Chapter 3
Are Performance Weights Beneficial? Investigating the Random Expert Hypothesis

Deniz Marti, Thomas A. Mazzuchi, and Roger M. Cooke

Abstract Expert elicitation plays a prominent role in fields where the data are scarce. As consulting multiple experts is critical in expert elicitation practices, combining various expert opinions is an important topic. In the Classical Model, uncertainty distributions for the variables of interest are based on an aggregation of elicited expert percentiles. Aggregation of these expert distributions is accomplished using linear opinion pooling relying on performance-based weights that are assigned to each expert. According to the Classical Model, each expert receives a weight that is a combination of the expert's statistical accuracy and informativeness for a set of questions, the values of which are unknown at the time the elicitation was conducted. The former measures "correspondence with reality," a measure of discrepancy between the observed relative frequencies of seed variables' values falling within the elicited percentile values and the expected probability based on the percentiles specified in the elicitation. The later gauges an expert's ability to concentrate high probability mass in small interquartile intervals. Some critics argue that this performance-based model fails to outperform the models that assign experts equal weights. Their argument implies that any observed difference in expert performance is just due to random fluctuations and is not a persistent property of an expert. Experts should therefore be treated equally and equally weighted. However, if differences in experts' performances are due to random fluctuations, then hypothetical experts created by randomly recombining the experts' assessments should perform statistically as well as the actual experts. This hypothesis is called the *random expert hypothesis*. This hypothesis is investigated using 44 post-2006 professional expert elicitation studies obtained through the TU Delft database. For each study, 1000 hypothetical expert panels are simulated whose elicitations are a random mix of all expert elicitations within that study. Results indicate that actual expert statistical accuracy performance is significantly better than that of randomly created experts. The study does not consider

D. Marti (✉) · T. A. Mazzuchi
The George Washington University, Washington, USA
e-mail: hazanmarti@gwu.edu

R. M. Cooke
Resources for the Future, Washington, USA

A. M. Hanea et al. (eds.), *Expert Judgement in Risk and Decision Analysis*,
International Series in Operations Research & Management Science 293,
https://doi.org/10.1007/978-3-030-46474-5_3

experts' informativeness but still provides strong support for performance-based weighting as in the Classical Model.

3.1 Introduction

Expert elicitation can play a prominent role in the decision-making process in risk assessment, system safety, reliability, and many other fields, particularly in fields where it is difficult to obtain data input (e.g., Einhorn 1974; Cooke and Goossens 2008; Mosleh et al. 1988; Singpurwalla 1988; Spetzler and von Holstein 1975; Wallsten and Budescu 1983; Otway and von Winterfeldt 1992; Aspinall 2010; Chap. 10, this volume). Disciplines that involve high levels of uncertainty combined with insufficient data include, but not are limited to, disaster management, epidemiology, intelligence, public and global health, environment, and security, all of which require robust probabilistic assessments (e.g., Ryan et al. 2012; Keeney and Von Winterfeldt 1989; Hald et al. 2016). In such fields, there might be cost and time considerations, as well as technical impracticalities to data collection, which result in limited scientific data. Sometimes, it is not practical to collect data due to the nature of events. Ultimately, absent or insufficient data lead to poor risk assessments and judgment, resulting in failure either to make informed decisions or to design reliable decision-making processes. Thus, in order to properly characterize the uncertainty in such fields, experts' inputs play a vital role (Cooke and Goossens 2008; Otway and von Winterfeldt 1992). Experts, in the absence of empirical data, are requested to provide information, which could be elicited in various forms such as probability elicitation, parameter estimation, and quantity estimation (Clemen and Winkler 1999). These forms of expert elicitations are essential for uncertainty characterization and risk and policy models.

The standard expert elicitation practice is to consult with multiple experts. Clemen and Winkler (1999) note that the reason for consulting multiple experts is to collect as much data as possible, which could be considered the same as a motivation to increase the sample size in an experiment. This raises the concern on how to fully encapsulate diverse expert judgments in a single input for the analysis. Morgan et al. (1992) noted that factors that lead to combining expert opinions must be chosen so that experts' knowledge can be optimally reflected in the ultimate outcome. Thus, the natural question that arises is "How should one combine multiple opinions?" While a significant body of literature has addressed this issue (see for review Ouchi 2004; Clemen 1989; Morgan et al. 1992), perhaps among the proposed methods, opinion pooling has been the most commonly used approach. Stone (1961) initially coined a strategy for combining opinions: opinion pooling, which was later substantially reviewed by other scholars (see for example, French 1981; Genest and Zidek 1986).

The linear opinion pool is a very practical and straightforward axiomatic method. It is, in fact, a weighted average of multiple probability distributions

$$f(\theta) = \sum_{i=1}^{n} w_i f_i(\theta) \tag{3.1}$$

Here, θ is the unknown quantity of interest, $f_i(\theta)$ is the density function of expert$_i$, w_i represents the weight assigned to expert$_i$, and n is the number of experts. The combined distribution is represented by $f(\theta)$, referred to as the decision maker's probability distribution. Each weight can be interpreted as the expert's relative contribution. If the decision maker has little evidence to judge experts' weights, then each weight is simply distributed equally to the experts, that is $w_i = 1/n$. This approach is called Equal Weighting (EW), and treats each expert equally. However, this approach does not give the decision maker the power to optimize the use of experts' opinions. The underlying assumption of using equal weights is that experts contribute equally. A pre-commitment to EW usually implies that experts' performance will not be measured at all. Consequently, the EW decision maker's performance cannot be validated. This potentially compromises the impact of expert judgment in science-based decision making.

The most prevalent approach addressing this concern is the Classical Model (Cooke 1991), which suggests a weighting mechanism that is based on experts' performances, rather than weighting experts equally. Some scholars argue that performance-based weighting does not outperform equal weighting in terms of the proposed performance criteria. Clemen (2008) provided the most thorough critique of the Classical Model. His results were based on a small sample of expert studies and thus were inconclusive. However, his work advanced the debate and motivated subsequent studies (e.g., Eggstaff et al. 2014; Colson and Cooke 2017), which eventually demonstrated the out-of-sample superiority of the Classical Model's performance-based approach, relative to equal weighting. Following on this work, this chapter seeks to evaluate the appropriateness of the *random expert hypothesis.* This work is novel in the sense of testing the fundamental premises of two aggregation approaches, EW and PW. Simply stated, this hypothesis investigates the claim that any expert's performance in performance-based weighting is due to chance. In this chapter, the random expert hypothesis will be evaluated with respect to the statistical accuracy measures for 44 most up-to-date datasets from the TU Delft database (Cooke and Goossens 2008).

3.1.1 Classical Model

The Classical Model is grounded in the argument that experts differ in terms of their performances—that is, in their ability to assess uncertainty and communicate it properly. Therefore, their performances should be quantified and then reflected in the weighting framework. The model addresses the naturally arising question of how experts' performance can be measured. The model proposes that the Decision Maker's distribution, (1), is obtained via performance-based weights whose values

are determined by the aforementioned measures of experts' statistical accuracy and informativeness. The performances on these two criteria are assessed via an elicitation procedure using predetermined seed variables whose exact values are known by the analyst.

The elicitation procedure involves requesting experts to provide their inputs for a predetermined number, say N, of seed variables, the values of which are usually known post hoc (Cooke 1991). The common practice is to ask experts for their estimates of 5^{th}, 50^{th}, and 95^{th} quantiles for seed variables though other percentile could be used as well. The two performance criteria are measured using these elicitations and the true realizations of the seed variables. The specified percentiles reflect the experts' judgments about this unknown quantity in terms of specified statistical bins. For example, by specifying an elicitation for the 5th percentile, $q_{5,}$ the expert considers that the probability that the true realization of the seed variable is smaller than q_5 is 0.05. Similarly, the 50th percentile, $q_{50,}$ suggests that the expert believes that there is 50% probability of observing the true value to be less than q_{50}, etc.

In addition to these assessed percentiles, the analyst specifies an overshoot percentage (commonly 10%, see Cooke 1991 for more details) in order to determine the complete support for the experts' distributions. Once elicitations are compiled, the analyst assesses the experts' performances using the true realizations of the seed variables. Specifically, the analyst determines how informative the expert distributions are relative to a minimally informative distribution on the support and how well the expert's uncertainty assessments via the specified percentile values match with the realization of the seed variables (i.e., statistical accuracy).

(1) Informativeness

Informativeness score gauges the additional contribution of the expert's elicitation relative to a background measure. That is, it answers the question of "does the expert provide any additional information than a minimally informative distribution?" To measure experts' performance with respect to this criterion, the analyst first combines the expert opinions for each seed variable into a single range, the lower and upper bounds of which are determined by, respectively, the minimum and the maximum of elicited values for each seed variable and the realization of these variables. Then, by using a 10% overshoot percentage, the entire cumulative distributions are computed for each expert. These elicited distributions for each expert are compared with a minimally informative background measure, usually the uniform distribution, which expresses complete uncertainty over the range. The more additional information an expert's distribution gives relative to the base knowledge, the higher the *information* scores he or she would receive.

(2) Statistical Accuracy

Statistical accuracy (a.k.a., calibration score) is a measure of the extent to which the expert's quantile assessment matches with reality. Cooke (1991) incorporated this idea into the model by using a hypothesis test. The null hypothesis is that the experts' percentile assessments correspond to reality. The p value associated with

this hypothesis constitutes the statistical accuracy score. That is, lower p value indicates less evidence about the experts' statistical accuracy performance. Following the computations below, the analyst determines the frequency of true realizations' occurrence in specified inter-quantile intervals, bounded by the specified quantiles.

3.1.2 The Debate on Aggregating Expert Elicitations Mechanisms: Performance-Based Weights (PW) Versus Equal Weights (EW)

The debate around how to aggregate expert elicitations revolve around two fundamental approaches: combining expert elicitations based on equal weights (EW) or based on their performance-based weights (PW). The Classical Model uses a performance-based approach. The model's main premise suggests that performance-based weighting mechanism ensures higher quality and improves the task for which the expert elicitation is done. There is a substantial body of knowledge that supports the use of performance weights (e.g., Aspinall et al. 2016; Bamber and Aspinall 2013; Colson and Cooke 2017; Wilson 2017). However, others have advocated the use of equal weights to combine expert elicitations (Clemen and Winkler 1999; Clemen 1989). They argued that equal weights perform as well as performance-based weights; therefore, there is no need to undertake an intensive expert elicitation procedure (e.g., Clemen 2008). Some of these critics failed to provide substantial evidence and details of their research procedure (e.g., replicable codes), so their findings are not considered to be conclusive.

Perhaps, among the EW advocates, the most productive contribution was Clemen (2008) who critiqued the Classical Model implementations for solely depending on in-sample validation. He argued that the concern about this validation technique was that it uses the dataset to determine the performances and also to validate the model. He suggested using out-of-sample validation and compared EW and PW. Specifically, Clemen (2008) performed a remove-one-at-a-time (ROAT) method, whereby seed variables are removed one at a time. Performance weights are computed based on the remaining seed variables, and these weights are used to predict the removed item. He found that PW failed to statistically outperform EW (PW outperformed EW in 9 out of 14 studies). Two concerns were raised about these findings: One, Clemen (2008) used a nonrandom sample and failed to justify his data choices. Second, the ROAT approach leads to systematic biases, whereby each removed item can penalize an expert who did poorly on that particular item (Cooke and Goossens 2008; Colson and Cooke 2017). This bias was addressed (Colson and Cooke 2017) by a more substantial approach, the cross-validation technique that uses a certain percentage of dataset, instead of a single seed variable. The dataset is split into a training set to determine the performance weights and a test set to predict the removed items. Eggstaff et al. (2014) performed an extensive cross-validation analysis on all possible sets of training and test variables and found that PW statistically outperforms EW.

These examples of previous studies confirmed the validity of the Classical Model; nonetheless, the debate continues. The model has been validated in studies that include different number of seed variables and experts (e.g., Tyshenko et al. 2011; Jaiswal et al. 2012; Bamber et al. 2016; Aspinall 2010; Aspinall et al. 2016). The debate so far focused on the validity of the Classical Model, in different validation approaches (i.e., in-sample, ROAT, and cross validation). However, it is also necessary to analyze the fundamental assumptions of the two competing approaches. No previous studies have tested the core distinction between the two camps of the debate: do the differences in performance reflect persistent differences in the experts, or are they an artifact caused by random influences introduced by the elicitation itself? For example, if the difference is due to one expert having a good day, or being influenced by domestic or professional stressors, or having more information about particular seed variable, etc., then the equal weighting scheme may be warranted. The EW approach assumes that any apparent differences in expert performance are due to such random influences and would not persist beyond the particular elicitation context. On the other hand, the PW approach suggests that performance differences reflect "properties of the experts," which persist beyond particular elicitation context. Focusing on the fundamental assumption that performance differences are persistent enables the formulation of this assumption as a testable statistical hypothesis termed the *Random Expert Hypothesis (REH)*: apparent differences in expert performance are due to random stressors affecting the elicitation.

3.2 Random Expert Hypothesis (REH)

The REH states that apparent differences in expert performance are due to random stressors of the elicitation. If this hypothesis were true, then randomly reallocating the assessments among the experts should have no effect on the performance of the expert panel. This "random scrambling" is precisely defined below. Under the REH, the scores of the best and worst performing experts in the original panel should be statistically indistinguishable from those of the best and worst experts after scrambling the assessments. The variation in expert scores in the original panel should be statistically indistinguishable from the variation in the scrambled panels. There are many ways of scrambling the experts' assessments and this allows a determination of the distributions of scores that result from randomly redistributing the stressors over the experts.

Note that random scrambling will have no effect on the EW combination. This underscores the fact that EW *implies* the REH. In consequence (modus tollens), if REH is (statistically) rejected, then so is EW. In this sense, REH provides a more powerful test of the assumption underlying the use of EW. Note also that if all experts in a panel are "equally good" or "equally bad," then the REH may actually be true for that panel. Indeed, this sometimes happens. The use of PW depends on the fact that such panels are in the minority. Testing the REH on a set of cases allows for gauging the size of that minority.

The REH was tested by a process of creating random panels of experts whose elicitations are derived from the experts within the original expert panel. For example, suppose an expert judgment panel includes ten experts, each of whom assessed 5th, 50th, and 95th percentiles for each seed variable. A hypothetical expert judgment panel would have ten randomly created experts, each of whose elicitations are randomly drawn without replacement from the original assessments for each variable. This process is repeated 1000 times. If there is not a systematic difference between randomly created experts and the original experts, as the REH implies, then one would expect that in approximately half of those 1000 runs, the original experts would outperform the random experts. The performance measure used in this study is statistical accuracy; informativeness and full performance weights will be considered in a future study.

Figure 3.1 displays the process of random expert creation for three experts and three seed variables. For example, Random Expert 1 takes the assessment of Original Expert 2 for Seed Variable #1, the assessment of Original Expert 1 for Seed Variable #2, and finally the assessment of Original Expert 3 for Seed Variable #3. Random Expert 2 chooses randomly from the remaining experts, and Random Expert 3 gets the remaining elicitations. Ultimately, a hypothetical expert judgment panel is composed by creating as many scrambled random experts as in the original experts.

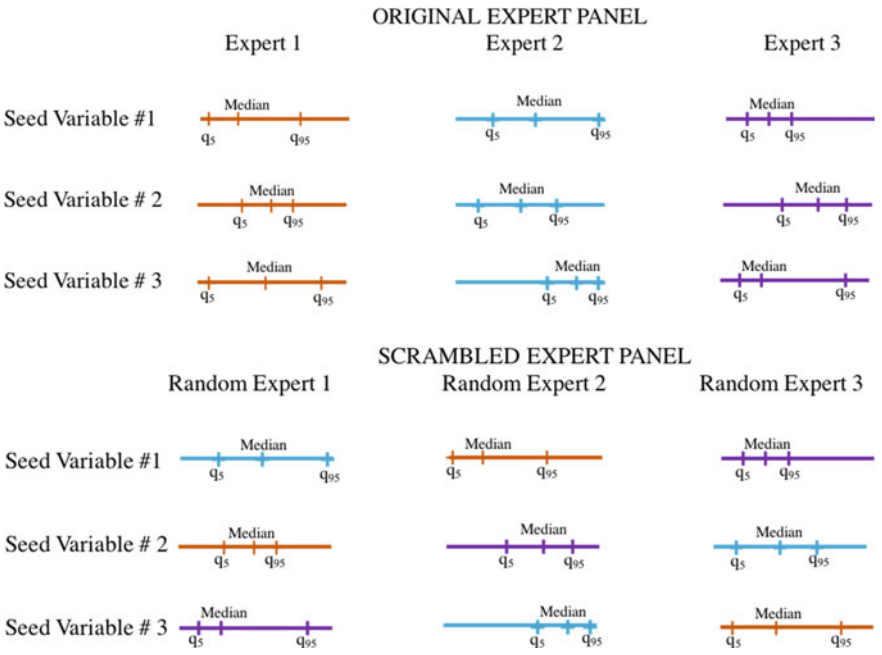

Fig. 3.1 An illustration of random expert creation process. q_5 corresponds to the 5^{th} percentile elicitation, the median corresponds to the 50^{th} percentile elicitation, and q_{95} corresponds to the 95^{th} percentile elicitation

The Classical Model assumes that expert performances in terms of statistical accuracy may vary systematically with respect to the persistent differences. Specifically, such persistent differences are the reasons why the "best performing expert" performs the best and the "poorest performing expert" performs the poorest. However, when the elicitations are scrambled for a very large number of runs, then scrambled experts should perform the same, statistically. That is, the "best performing expert" does no longer perform as much better than the other experts; the "poorest performing expert" does no longer perform as much poorly than the rest. In other words, the scrambling process eliminates the systematic variation, which implies a smaller standard deviation.

If REH is false, then the original expert panel should look statistically different from the population of scrambled panels. The systematic differences among experts, as posited by the Classical Model, lead to a larger average score and smaller standard deviation of the score. The maximum score (i.e., the best performing expert's score) of the original panel is expected to be higher than that of the scrambled panel. Similarly, the minimum score (i.e., the poorest performing expert's score) of the original panel is expected to be lower than that of the scrambled panel.

There are a number of ways in which a test could be constructed to examine whether the original expert panel comes from the same distribution as the scrambled panels. In this study, four tests of REH are identified. Specifically, if REH were true, then

(1) The probability is 50% that the average of the experts' statistical accuracies in the original panel is higher than that of a scrambled panel
(2) The probability is 50% that the standard deviation of the experts' statistical accuracies in the original panel is higher than that of a scrambled panel
(3) The probability is 50% that the maximum of the experts' statistical accuracies in the original panel is higher than that of a scrambled panel
(4) The probability is 50% that the minimum of the experts' statistical accuracies in the original panel is lower than that of a scrambled panel.

These predictions of REH were tested based on experts' statistical accuracy performances measured by the statistical accuracy score. The statistical accuracy score is the focus since it is the main characteristics of the performance-based weights (see the Cooke 1991 for discussion), while the information score has a role of modulating the statistical accuracy score.

3.3 Expert Judgment Data

TU Delft database provides extensive datasets of expert elicitations that were conducted based on Classical Model framework (Cooke and Goossens 2008). This database has been recently updated with new studies that were performed starting from 2006 to 2015 (Colson and Cooke 2017). As summarized by Colson and Cooke (2017), these studies were done by organizations such as Bristol University,

the British Government, United States Department of Homeland Security, World Health Organization, and the US Environmental Protection Agency, etc. Studies were performed in two formats of structured expert judgment, in three percentiles and five percentiles. Experts are asked to elicit the seed variables for 5th, 50th, and 95th percentiles in the former format, and 5th, 25th, 50th, 75th, and 95th percentiles in the latter format. These elicitations were compiled by Cooke and Goosens (2008) and made available to the researchers and recently updated (available at http://rogermcooke.net/). This study focuses on all 44 datasets that are available in the new post-2006 expert judgment database. 27 of these 44 datasets came from studies, which were performed in three-percentile format, and 17 of the 44 datasets were performed in five-percentile format: experts were asked to provide five percentiles for the elicited variables.

Table 3.1 summarizes the names, the percentile format, number of experts, number of seed variables, and associated references for each expert judgment panel. The studies are across wide range of domains such as environmental risk, bioterrorism, air traffic control, and volcano eruptions. The number of experts in the panels of these studies ranged from 4 to 21, and the number of seed variables ranged from 8 to 48. The three-percentile format data has 298 experts who elicited 386 seed variables in total, which yielded 4597 elicitations in total. The five-percentile format data has 111 experts who elicited 170 seed variables in total, which yielded a total of 1117 elicitations.

3.4 Hypothesis Testing

44 studies presented in Table 3.1 are used to test the random expert hypothesis. For each study, hypothetical expert judgment panels consisting of randomly scrambled experts are simulated in 1000 runs. The extent to which this data support the REH can be statistically examined by a Binomial test for each of the four statistical metrics, namely, average, standard deviation, maximum, and the minimum scores of expert panels for each study.

$$H_0 : r = 0.5$$

$$H_a : r > (<) 0.5$$

where r is the percentage of the studies in which the original experts outperform the random experts.

r is the success probability in which the success, "outperformance," is defined as follows:

1. The average statistical accuracy score of the original expert panels is higher than that of a scrambled expert panel

Table 3.1 Expert judgment studies are illustrated with the number of seed variables and experts, and percentile formats

Study	Percentile format	# of experts	# of seed variables	Subject
UMD	3	9	11	Nitrogen removal in Chesapeake Bay
USGS	3	18	32	Volcanos
arsenic	3	9	10	Air quality levels for arsenic
Biol Agents	3	9	10	Human dose–response curves for bioterror agents
Geopolit	3	9	16	Geopolitics
ATCEP	3	5	10	Air traffic controllers human error
Daniela	3	4	10	Fire prevention and control
eBBP	3	14	15	XMRV blood/tissue infection transmission risks
create	3	7	10	Terrorism
effErupt	3	14	8	Icelandic fissure eruptions: source characterization
erie	3	10	15	Establishment of Asian Carp in Lake Erie
FCEP	3	5	8	Flight crew human error
Sheep	3	14	15	Risk management policy for sheep scab control
Hemophilia	3	18	8	Hemophilia
Liander	3	11	10	Underground cast iron gas-lines
PHAC	3	10	12	Additional CWD factors
TOPAZ	3	21	16	Tectonic hazards for radwaste siting in Japan
SPEED	3	14	16	Volcano hazards (Vesuvius and Campi Flegrei, Italy)
TDC	3	18	17	Volcano hazards (Tristan da Cunha)

(continued)

Table 3.1 (continued)

Study	Percentile format	# of experts	# of seed variables	Subject
GL	3	9	13	Costs of invasive species in Great Lakes
Goodheart	3	5	10	Airport safety
Ice	3	10	11	Sea level rise from ice sheets melting due to global warming
puig-gdp	3	9	13	Emission forecasts from Mexico
puig-oil	3	6	19	Oil emissions and prices
YTBID (CDC)	3	14	48	Return on investment for CDC warnings
Gerestenberger	3	12	13	Probabilistic seismic-hazard model for canterbury
CWD	3	14	10	Infection transmission risks: Chronic wasting disease from deer to humans
Nebraska	5	4	10	Grant effectiveness, child health insurance enrollment
San Diego	5	7	10	Effectiveness of surgical procedures
BFIQ	5	7	11	Breastfeeding and IQ
France	5	5	10	Future antimicrobial resistance in France
Italy	5	4	8	Future antimicrobial resistance in Italy
Spain	5	5	10	Future antimicrobial resistance in Spain
UK	5	6	10	Future antimicrobial resistance in UK
Arkansas	5	4	10	Grant effectiveness, child health insurance enrollment
CoveringKids	5	5	10	Grant effectiveness, child health insurance enrollment
dcpn_Fistula	5	8	10	Effectiveness of obstetric fistula repair

(continued)

Table 3.1 (continued)

Study	Percentile format	# of experts	# of seed variables	Subject
Florida	5	7	10	Grant effectiveness, child health insurance enrollment
Illinois	5	5	10	Grant effectiveness, child health insurance enrollment
Obesity	5	4	10	Grant effectiveness, childhood obesity
Tobacco	5	7	10	Grant effectiveness, childhood obesity
Washington	5	5	10	Grant effectiveness, child health insurance enrollment
cdc-roi	5	20	10	Return on investment for CDC warnings
IQ-earn	5	8	11	Effects of increases in IQ in India on the present value of Lifetime earnings

Note The references to the data can be found in the Appendix

2. The standard deviation of statistical accuracy scores of the original expert panels is higher than that of a scrambled expert panel
3. The maximum statistical accuracy scores of the original expert panels is higher than that of a scrambled expert panel
4. The minimum statistical accuracy scores of the original expert panels is lower than that of a scrambled expert panel.

3.5 Results

The data were analyzed in two different formats: (1) in three-percentile format data, including all 44 available datasets (thus five-percentile datasets were converted to three-percentile datasets), (2) in five-percentile format data, including only five-percentile elicitations.

3.5.1 The Analysis of the Three-Percentile Format Data

The average, standard deviation, the maximum, and the minimum scores of the original experts are compared with those of the random experts in each randomly

created 1000 expert panels. The Binomial tests are performed for all 44 datasets available in three-percentile format.

The statistical accuracy scores were computed for the three-percentile format data, consisting of 27 studies that were originally performed as a three-percentile format, and 17 five-percentile studies that were converted to three-percentile format. Table 3.2 provides the statistical summaries of the original experts' statistical accuracy scores: summaries, average, standard deviation, maximum, and minimum scores of expert panels.

Then, four statistical metrics—average, standard deviation, maximum, and minimum of the statistical accuracy scores—were computed for the original expert panels and for each of the 1000 scrambled expert panels. Then, for each expert panel, the percentage that the original experts' corresponding statistics ranked higher than (lower for the minimum) those of the 1000 scrambled expert panels was determined. Under the REH, the original expert panels' metrics should be ranked above (below) those of the scrambled expert panels 50% of the time. Table 3.3 illustrates the actual percentages determined for each dataset.

For example, in Table 3.3, the corresponding percentage for the average score in the study UMD is shown as 99.7%. This indicates that in 99.7% of the scrambled panels (997 out of 1000 simulation runs), the average scores of the original experts are greater than those of the randomly scrambled experts. Similarly, in the UMD study, in 96.4% of the scrambled panels (964 out of 1000 simulation runs), the standard deviation of the experts in the original panel are greater than those of scrambled experts, indicating a larger variation in the original expert score in most cases. The best performing expert in UMD study outperforms the best performing expert of the scrambled panels in 95.4% of the time. This means that, in 954 out of 1000 simulation runs randomly created expert panels, the best performing experts are outperformed by the original best performing expert. Finally, 100% for the minimum score displayed in Table 3.3 shows that the minimum score of the original expert panel was lower than those of all scrambled panels, indicating that the score of the poorest performing expert of the original panel performed the poorest compared to all random experts.

Figure 3.2 shows that in 16 out of 44 studies, the original experts outperformed more than 95% (i.e., 950 out of 1000 simulation runs) of the scrambled expert panels. Similarly, in 5 studies, the original experts outperformed the scrambled experts in 85–95% of the time. In total, in 33 out of 44 studies, the original experts' average scores ranked higher than those of the scrambled experts from 1000 expert panels at least 80% of the time.

Figure 3.3 shows that, in 10 studies, the standard deviation of the experts' statistical accuracy scores in the original panel is larger than those in more than 95% of the 1000 randomly created expert panels. In 28 out of 44 studies, the variation in the original expert data is larger than the variation in the scrambled expert panels at least 80% of the time.

Figure 3.4 shows that, in 10 studies, the best performing expert in the original expert panel outperforms the best performing expert in the random expert panels in more than 95% of the time. In 26 out of 44 studies, the best performing original

Table 3.2 Statistical accuracy scores of the original experts for the three-percentile format elicitation data

Study No.	Study name	Average	Standard deviation	Max	Min
1	UMD	1.33E-01	2.69E-01	7.06E-01	3.21E-14
2	USGS	3.15E-03	1.17E-02	5.55E-02	7.12E-13
3	arsenic	4.84E-03	1.18E-02	3.57E-02	9.86E-07
4	Biol Agents	5.70E-02	1.16E-01	3.11E-01	1.42E-06
5	Geopolit	5.10E-02	1.01E-01	2.30E-01	1.42E-06
6	ATCEP	2.99E-02	4.47E-02	1.01E-01	1.42E-06
7	Daniela	1.88E-01	2.57E-01	5.54E-01	4.35E-07
8	eBBP	2.00E-01	2.63E-01	8.33E-01	8.91E-06
9	create	3.57E-03	6.37E-03	1.71E-02	8.91E-06
10	effErupt	2.91E-02	5.46E-02	1.85E-01	8.91E-06
11	erie	2.27E-01	2.46E-01	6.61E-01	1.08E-08
12	FCEP	1.75E-01	2.84E-01	6.64E-01	5.12E-05
13	Sheep	5.64E-02	1.70E-01	6.43E-01	1.62E-11
14	hemophilia	1.88E-01	2.28E-01	6.64E-01	2.66E-04
15	Liander	3.18E-04	8.37E-04	2.81E-03	3.50E-08
16	PHAC	9.71E-03	2.46E-05	7.50E-05	2.43E-10
17	TOPAZ	3.08E-02	1.00E-01	2.43E-10	4.42E-12
18	SPEED	1.83E-02	6.03E-02	2.27E-01	2.88E-12
19	TDC	1.03E-01	2.72E-01	9.89E-01	1.02E-12
20	GL	6.13E-02	1.51E-01	4.54E-01	1.91E-09
21	Goodheart	1.47E-01	2.76E-01	7.07E-01	7.99E-04
22	Ice	8.53E-02	1.50E-01	3.99E-01	5.84E-06
23	puig-gdp	3.68E-02	9.16E-02	2.77E-01	5.04E-12
24	puig-oil	1.72E-03	4.17E-03	1.02E-02	3.27E-12
25	YTBID (CDC)	1.43E-01	2.23E-01	9.68E-01	5.80E-07
26	Gerestenberger	6.35E-02	6.29E-02	1.52E-01	1.88E-05
27	CWD	7.62E-02	1.47E-01	4.93E-01	1.07E-06
28	Nebraska	1.89E-03	3.71E-03	7.46E-03	4.54E-10
29	San Diego	3.45E-04	5.91E-04	1.31E-03	8.36E-11
30	BFIQ	1.24E-01	2.33E-01	6.38E-01	2.28E-04
31	France	1.56E-01	3.09E-01	7.07E-01	1.54E-07
32	Italy	1.70E-01	3.14E-01	6.40E-01	5.86E-07
33	Spain	4.70E-06	9.07E-06	2.08E-05	1.29E-10
34	UK	1.49E-01	2.72E-01	6.83E-01	6.17E-09
35	Arkansas	8.00E-02	1.56E-01	3.14E-01	1.07E-06
36	CoveringKids	2.76E-01	3.02E-01	6.83E-01	9.86E-07

(continued)

Table 3.2 (continued)

Study No.	Study name	Average	Standard deviation	Max	Min
37	dcpn_Fistula	6.54E-04	1.13E-03	2.81E-03	9.86E-07
38	Florida	2.24E-02	2.36E-02	4.70E-02	5.21E-06
39	Illinois	1.75E-02	3.23E-02	7.50E-02	5.45E-08
40	Obesity	6.67E-02	9.06E-02	1.92E-01	2.47E-10
41	Tobacco	2.06E-01	2.39E-01	6.83E-01	5.99E-03
42	Washington	6.29E-02	1.04E-01	2.44E-01	5.99E-04
43	cdc-roi	1.08E-01	1.46E-01	4.93E-01	3.50E-08
44	IQ-earn	6.88E-02	1.26E-01	3.70E-01	1.70E-07

Note First 27 datasets were expert elicitations based on three-percentile format (5^{th}, 50^{th}, and 95^{th} percentiles) and last 17 studies were converted into three-percentile format by truncating the 25^{th} and the 75^{th} percentiles)

Table 3.3 The percentage of original experts' corresponding statistics ranked higher than (lower for the minimum) those of the 1000 randomly created expert panels (the entire available data in three-percentile format)

Study No.	Study name	Average (%)	Standard deviation (%)	Max (%)	Min (%)
1	UMD	99.70	96.40	95.40	100.00
2	USGS	86.60	84.50	79.40	80.10
3	arsenic	57.80	60.80	56.50	43.40
4	Biol Agents	84.20	73.10	60.30	69.80
5	Geopolit	87.20	82.70	76.30	54.80
6	ATCEP	95.80	94.70	93.90	99.50
7	Daniela	91.90	64.70	63.60	99.70
8	eBBP	99.10	91.40	83.30	88.30
9	create	23.00	34.70	20.80	13.10
10	effErupt	85.90	80.50	54.00	88.80
11	erie	87.10	71.10	75.00	100.00
12	FCEP	93.30	84.50	85.00	92.00
13	Sheep	98.80	97.70	97.80	99.20
14	hemophilia	90.40	77.30	24.20	34.70
15	Liander	21.20	26.50	25.30	36.40
16	PHAC	56.60	48.70	22.50	99.70
17	TOPAZ	98.00	98.00	98.00	8.00
18	SPEED	97.90	97.50	97.50	97.60
19	TDC	100.00	100.00	97.50	99.10
20	GL	100.00	99.40	99.10	98.80
21	Goodheart	82.70	83.90	83.10	34.70
22	Ice	95.00	91.50	82.10	55.00

(continued)

Table 3.3 (continued)

Study No.	Study name	Average (%)	Standard deviation (%)	Max (%)	Min (%)
23	puig-gdp	96.70	96.40	96.30	99.30
24	puig-oil	97.20	97.20	97.20	73.60
25	YTBID (CDC)	97.90	94.20	80.30	88.70
26	Gerestenberger	6.35	6.29	15.19	73.80
27	CWD	82.90	79.80	78.40	71.50
28	Nebraska	76.80	78.70	78.70	97.80
29	San Diego	91.10	90.40	85.10	74.10
30	BFIQ	80.10	90.40	80.60	48.30
31	France	99.80	98.90	98.90	97.20
32	Italy	80.50	85.80	81.60	99.70
33	Spain	59.40	40.80	35.70	88.30
34	UK	96.30	89.10	88.50	99.90
35	Arkansas	97.90	96.20	95.20	81.90
36	CoveringKids	96.00	85.00	62.90	98.70
37	dcpn_Fistula	11.60	14.90	9.30	17.10
38	Florida	54.20	33.70	14.70	60.30
39	Illinois	79.40	72.80	72.70	76.40
40	Obesity	93.10	90.50	90.50	99.90
41	Tobacco	40.40	49.30	36.30	58.10
42	Washington	26.70	40.80	37.70	50.90
43	cdc-roi	94.80	61.10	58.10	91.60
44	IQ-earn	2.30	16.60	26.40	99.60

expert outperformed the best performing random expert in at least 80% of the 1000 scrambled expert panels.

Figure 3.5 shows that, in 19 studies, the poorest performing expert in the original expert panel performed poorer than the poorest performing expert in the randomly created expert panels. In 26 out of 44 studies, the original experts' minimum score is lower than the random expert panel's minimum score in at least 80% of the 1000 expert panels.

In above results, the percentage score may be a function of the number of experts and the corresponding spread in the calibration scores of the experts. The exact determination is a subject of future research. However, the empirical results from above suggest that the REH may not be appropriate. To more formally test the REH, a statistical test is needed. The test selected was the Binomial test due to its appropriateness for dichotomous outcomes and its nonparametric nature.

The Binomial test results show that the average of the statistical accuracy results of the original experts outperformed the randomly created experts more than 50% of the time, in three statistical metrics: on average ($p = 2.65\text{E-}06$), on standard deviation

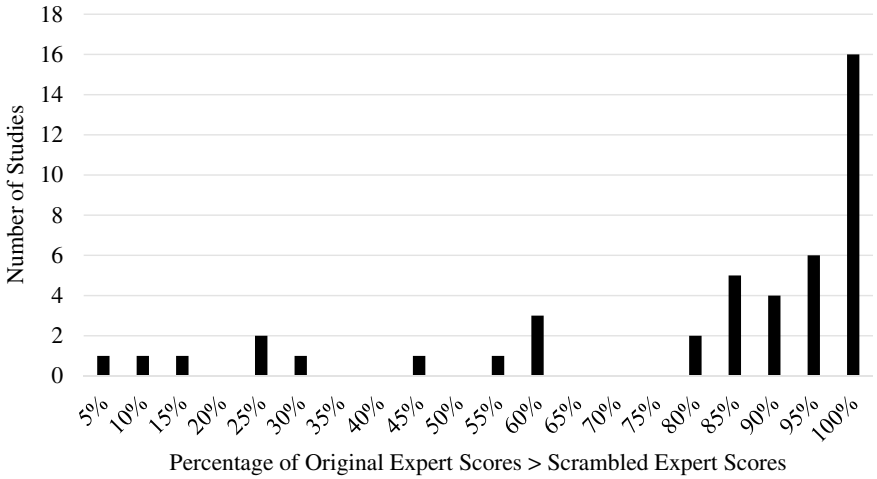

Fig. 3.2 Distribution of percentage of original experts' average statistical accuracy scores ranked higher among those of scrambled experts in 1000 hypothetical expert panels based on 44 studies

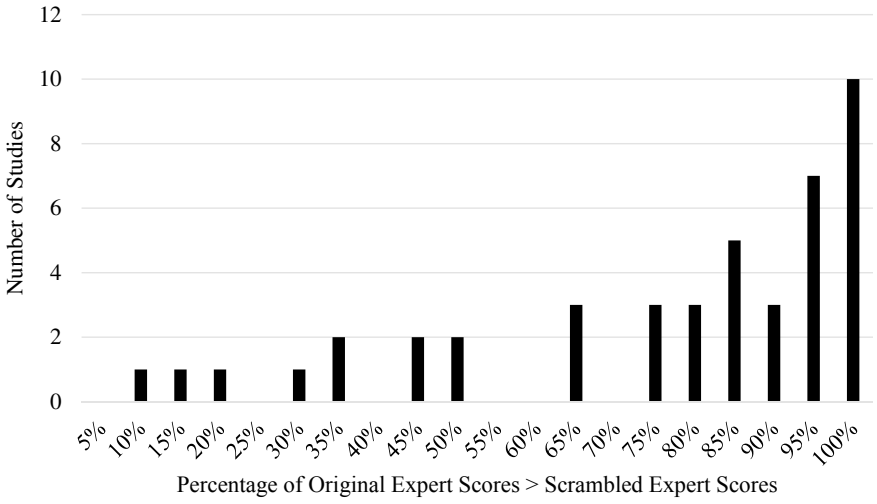

Fig. 3.3 Distribution of percentage of the standard deviation of the original experts' statistical accuracy scores ranked higher among those of scrambled experts in 1000 hypothetical expert panels based on 44 studies

($p = 1.94\text{E-}04$), and on maximum scores ($p = 6.3\text{E-}04$). Also, the minimum of the original experts performed significantly poorer than the poorest performing randomly created experts more than 50% of the time ($p = 1.27\text{E-}05$).

Overall, the results of the random expert hypothesis testing show that, in a significant number of studies, the scrambled experts fail to perform as well as the original

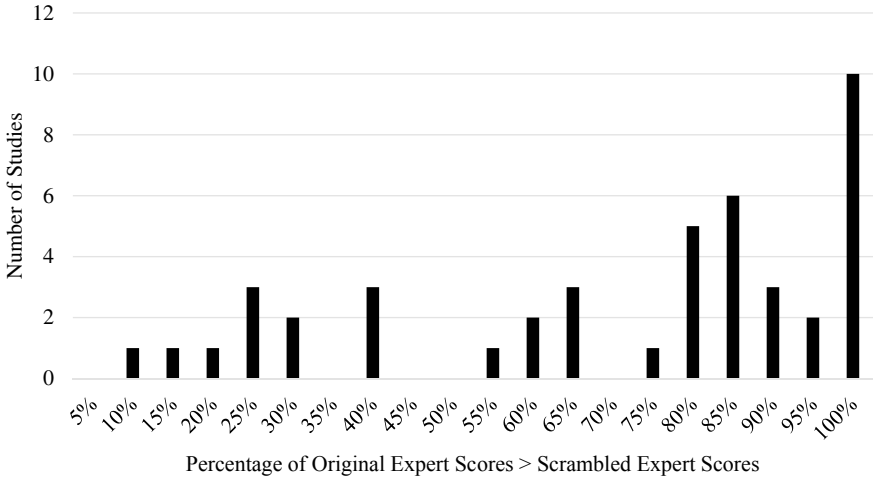

Fig. 3.4 Distribution of percentage of the maximum of the original experts' statistical accuracy scores ranked higher among those of scrambled experts in 1000 hypothetical expert panels based on 44 studies

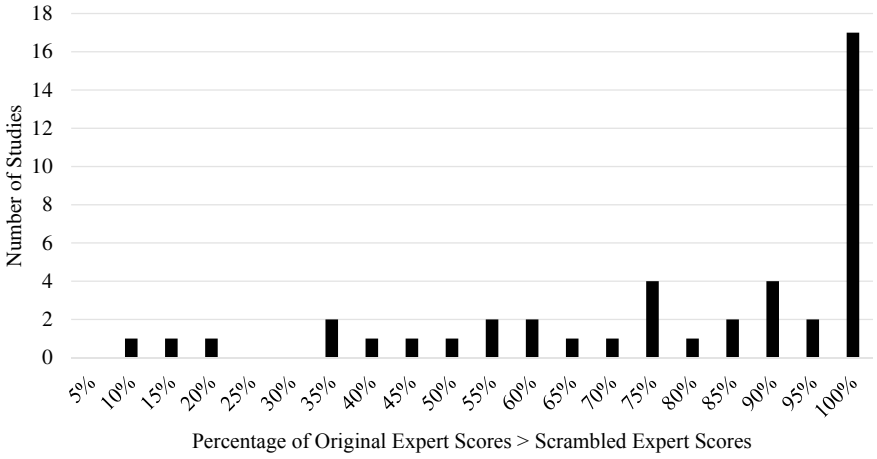

Fig. 3.5 Distribution of percentage of the minimum of the original experts' statistical accuracy scores ranked lower among those of scrambled experts in 1000 hypothetical expert panels based on 44 studies

experts. Specifically, in most studies, the original experts outperformed the scrambled experts in an overwhelmingly large percentage of the hypothetical expert panels. Binomial test results suggest that the original experts ranked higher than the scrambled experts in three statistical summaries, the average, standard deviation, and the maximum of statistical accuracy score, and ranked lower than in terms of the

minimum score. This indicates that the hypothesis that expert performances occur due to randomness is extremely unlikely.

3.5.2 Analysis of the Five-Percentile Format Data

The conventional elicitation format in Structured Expert Judgment practices is three-percentile format; however, in some cases, analysts would prefer five-percentile format where they ask experts their elicitations in five percentiles such as 5^{th}, 25^{th}, 50^{th}, 75^{th}, and 95^{th} percentiles. Therefore, it is deemed important to test the REH on alternative elicitation formats. In this section, the same analyses performed in the previous section to the entire dataset available were computed for 17 studies that were originally performed as five-percentile format.

Table 3.4 shows statistical accuracy scores of five-percentile elicited data. The summary statistics shown in the table were incorporated into the next analyses where the corresponding statistics of the scrambled panels were compared with the original experts. Table 3.5 shows the percentages that the average, standard deviation, and the maximum of the original experts outperformed the random expert panels, and the

Table 3.4 Average, standard deviation, max, and min of original experts' statistical accuracy scores for the five-percentile format elicitation data

Study No.	Study name	Average	Standard deviation	Max	Min
28	Nebraska	8.35E-03	1.64E-02	3.30E-02	7.34E-09
29	San Diego	6.97E-04	1.43E-03	3.82E-03	1.02E-09
30	BFIQ	1.45E-01	2.56E-01	6.92E-01	3.02E-04
31	France	1.37E-01	2.88E-01	6.52E-01	1.99E-07
32	Italy	1.37E-01	2.88E-01	6.52E-01	1.99E-07
33	Spain	7.02E-06	1.00E-05	2.24E-05	1.02E-09
34	UK	6.42E-02	9.21E-02	1.85E-01	1.96E-08
35	Arkansas	1.93E-02	3.39E-02	6.98E-02	1.15E-05
36	CoveringKids	3.28E-01	3.40E-01	7.56E-01	6.23E-06
37	dcpn_Fistula	1.81E-03	3.10E-03	7.62E-03	6.23E-06
38	Florida	3.81E-02	4.63E-02	1.25E-01	1.18E-05
39	Illinois	3.68E-02	5.48E-02	1.32E-01	3.32E-07
40	Obesity	1.66E-01	2.11E-01	4.40E-01	4.09E-09
41	Tobacco	2.06E-01	2.61E-01	6.88E-01	1.05E-03
42	Washington	3.14E-02	3.09E-02	6.98E-02	3.82E-03
43	cdc-roi	1.30E-01	2.25E-01	7.20E-01	2.18E-07
44	IQ-earn	7.96E-02	1.56E-01	4.54E-01	6.97E-07

Table 3.5 The percentage of original experts' corresponding statistics ranked higher (lower for minimum) than those of the 1000 scrambled expert panels for five-percentile format elicitation data

Study name	Average (%)	Standard deviation (%)	Max (%)	Min (%)
Nebraska	85.30	85.60	85.60	97.70
San Diego	88.60	88.60	88.70	79.50
BFIQ	82.50	87.70	86.60	59.20
France	99.70	99.70	99.70	99.30
Italy	96.50	97.60	98.20	100.00
Spain	62.30%	50.30	47.90	90.40
UK	48.30	51.30	45.00	99.90
Arkansas	72.70	74.90	74.80	76.00
CoveringKids	96.40	82.00	75.50	98.60
dcpn_Fistula	18.80	20.90	15.60	10.70
Florida	50.00%	32.10	29.30	71.30
Illinois	80.10	70.90	72.20	82.60
Obesity	99.20	94.90	94.90	99.80
Tobacco	32.20	50.80	45.70	91.10
Washington	3.50	4.00	3.00	18.80
cdc-roi	96.50	92.10	77.00	94.80
IQ-earn	2.60	17.10	17.00	99.80

minimum of the original expert score is lower than the minimum of the scrambled experts.

Figure 3.6 shows that, in 7 out of 17 studies, the original experts outperformed more than 95% (i.e., 950 out of 1000 simulation runs) of the scrambled expert panels. In total, in 11 out of 17 studies, the original experts' average scores ranked higher than those of the scrambled experts from 1000 expert panels at least 80% of the time.

Figure 3.7 shows that, in 6 studies, the standard deviation of the experts' statistical accuracy scores in the original panel is larger than those in more than 95% of the 1000 randomly created expert panels. In 10 out of 17 studies, the variation in the original expert data is larger than the variation in the scrambled expert panels at least 80% of the time.

Figure 3.8 shows that, in 3 studies, the best performing expert in the original expert panel outperforms the best performing expert in the random expert panels in more than 95% of the time. In 9 out of 17 studies, the best performing original expert outperformed the best performing random expert in at least 80% of the 1000 scrambled expert panels.

Figure 3.9 shows that, in 8 studies, the poorest performing expert in the original expert panel performed poorer than the poorest performing expert in the randomly created expert panels. In 11 out of 17 studies, the original experts' minimum score is lower than the random expert panel's minimum score in at least 80% of the 1000 expert panels.

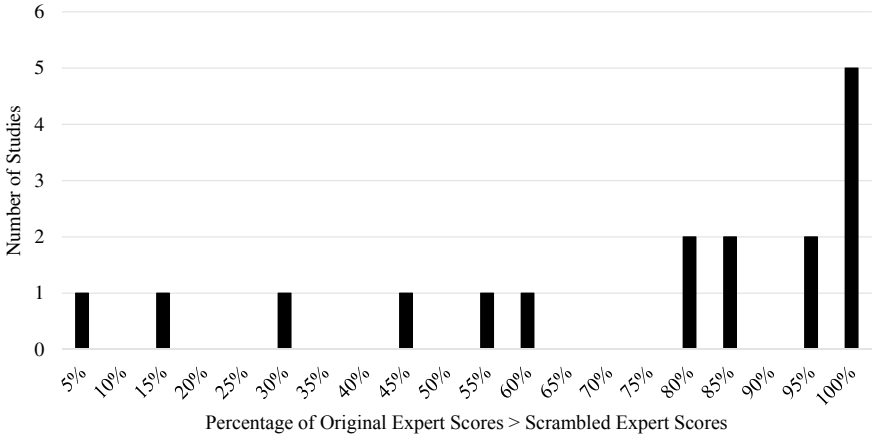

Fig. 3.6 Distribution of percentage of the average statistical accuracy of the original experts' statistical accuracy scores ranked higher among those of scrambled experts in 1000 hypothetical expert panels based on 17 studies that were originally elicited in five-percentile formats

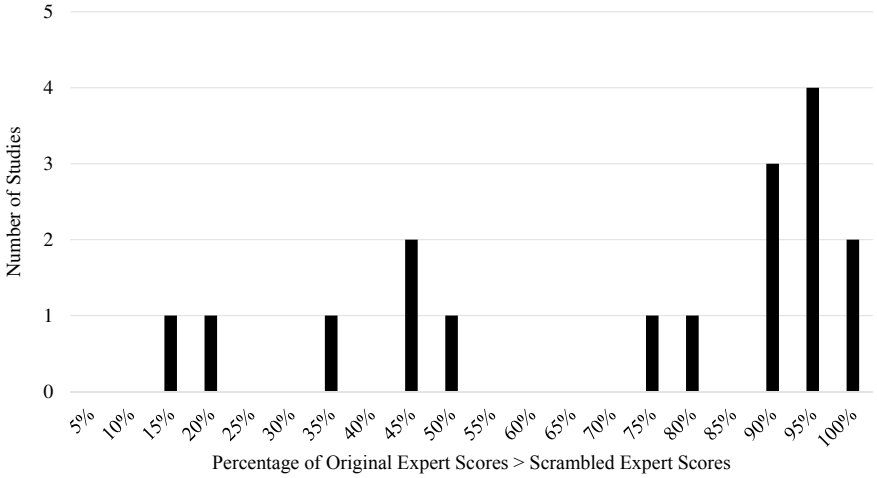

Fig. 3.7 Distribution of percentage of the standard deviation of the original experts' statistical accuracy scores ranked higher among those of scrambled experts in 1000 hypothetical expert panels based on 17 studies that were originally elicited in five-percentile formats

The Binomial results show that the average statistical accuracy scores of the original experts outperformed the randomly created experts in more than 50% of the 17 studies ($p = 0.024$). However, the Binomial test results show that the proportions of the studies in which standard deviation and maximum scores of the original experts outperform those of random experts were not statistically significant ($p = 0.167$ and $p = 0.17$, respectively). Finally, the Binomial test results indicate a significant

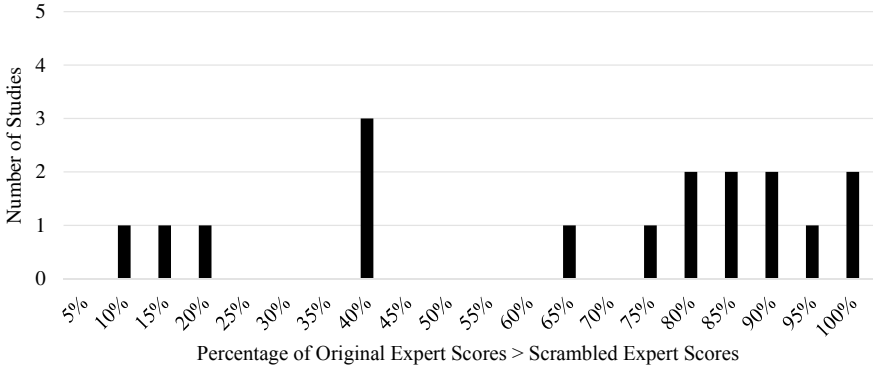

Fig. 3.8 Distribution of percentage of the maximum of the original experts' statistical accuracy scores ranked higher among those of scrambled experts in 1000 hypothetical expert panels based on 17 studies that were originally elicited in five-percentile formats

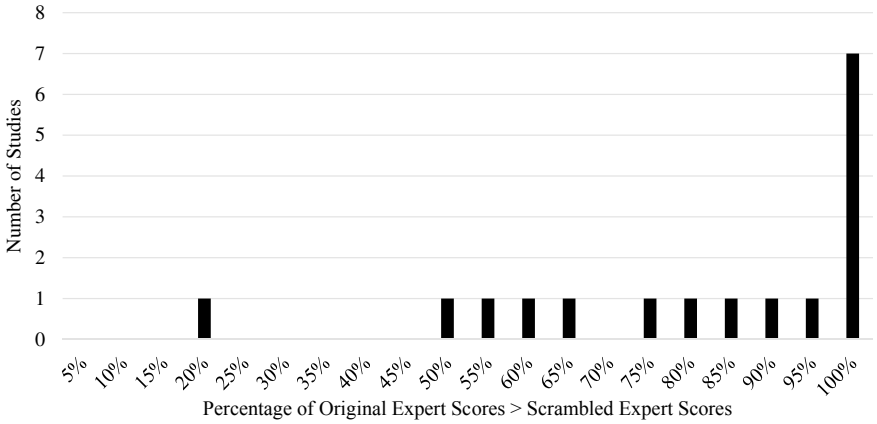

Fig. 3.9 Distribution of percentage of the minimum of the original experts' statistical accuracy scores ranked lower among those of scrambled experts in 1000 hypothetical expert panels based on 17 studies that were originally elicited in five-percentile formats

proportion of the studies in which the minimum of the original expert statistical accuracy scores was outperformed by the minimum of the random experts ($p = 0.00117$). As expected, the statistical tests with only 17 studies have much lower power.

3.5.3 A Sign Test Between the Three-Percentile Format and Five-Percentile Format Elicitation Data

Studies that were originally conducted to collect as in five percentiles (i.e., 5^{th}, 25^{th}, 50^{th}, 75^{th}, and 95^{th} percentiles) format were converted into the three-percentile format (5^{th}, 50^{th}, and 95^{th} percentiles). The average statistical accuracy scores of original experts in both formats were computed and compared by a two-sided sign test. This test was done in R, using "Wilcoxon Rank Sum and Signed Rank Test" function. The test results show that the difference between three and five-percentile formats were not statistically different ($W = 144$, $p = 1$). The corresponding test results for the standard deviation ($W = 145$, $p = 1$), maximum ($W = 137$, $p = 0.81$), and minimum ($W = 118$, $p = 0.37$) were also not significant, indicating that when experts are asked their elicitations in either three or five-percentile formats, their statistical accuracy did not significantly change. This implies that the number of probability bins and in turn bin range (e.g., whether covers 25% or 45%) do not significantly influence experts' statistical accuracy.

Furthermore, the sign test was performed to test whether the original experts' outperformance percentages differed in three-percentile format than in five-percentile format. Sign test results show that there was not a statistical difference between the three-percentile format analysis and five-percentile format analysis in terms of percentages that the original experts outperform the scrambled experts in 1000 simulations ($W = 146.5$, $p = 0.96$). Similar analysis was done for standard deviation ($W = 146$, $p = 0.97$; i.e., the percentage that the original experts outperform the random experts in their standard deviation), for the maximum ($W = 141$, $p = 0.92$; i.e., the percentage that the original experts outperform the scrambled experts in their maximum scores), and for the minimum ($W = 131$ $p = 0.65$; i.e., the percentage that the minimum score of the original experts is less than the minimum score of the random expert panels).

3.6 Concluding Remarks

This book chapter addresses the fundamental limitation of the equal weighting approach, namely that experts are expected to be interchangeable. This assumption has severe implications because it treats the best performing experts equally with the poor performing experts. Specifically, it leads to a depreciation of the maximum value of the expert input by undervaluing useful expert elicitations and overvaluing redundant or misleading elicitations of poorly performing experts. In order to address the aforementioned limitation of the equal weighting approach, the random expert hypothesis was used to test if experts should be treated equally. The results provide strong evidence that the original expert panels outperform randomly created experts. Specifically, the performances of the original experts with those of randomly scrambled experts were compared in terms of their statistical accuracy. Results show that

the original experts perform better than the randomly created experts; their statistical accuracy scores spread more since there are good and poor performing experts, which illustrates the potential problem of the equal weight approach. It may not be reasonable to assign all experts equal weights.

The present study also tested whether the results are replicated in the different elicitation format, specifically three versus five-percentile format. This analysis has significant practical implications. Showing the differences in statistical accuracy in different elicitation formats offers valuable insights to analysts so that they can decide the number of bins that they would ask experts to elicit. If there are performance differences between the three and five-quantile formats, they are too small to be detected with the current dataset. This question could be revisited in the future as more data become available.

This study focused on comparing performances in terms of statistical accuracy scores. As proposed by the Classical Model (e.g., Cooke 1991; Cooke et al. 2008), the statistical accuracy score is the dominant component in expert decision weight computations. Specifically, the Classical Model gives the power to the analyst to exclude the assessment of an expert whose statistical accuracy performance is less than a given threshold. In other words, it is the statistical accuracy that determines whether an experts' input is included into the analysis. As aforementioned, the information score functions serve as a modulating factor for evaluating expert performances. There may be cases where experts can provide large intervals indicating greater uncertainty in their estimates, which would still guarantee a high statistical accuracy score yet may not be as informative. Information score is an effective way to penalize those experts. Therefore, it is encouraged to investigate the random expert hypothesis based on decision weights that encompasses both statistical accuracy and information score. In future studies, thorough analyses including large dataset will be analyzed.

Finally, it is useful to compare this study with previous cross-validation studies (Eggstaff et al. 2014; Colson and Cooke 2017). Those studies considered all non-trivial splits of the statistical accuracy variables into training and test sets. The Classical Model performance weight was initialized on each training set and compared to equal weighting on the respective test sets. Although these studies showed significant out-of-sample superiority for performance weighting, the results were tempered by the fact that the performance weighting based on each training set is not the same as the performance weighting based on all variables. There was an out-of-sample penalty for statistical accuracy which decreased with training set size, but which obviously could not be eliminated. Hence, the superiority of performance weighting was largely driven by the higher informativeness of the performance weighted decision maker. The present results utilize the full set of statistical accuracy variables and do not consider informativeness. This suggests that performance weighting is also superior with respect to statistical accuracy in addition to informativeness. Working this out is a task for future research.

Appendix

Data references table

Study name	References
UMD	Koch, Benjamin J., Filoso, S., Cooke, R. M. Hosen, J. D., Colson, A.R. Febria, Catherine M., Palmer, M. A., (2015) Nitrogen in stormwater runoff from Coastal Plain watersheds: The need for empirical data, reply to Walsh, Elementa DOI https://doi.org/10.12952/journal.elementa.000079. https://www.elementascience.org/articles/79 Koch, Benjamin J., Febria, Catherine M., Cooke, Roger M. Hosen, Jacob D., Baker, Matthew E., Colson, Abigail R. Filoso, Solange, Hayhoe, Katharine, Loperfido, J.V., Stoner, Anne M.K., Palmer, Margaret A., (2015) Suburban watershed nitrogen retention: Estimating the effectiveness of storm water management structures, Elementa, https://doi.org/10.12952/journal.elementa.000063 https://www.elementascience.org/articles/63
USGS	Newhall, C. G., & Pallister, J. S. (2015). Using multiple data sets to populate probabilistic volcanic event trees. In Volcanic Hazards, Risks and Disasters (pp. 203–232)
arsenic	Hanzich, J.M. (2007) Achieving Consensus: An Analysis Of Methods To Synthesize Epidemiological Data For Use In Law And Policy. Department of Public Health & Primary Care, Institute Of Public Health, University of Cambridge; unpublished MPhil thesis, 66 pp + appendices
Biol Agents	Aspinall & Associates (2006). REBA Elicitation. Commercial-in-confidence report, pp. 26
Geopolit	Ismail and Reid (2006). "Ask the Experts" presentation
ATCEP	Morales-Nápoles, O., Kurowicka, D., & Cooke, R. (2008). EEMCS final report for the causal modeling for air transport safety (CATS) project
Daniela	Forys, M.B., Kurowicka, D., Peppelman, B.(2013) "A probabilistic model for a gas explosion due to leakages in the grey cast iron gas mains" Reliability Engineering & System Safety volume 119, issue, year 2013, pp. 270–279
eBBP	Tyshenko, M.G., S. ElSaadany, T. Oraby, M. Laderoute, J. Wu, W. Aspinall and D. Krewski (2011) Risk Assessment and Management of Emerging Blood-Borne Pathogens in Canada: Xenotropic Murine Leukaemia Virus-Related Virus as a Case Study for the Use of a Precautionary Approach. Chapter in: *Risk Assessment* (ISBN 979-953-307-765-8) Cashman, N.R., Cheung, R., Aspinall, W., Wong, M. and Krewski, D. (2014) Expert Elicitation for the Judgment of Prion Disease Risk Uncertainties associated with Urine-derived and Recombinant Fertility Drugs. Submitted to: Journal of Toxicology and Environmental Health
create	Bier V.M, Kosanoglu, F, Shin J, unpublished data, nd
effErupt	Aspinall, W.P. (2012) Comment on "Social studies of volcanology: knowledge generation and expert advice on active volcanoes" by Amy Donovan, Clive Oppenheimer and Michael Bravo [*Bull Volcanol* (2012) 74:677–689] Bulletin of Volcanology, 74, 1569–1570. https://doi.org/10.1007/s00445-012-0625-x

(continued)

(continued)

Study name	References
erie	Colson, Abigail R., Sweta Adhikari, Ambereen Sleemi, and Ramanan Laxminarayan. (2015) "Quantifying Uncertainty in Intervention Effectiveness with Structured Expert Judgment: An Application to Obstetric Fistula." BMJ Open, 1–8. https://doi.org/10.1136/bmjopen-2014-007233 Cooke, R.M., Wittmann, M.E., Lodge, D.M., Rothlisberger, J.D., Rutherford E.S., Zhang, H. and Mason, D.M. (2014) "Out-of-Sample Validation for Structured Expert Judgment of Asian Carp Establishment in Lake Erie", Integrated Environmental Assessment and Management, open access. DOI: https://doi.org/10.1002/ieam.1559 Zhang, H, Rutherford E.S., Mason, D.M., Breck, J,T,, Wittmann M.E., Cooke R.M., Lodge D.M., Rothlisberger J.D., Zhu X., and Johnson, T B., (2015) Forecasting the Impacts of Silver and Bighead Carp on the Lake Erie Food Web, Transactions of the American Fisheries Society, Volume 145, Issue 1, pp 136–162, https://doi.org/10.1080/00028487.2015.1069211
FCEP	Leontaris, G., & Morales-Nápoles, O. (2018). ANDURIL—A MATLAB toolbox for ANalysis and Decisions with UnceRtaInty: Learning from expert judgments. SoftwareX, 7, 313–317
Sheep	Hincks, T., Aspinall, W. and Stone, J. (2015) Expert judgement elicitation exercise to evaluate Sheep Scab control measures: Results of the Bayesian Belief Network analysis. University of Bristol PURE Repository Working Paper (forthcoming)
hemophilia	Fischer K, Lewandowski D, Janssen MP. Estimating unknown parameters in haemophilia using expert judgement elicitation. Haemophilia. 2013 Sep;19(5):e282–e288
Liander	Forys, M.B., Kurowicka, D., Peppelman, B.(2013) "A probabilistic model for a gas explosion due to leakages in the grey cast iron gas mains" Reliability Engineering & System Safety volume 119, issue, year 2013, pp. 270–279
PHAC	Oraby,T., Tyshenko, M.G., Westphal, M., Darshan, S., Croteau, M., Aspinall, W., Elsaadany, S., Cashman, N. and Krewski, D. (2011) Using Expert Judgments to Improve Chronic Wasting Disease Risk Management in Canada. Journal of Toxicology and Environmental Health, in press. Volume 74, Issue 2-4, 2011 Special Issue: Prion Research in Perspective 2010
TOPAZ	Scourse, E., Aspinall, W.P. and Chapman, N. (2014) Using expert elicitation to characterise long-term tectonic risks to radioactive waste repositories in Japan. Journal of Risk Research, https://doi.org/10.1080/13669877.2014.971334
SPEED	Hicks, A., Barclay, J., Simmons, P. and Loughlin, S. (2014). "An interdisciplinary approach to volcanic risk reduction under conditions of uncertainty: a case study of Tristan da Cunha." Nat. Hazards Earth Syst. Sci. 14(7): 1871-1887. https://doi.org/10.5194/nhess-14-1871-2014. www.nat-hazards-earth-syst-sci-discuss.net/1/7779/2013/ Bevilacqua, A., Isaia, R., Neri, A., Vitale, S., Aspinall, W.P. and eight others (2015) Quantifying volcanic hazard at Campi Flegrei caldera (Italy) with uncertainty assessment: I. Vent opening maps. Journal of Geophysical Research—Solid Earth; AGU. https://doi.org/10.1002/2014jb011775
TDC	Scourse, E., Aspinall, W.P. and Chapman, N. (2014) Using expert elicitation to characterise long-term tectonic risks to radioactive waste repositories in Japan. Journal of Risk Research, https://doi.org/10.1080/13669877.2014.971334

(continued)

(continued)

Study name	References
GL	Rothlisberger,J.D. Finnoff, D.C. Cooke,R.M. and Lodge, D.M. (2012) "Ship-borne nonindigenous species diminish Great Lakes ecosystem services" Ecosystems (2012) 15: 462–476 https://doi.org/10.1007/s10021-012-9522-6 Rothlisberger, J.D., Lodge, D.M. Cooke, R.M. and Finnoff, D.C. (2009) "Future declines of the binational Laurentian Great Lakes fisheries: recognizing the importance of environmental and cultural change" *Frontiers in Ecology and the Environment;* https://doi.org/10.1890/090002
Goodheart	Goodheart, B. (2013). Identification of causal paths and prediction of runway incursion risk by means of Bayesian belief networks. Transportation Research Record: Journal of the Transportation Research Board, (2400), 9–20.
Ice	Bamber, J.L., and Aspinall, W.P., (2012) An expert judgement assessment of future sea 1evel rise from the ice sheets, Nature Climate Change, PUBLISHED ONLINE: January 6, 2012| https://doi.org/10.1038/nclimate1778. http://www.nature.com/nclimate/journal/vaop/ncurrent/full/nclimate1778.html
puig-gdp	Puig, D., Morales-Nápoles, O., Bakhtiari, F., & Landa, G. (2018). The accountability imperative for quantifying the uncertainty of emission forecasts: evidence from Mexico. Climate Policy, 18(6), 742–751
puig-oil	Puig, D., Morales-Nápoles, O., Bakhtiari, F., & Landa, G. (2018). The accountability imperative for quantifying the uncertainty of emission forecasts: evidence from Mexico. Climate Policy, 18(6), 742–751
YTBID (CDC)	Colson, Abigail R., M.A. Cohen, S. Regmi, A. Nandi, R. Laxminarayan (2015) "Structured Expert Judgment for Informing the Return on Investment in Surveillance: The Case of Environmental Public Health Tracking." Working Paper. Center for Disease Dynamics, Economics & Policy
Gerestenberger	Gerstenberger, M. C., et al. (2016). "A Hybrid Time-Dependent Probabilistic Seismic-Hazard Model for Canterbury, New Zealand." Seismological Research Letters. Vol. 87 Doi: https://doi.org/10.1785/0220160084 Gerstenberger, M.C.; McVerry, G.H.; Rhoades, D.A.; Stirling, M.W. (2014) Seismic hazard modeling for the recovery of Christchurch, New Zealand.*Earthquake Spectra, 30(1):* 17–29; https://doi.org/10.1193/021913eqs037m Gerstenberger, M.C.; Christophersen, A.; Buxton, R.; Allinson, G.; Hou, W.; Leamon, G.; Nicol, A. (2013) Integrated risk assessment for CCS. p. 2775–2782; https://doi.org/10.1016/j.egypro.2013.06.162 IN: Dixon, T.; Yamaji, K. (eds) *11th International Conference on Greenhouse Gas Control Technologies, 18th-22nd November 2012, Kyoto International Conference Center, Japan.* Elsevier. *Energy procedia 37*

(continued)

(continued)

Study name	References
CWD	Tyshenko, M.G., ElSaadany, S., Oraby, T., Darshan, S., Catford, A., Aspinall, W., Cooke, R. and Krewski, D. (2012) Expert judgement and re-elicitation for prion disease risk uncertainties. *International Journal of Risk Assessment and Management*, 16(1–3), 48–77. https://doi.org/10.1504/ijram.2012.047552 Tyshenko, M.G., S. ElSaadany, T. Oraby, S. Darshan, W. Aspinall, R. Cooke, A. Catford, and D. Krewski (2011) Expert elicitation for the judgment of prion disease risk uncertainties. *J Toxicol Environ Health* A.; 74(2–4):261–285 Oraby,T., Tyshenko, M.G., Westphal, M., Darshan, S., Croteau, M., Aspinall, W., Elsaadany, S., Cashman, N. and Krewski, D. (2011) Using Expert Judgments to Improve Chronic Wasting Disease Risk Management in Canada. Journal of Toxicology and Environmental Health, in press. Volume 74, Issue 2–4, 2011 Special Issue: Prion Research in Perspective 2010
Nebraska	Attribution study for Robert Wood Johnson Covering Kids & Families in Pennsylvania, Washington, Nebraska, Illinois, Arkansas, and Florida, conducted by Center for Disease Dynamics, Economics & Policy, 2012
San Diego	Attribution study for Robert Wood Johnson Covering Kids & Families in Pennsylvania, Washington, Nebraska, Illinois, Arkansas, and Florida, conducted by Center for Disease Dynamics, Economics & Policy, 2012
BFIQ	Colson, A. Cooke, R.M., Lutter, Randall, (2016) How Does Breastfeeding Affect IQ? Applying the Classical Model of Structured Expert Judgment, Resources for the Future, RFF DP16–28 http://www.rff.org/research/publications/how-does-breastfeeding-affect-iq-applying-classical-model-structured-expert
France	Abigail R. Colson, Itamar Megiddo, Gerardo Alvarez-Uria, Sumanth Gandra, Tim Bedford, Alec Morton, Roger M. Cooke, Ramanan Laxminarayan (ns). "Quantifying Uncertainty about Future Antimicrobial Resistance: Comparing Structured Expert Judgment and Statistical Forecasting Methods."
Italy	Abigail R. Colson, Itamar Megiddo, Gerardo Alvarez-Uria, Sumanth Gandra, Tim Bedford, Alec Morton, Roger M. Cooke, Ramanan Laxminarayan (ns). "Quantifying Uncertainty about Future Antimicrobial Resistance: Comparing Structured Expert Judgment and Statistical Forecasting Methods."
Spain	Abigail R. Colson, Itamar Megiddo, Gerardo Alvarez-Uria, Sumanth Gandra, Tim Bedford, Alec Morton, Roger M. Cooke, Ramanan Laxminarayan (ns). "Quantifying Uncertainty about Future Antimicrobial Resistance: Comparing Structured Expert Judgment and Statistical Forecasting Methods."
UK	Abigail R. Colson, Itamar Megiddo, Gerardo Alvarez-Uria, Sumanth Gandra, Tim Bedford, Alec Morton, Roger M. Cooke, Ramanan Laxminarayan (ns). "Quantifying Uncertainty about Future Antimicrobial Resistance: Comparing Structured Expert Judgment and Statistical Forecasting Methods."
Arkansas	Attribution study for Robert Wood Johnson Covering Kids & Families in Pennsylvania, Washington, Nebraska, Illinois, Arkansas, and Florida, conducted by Center for Disease Dynamics, Economics & Policy, 2012.
CoveringKids	Attribution study for Robert Wood Johnson Covering Kids & Families in Pennsylvania, Washington, Nebraska, Illinois, Arkansas, and Florida, conducted by Center for Disease Dynamics, Economics & Policy, 2012

(continued)

(continued)

Study name	References
dcpn_Fistula	Aspinall,W. Devleesschauwer, B. Cooke, R.M., Corrigan,T., Havelaar, A.H., Gibb, H., Torgerson, P., Kirk, M., Angulo, F., Lake, R., Speybroeck, N., and Hoffmann, S. (2015) World Health Organization estimates of the relative contributions of food to the burden of disease due to selected foodborne hazards: a structured expert elicitation. PLOS ONE,: January 19, 2016 https://doi.org/10.1371/journal.pone.0145839
Florida	Attribution study for Robert Wood Johnson Covering Kids & Families in Pennsylvania, Washington, Nebraska, Illinois, Arkansas, and Florida, conducted by Center for Disease Dynamics, Economics & Policy, 2012
Illinois	Attribution study for Robert Wood Johnson Covering Kids & Families in Pennsylvania, Washington, Nebraska, Illinois, Arkansas, and Florida, conducted by Center for Disease Dynamics, Economics & Policy, 2012
Obesity	Colson, Abigail R., R.M. Cooke, R. Laxminarayan. (2015) "Attributing Impact to a Charitable Foundation's Programs with Structured Expert Judgment." Working Paper. Center for Disease Dynamics, Economics & Policy
Tobacco	Colson, Abigail R., R.M. Cooke, R. Laxminarayan. (2015) "Attributing Impact to a Charitable Foundation's Programs with Structured Expert Judgment." Working Paper. Center for Disease Dynamics, Economics & Policy
Washington	Attribution study for Robert Wood Johnson Covering Kids & Families in Pennsylvania, Washington, Nebraska, Illinois, Arkansas, and Florida, conducted by Center for Disease Dynamics, Economics & Policy, 2012
cdc-roi	Colson, Abigail R., M.A. Cohen, S. Regmi, A. Nandi, R. Laxminarayan (2015) "Structured Expert Judgment for Informing the Return on Investment in Surveillance: The Case of Environmental Public Health Tracking." Working Paper. Center for Disease Dynamics, Economics & Policy
IQ-earn	Randall Lutter, Abigail Colson, and Roger Cooke (ns), (ns), "Effects of Increases in IQ in India on the Present Value of Lifetime Earnings

References

Aspinall, W. (2010). A route to more tractable expert advice. *Nature, 463*(7279), 294.

Aspinall, W. P., Cooke, R. M., Havelaar, A. H., Hoffmann, S., & Hald, T. (2016). Evaluation of a performance-based expert elicitation: WHO global attribution of foodborne diseases. *PLoS ONE, 11*(3), e0149817.

Bamber, J. L., & Aspinall, W. P. (2013). An expert judgement assessment of future sea level rise from the ice sheets. *Nature Climate Change, 3*(4), 424.

Bamber, J. L., Aspinall, W. P., & Cooke, R. M. (2016). A commentary on "how to interpret expert judgment assessments of twenty-first century sea-level rise" by Hylke de Vries and Roderik SW van de Wal. *Climatic Change, 137*(3–4), 321–328.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5*(4), 559–583.

Clemen, R. T. (2008). Comment on Cooke's classical method. *Reliability Engineering & System Safety, 93*(5), 760–765.

Clemen, R. T., & Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis, 19*(2), 187–203.
Colson, A. R., & Cooke, R. M. (2017). Cross validation for the classical model of structured expert judgment. *Reliability Engineering & System Safety, 163,* 109–120.
Cooke, R. (1991). *Experts in uncertainty: Opinion and subjective probability in science.* Oxford University Press on Demand.
Cooke, R. M., & Goossens, L. L. (2008). TU Delft expert judgment data base. *Reliability Engineering & System Safety, 93*(5), 657–674.
Eggstaff, J. W., Mazzuchi, T. A., & Sarkani, S. (2014). The effect of the number of seed variables on the performance of Cooke′ s classical model. *Reliability Engineering & System Safety, 121,* 72–82.
Einhorn, H. J. (1974). Expert judgment: Some necessary conditions and an example. *Journal of Applied Psychology, 59*(5), 562.
French, S. (1981). Consensus of opinion. *European Journal of Operational Research, 7*(4), 332–340.
Genest, C., & Zidek, J. V. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science, 1*(1), 114–135.
Hald, T., Aspinall, W., Devleesschauwer, B., Cooke, R., Corrigan, T., Havelaar, A. H., et al. (2016). World Health Organization estimates of the relative contributions of food to the burden of disease due to selected foodborne hazards: A structured expert elicitation. *PLoS ONE, 11*(1), e0145839.
Jaiswal, K. S., Aspinall, W., Perkins, D., Wald, D., & Porter, K. A. (2012). Use of expert judgment elicitation to estimate seismic vulnerability of selected building types. In *15th World Conference on Earthquake Engineering (WCEE).* Lisbon, Portugal, Sept (pp. 24–28).
Keeney, R. L., & Von Winterfeldt, D. (1989). On the uses of expert judgment on complex technical problems. *IEEE Transactions on Engineering Management, 36*(2), 83–86.
Morgan, M. G., Henrion, M., & Small, M. (1992). *Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis.* Cambridge university press.
Mosleh, A., Bier, V. M., & Apostolakis, G. (1988). A critique of current practice for the use of expert opinions in probabilistic risk assessment. *Reliability Engineering & System Safety, 20*(1), 63–85.
Otway, H., & von Winterfeldt, D. (1992). Expert judgment in risk analysis and management: process, context, and pitfalls. *Risk Analysis, 12*(1), 83–93.
Ouchi, F. (2004). A literature review on the use of expert opinion in probabilistic risk analysis.
Ryan, J. J., Mazzuchi, T. A., Ryan, D. J., De la Cruz, J. L., & Cooke, R. (2012). Quantifying information security risks using expert judgment elicitation. *Computers & Operations Research, 39*(4), 774–784.
Singpurwalla, N. D. (1988). Foundational issues in reliability and risk analysis. *SIAM Review, 30*(2), 264–282.
Spetzler, C. S., & Stael von Holstein, C. A. S. (1975). Exceptional paper—probability encoding in decision analysis. *Management Science, 22*(3), 340–358.
Stone, M. (1961). The opinion pool. *The Annals of Mathematical Statistics*, 1339–1342.
Tyshenko, M. G., ElSaadany, S., Oraby, T., Darshan, S., Aspinall, W., Cooke, R., et al. (2011). Expert elicitation for the judgment of prion disease risk uncertainties. *Journal of Toxicology and Environmental Health, Part A, 74*(2–4), 261–285.
Wallsten, T. S., & Budescu, D. V. (1983). State of the art—encoding subjective probabilities: A psychological and psychometric review. *Management Science, 29*(2), 151–173.
Wilson, K. J. (2017). An investigation of dependence in expert judgement studies with multiple experts. *International Journal of Forecasting, 33*(1), 325–336.