–51.

[44] Bamber JL, Oppenheimer M, Kopp RE, Aspinall WP, Cooke RM. Ice sheet contributions to future sea-level rise from structured expert judgment. Proc Natl Acad Sci 2019;116(23):11195–200.

[45] Magnan AK, Bell R, Duvat VK, Ford JD, Garschagen M, Haasnoot M, et al. Status of global coastal adaptation. Nature Clim Change 2023;1–9.

[46] Barons MJ, Aspinall W. Anticipated impacts of Brexit scenarios on UK food prices and implications for policies on poverty and health: a structured expert judgement approach. BMJ Open 2020;10(3):e032376.

[47] Colonna KJ, Nane GF, Choma EF, Cooke RM, Evans JS. A retrospective assessment of COVID-19 model performance in the USA. Royal Soc Open Sci 2022;9(10):220021.

[48] Hanea AM, Nane GF. An in-depth perspective on the classical model. In: Hanea AM, Nane GF, Bedford T, French S, editors. Expert judgement in risk and decision analysis. Cham: Springer International Publishing; 2021, p. 225–56.

[49] Quigley J, Colson A, Aspinall W, Cooke RM. Elicitation in the classical model. In: Dias LC, Morton A, Quigley J, editors. Elicitation: The science and art of structuring judgement. Cham: Springer International Publishing; 2018, p. 15–36.

[50] Pieter't Hart CM, Leontaris G, Morales-Nápoles O. Update (1.1) to ANDURIL—A MATLAB toolbox for ANalysis and decisions with UnceRtaInty: Learning from expert judgments: ANDURYL. SoftwareX 2019;10:100295.

[51] Rongen G, Pieter't Hart CM, Leontaris G, Morales-Nápoles O. Update (1.2) to ANDURIL and ANDURYL: Performance improvements and a graphical user interface. SoftwareX 2020;12:100497.

[52] Stewart MG, Melchers RE. Simulation of human error in a design loading task. Struct Saf 1988;5(4):285–97.

[53] Stewart MG, Melchers RE. Error control in member design. Struct Saf 1989;6(1):11–24.

[54] Stewart MG. Modeling human performance in reinforced concrete beam construction. J Constr Eng Manag 1993;119(1):6–22.

[55] Park J, Kim Y, Jung W. Calculating nominal human error probabilities from the operation experience of domestic nuclear power plants. Reliab Eng Syst Saf 2018;170:215–25.

[56] Williams J. A data-based method for assessing and reducing human error to improve operational performance. In: Conference Record for 1988 IEEE fourth conference on human factors and power plants. IEEE; 1988, p. 436–50.

[57] Gertman DI, Blackman HS, Haney LN, Seidler KS, Hahn HA. INTENT: a method for estimating human error probabilities for decisionbased errors. Reliab Eng Syst Saf 1992;35(2):127–36.

[58] Hollnagel E. Cognitive reliability and error analysis method. Elsevier; 1998.

[59] Gertman D, Blackman H, Marble J, Byers J, Smith C. The SPAR-H human reliability analysis method. Tech. rep. NUREG/CR-6883, INL/EXT-05-00509, Idaho, USA: Idaho National Laboratory; 2005.

[60] Taylor C. The Petro-HRA guideline. Tech. rep. IFE/HR/E-2017/001, Institute for Energy Technology; 2017.

[61] Akyuz E, Celik M, Cebi S. A phase of comprehensive research to determine marine-specific EPC values in human error assessment and reduction technique. Saf Sci 2016;87:63–75.

[62] Laumann K, Rasmussen M. 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. Reliab Eng Syst Saf 2016;145:287–300.

[63] Kim AR, Park J, Kim Y, Kim J, Seong PH. Quantification of performance shaping factors (PSFs)' weightings for human reliability analysis (HRA) of low power and shutdown (LPSD) operations. Ann Nucl Energy 2017;101:375–82.

[64] Kim Y, Park J, Jung W, Choi SY, Kim S. Estimating the quantitative relation between PSFs and HEPs from full-scope simulator data. Reliab Eng Syst Saf 2018;173:12–22.

[65] Kim Y, Park J, Presley M. Selecting significant contextual factors and estimating their effects on operator reliability in computer-based control rooms. Reliab Eng Syst Saf 2021;213:107679.

[66] Liu P, Qiu Y, Hu J, Tong J, Zhao J, Li Z. Expert judgments for performance shaping factors' multiplier design in human reliability analysis. Reliab Eng Syst Saf 2020;194:106343.

[67] Paglioni VP, Groth KM. Dependency definitions for quantitative human reliability analysis. Reliab Eng Syst Saf 2022;220:108274.

[68] Liu P, Qiu Y, Hu J, Tong J, Zhao J, Li Z. Expert judgments for performance shaping factors' multiplier design in human reliability analysis. Reliab Eng Syst Saf 2020;194:106343.

[69] Taylor-Adams S, Kirwan B. Human reliability data requirements. Int J Qual Reliab Manag 1995.

[70] Kirwan B, Martin B, Rycraft H, Smith A. Human error data collection and data generation. Int J Qual Reliab Manag 1990.

[71] Swain AD. Human reliability analysis: Need, status, trends and limitations. Reliab Eng Syst Saf 1990;29(3):301–13.

[72] He X, Wang Y, Shen Z, Huang X. A simplified CREAM prospective quantification process and its application. Reliab Eng Syst Saf 2008;93(2):298–306.

[73] Pasman HJ, Rogers WJ. How to treat expert judgment? With certainty it contains uncertainty! J Loss Prev Process Ind 2020;66:104200.

[74] O'Hagan A. Expert knowledge elicitation: subjective but scientific. Amer Statist 2019;73:69–81.

[75] Cooke RM, Goossens LL. TU Delft expert judgment data base. Reliab Eng Syst Saf 2008;93(5):657–74.

[76] Hanea AM, McBride MF, Burgman MA, Wintle BC. The value of performance weights and discussion in aggregated expert judgments. Risk Anal 2018;38(9):1781–94.

[77] Clemen RT. Comment on Cooke's classical method. Reliab Eng Syst Saf 2008;93(5):760–5.

[78] Cooke RM, Marti D, Mazzuchi T. Expert forecasting with and without uncertainty quantification and weighting: What do the data say? Int J Forecast 2021;37(1):378–87.

[79] Bolger F, Rowe G. The aggregation of expert judgment: Do good things come to those who weight? Risk Anal 2015;35(1):5–11.

[80] Cooke RM. The aggregation of expert judgment: do good things come to those who weight? Risk Anal 2015;35(1):12–5.

[81] Eggstaff JW, Mazzuchi TA, Sarkani S. The effect of the number of seed variables on the performance of Cooke's classical model. Reliab Eng Syst Saf 2014;121:72–82.

[82] Colson AR, Cooke RM. Cross validation for the classical model of structured expert judgment. Reliab Eng Syst Saf 2017;163:109–20.

[83] Marti D, Mazzuchi TA, Cooke RM. Are performance weights beneficial? Investigating the random expert hypothesis. In: Hanea AM, Nane GF, Bedford T, French S, editors. Expert judgement in risk and decision analysis. Cham: Springer International Publishing; 2021, p. 53–82.

[84] Aspinall W. A route to more tractable expert advice. Nature 2010;463(7279):294–5.

[85] Garthwaite PH, Kadane JB, O'Hagan A. Statistical methods for eliciting probability distributions. J Amer Statist Assoc 2005;100(470):680–701.