# Chapter 5
# IDEA for Uncertainty Quantification

**Anca M. Hanea, Mark Burgman, and Victoria Hemming**

**Abstract** It is generally agreed that an elicitation protocol for quantifying uncertainty will always benefit from the involvement of more than one domain expert. The two key mechanisms by which judgements may be pooled across experts are through striving for consensus, via behavioural aggregation, where experts share and discuss information, and via mathematical methods, where judgements are combined using a mechanistic rule. Mixed approaches combine elements of both deliberative (behavioural) and mechanical (mathematical) styles of aggregation.

This chapter outlines a mixed-aggregation protocol called IDEA. It synthesises specific elements from several of the *classical* structured expert judgement approaches. IDEA encourages experts to Investigate, Discuss, and Estimate, and concludes with a mathematical Aggregation of judgements.

## 5.1 Introduction

Several elicitation protocols developed over the last decades have been deployed successfully in political science, infrastructure planning, volcanology, etc. (e.g. Aspinall 2010; Aspinall and Cooke 2013; Bolger et al. 2014; Cooke and Goossens 2008; O'Hagan et al. 2006). The protocols detailed in Chaps. 2 and 3 of this book (see Quigley et al. 2018 and Gosling 2018 respectively) are two of the most notable examples of structured protocols that follow thoroughly documented methodological rules. They differ in several aspects, including the way interaction between experts is handled, and the way in which experts' judgements are pooled.

The Classical (Cooke's) Model detailed in Chap. 2 of this book (Quigley et al. 2018) uses mathematical aggregation. In mathematical aggregation approaches, interaction between experts is generally limited to training and briefing (e.g. Valverde 2001; Cooke 1991), since it is believed that more interaction may induce

A.M. Hanea (✉) • V. Hemming
CEBRA, University of Melbourne, Parkville, VIC, Australia
e-mail: ahanea@unimelb.edu.au; hemmingv@student.unimelb.edu.au

M. Burgman
Centre for Environmental Policy, Imperial College London, London, UK
e-mail: m.burgman@imperial.ac.uk



L.C. Dias et al. (eds.), *Elicitation*, International Series in Operations Research & Management Science 261, https://doi.org/10.1007/978-3-319-65052-4_5

dependence between elicited judgements (e.g. O'Hagan et al. 2006), adversely affecting them. Chapter 9 of this book (Wilson and Farrow 2018) discusses the aggregation of correlated judgements in detail; here we touch on this subject very briefly.

The main advantage of mathematical aggregation is that it makes aggregation explicit and auditable. The choice of the aggregation rule is nevertheless difficult. Different rules possess different properties and it is not possible to have all desirable properties in one rule (Clemen and Winkler 1999). The Classical Model uses an unequally weighted linear pool, distinguished by the use of calibration variables to derive performance based weights. Techniques for testing and evaluating experts' performances necessarily play an important role in exploring the performance of experts. Commonly used metrics are designed to be objective. However, different metrics focus on (and measure) different attributes of performance.

Another class of methods of aggregating experts' judgements is referred to as *behavioural aggregation*, and involves striving for consensus via deliberation (O'Hagan et al. 2006). The Sheffield protocol, detailed in Chapter 3 of the book (Gosling 2018), is an example. When experts disagree, the advocates of behavioural aggregation recommend a discussion between the experts with divergent opinions, resulting in a "self-weighting" through consensus.[1] But this comes at the cost of verifiability and reproducibility. Moreover, such interaction is prone to group dynamic biases including overconfidence, polarisation of judgements and groupthink (Kerr and Tindale 2011).

*Mixed approaches* combine behavioural and mathematical aggregation techniques. The most common mixed approach is the Delphi protocol (Rowe and Wright 2001), in which experts receive feedback over successive question rounds through a facilitator, in the form of other group members' judgements. Experts remain anonymous and do not interact with one another directly. As originally conceived, the Delphi method strives to reach consensus after a relatively small number of rounds (Dalkey 1969), though in modern usages achieving consensus is not necessarily the primary aim (e.g. von der Gracht 2012). While research supports a general conclusion that Delphi methods can improve accuracy over successive rounds, this is by no means guaranteed. Critical reviews suggest that even though individual judgements may converge (von der Gracht 2012), this convergence does not necessarily lead to greater accuracy (e.g. Murphy et al. 1998; Bolger et al. 2011). Moreover, the Delphi method is widely used for the elicitation of point estimates rather than probability distributions.

The IDEA protocol described in this chapter synthesizes specific elements from all the approaches described above. In doing so, it aims to minimize the

[1]However, where a group consensus judgement cannot be reached, individual expert distributions can be elicited and combined using a mathematical aggregation technique. Or alternatively, where consensus is not the aim, the resulting spread of expert viewpoints following discussion can be maintained and presented to decision-makers (Morgan 2015).

disadvantages of existing approaches and optimise their advantages. The majority of elements that characterise IDEA are not new; its novel contribution is in the structured approach to the combination of these elements.

The reminder of this chapter is organised as follows: Section 5.2 introduces the IDEA protocol, Sect. 5.3 discusses the analysis of expert data collected using IDEA and Sect. 5.4 offers guidance for facilitators to use IDEA to elicit and quantify uncertainty.

## 5.2 The IDEA Protocol

The acronym *IDEA* arises from the combination of the key features of the protocol that distinguish it from other structured elicitation procedures: it encourages experts to *Investigate* and estimate individual first round responses, *Discuss*, *Estimate* second round responses, following which judgements are combined using mathematical *Aggregation* (Hanea et al. 2016).

An outline of the basic approach is as follows. First, experts provide private, individual estimates in response to the questions posed to them. They receive feedback in the form of the judgements of the other experts. With the assistance of a facilitator, the experts discuss their initial estimates with the others, sharing information, clarifying terms, and establishing a shared understanding of the problem. This discussion stage may take place remotely (e.g. Wintle et al. 2012; McBride et al. 2012; Hanea et al. 2016) or face-to-face (e.g. Burgman et al. 2011). During the discussion stage, ideally the anonymity of the individual estimates is maintained to counter possible unwanted dominance and halo effects. Experts are asked to revise their judgements in light of this discussion and make a second, private and anonymous estimate. These second round estimates are finally combined mathematically (see Fig. 5.1).

The motivation behind the use of the IDEA protocol is that while interaction between experts can be detrimental during the initial development of arguments and

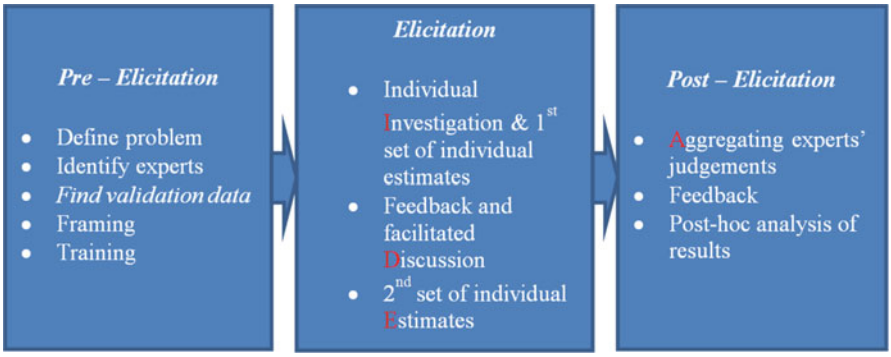


**Fig. 5.1** The IDEA protocol

responses, its use during the evaluation stage can be beneficial: allowing experts to better clarify reasoning and assumptions, and to benefit from the gains arising from well-functioning behavioural groups. The controlled interaction and feedback allow for exchange of information independent of its source, thereby removing some of the more negative aspects of behavioural groups. Using a (final) mathematical aggregation lessens the pressure for experts to reach consensus. In making their estimates for each question, experts answer using either a 4-step format for eliciting information about quantities, or a 3-step format for eliciting probabilities of binary variables (Burgman 2016). These formats draw on empirical findings from cognitive psychology and they have been shown to mitigate overconfidence (Speirs-Bridge et al. 2010; Soll and Klayman 2004).

### *5.2.1 Eliciting Probabilities*

When eliciting probabilities of binary variables (or event' occurrences), IDEA uses three questions, termed a 3-step format, one for a *best estimate* and the other two for an interval that captures uncertainty around it. The bounds are asked for before the best estimate, to get experts to think about the extreme conditions. The first two questions are prefaced with statements that urge them to think about evidence that points in one direction, and then the other, as shown in Fig. 5.2.

Other approaches, including Cooke's protocol, ask the experts to assign events to probability bins $b_i = (p_i, 1 - p_i)$, where $p_i$ corresponds to the probability of occurrence. Bins can have the following form: $b_1 = (0.1, 0.9)$, $b_2 = (0.2, 0.8)$, $b_3 = (0.3, 0.7)$, etc. if the continuous probability of occurrence scale is discretized into ten intervals. An expert assigns an event to the $b_2$ bin if their best estimate (about the probability of occurrence) is anywhere between 0.1 and 0.2. So, in a way, these approaches only ask for *best estimates*, acknowledging the imprecision in the experts' judgements by allowing a fixed interval around them (equal to the respective bin's length).

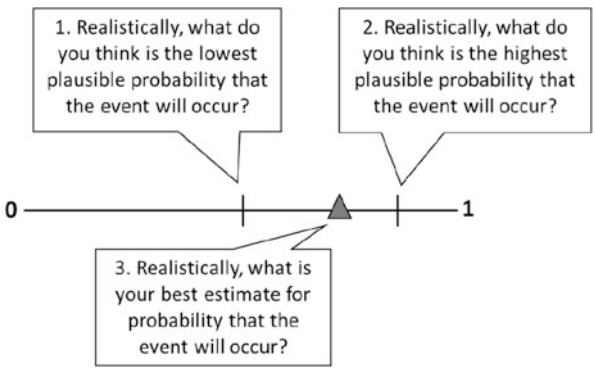


**Fig. 5.2** The 3-step format

The probabilities of binary variables can sometimes be interpreted in terms of relative frequencies. It is then legitimate to ask experts to quantify their degree of belief using a subjective distribution. In this case the upper and lower bounds asked for in the 3-step format may be thought of as quantiles of this subjective probability distribution. However, when the relative frequency interpretation is not appropriate the 3-step format may be criticised for lacking operational definitions for the upper and lower bounds. We emphasize that in such cases the bounds are elicited to improve thinking about the best estimates. They are not used in a probabilistic framework.

In both situations, if questions resolve within the time frame of the study, and using the experts' best estimates only, experts' performances can be assessed in terms of accuracy and calibration. For calibration measures, the best estimates are placed in probability bins. For example, using the notation above, best estimates between 0.2 and 0.3 are assigned to bin $b_3$. This construction allows the evaluation of calibration measures used in other protocols, e.g. Cooke's protocol. Sections 5.3.2 and 5.3.3 discuss a comparison of such measures evaluated using a dataset detailed later in this chapter.

### *5.2.2 Eliciting Quantiles of Probability Distributions*

When IDEA is used to elicit continuous quantities (continuous random variables) this procedure uses four questions to elicit the values of variables (corresponding to different quantiles), termed a 4-step format. This approach draws on research from psychology on the effects of question formats, mitigating much of the overconfidence typically observed in expert estimates (e.g. Soll and Klayman 2004; Speirs-Bridge et al. 2010).

In the 4-step format (like the 3-step format above) bounds are elicited before asking for the best estimate, to encourage experts think about extreme values, and to prevent them from anchoring on their best estimate. The first three questions are used to elicit three values of the variable, corresponding to three different quantiles, and the fourth question is used to identify the probabilities corresponding to the upper and lower quantiles specified by the experts (Fig. 5.3).

The best estimate corresponds to the median.[2] The lower and upper bounds correspond to upper and lower quantiles (denoted $q_l$ and $q_u$), such that their difference corresponds to the specified confidence level. If, for example, an expert provides a 50% confidence level, $q_l$ and $q_u$ will be taken to be the first and the third quartiles. When experts provide different confidence levels, their estimates are

[2]The best estimate may be also interpreted as the mode of the distribution. Methods for building a distribution that complies with the mode and two specified quantiles are proposed in Salomon (2013). However the interpretation of the best estimate and its use in constructing a distribution should be clearly specified prior to the elicitation.

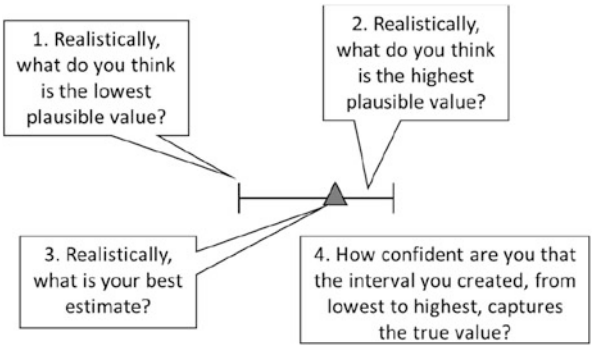


**Fig. 5.3** The 4-step format

rescaled to a consistent confidence level (e.g. 90% confidence) such that experts' distributions can be further compared and aggregated. Several methods may be used to rescale to a fixed pair of quantiles, ranging from a simple linear extrapolation to fitting a parametric distribution to the elicited quantiles and extracting the required quantiles from the fit. The sensitivity of an aggregated distribution (calculated for example as a weighted combination of individual rescaled expert distributions) to the choice of the rescaling method is assumed low (as supported by anecdotal evidence). However this topic requires additional research.

A slightly different version of this procedure, where the elicited quantiles are fixed, corresponds to the way questions are asked in the Sheffield method and in Cooke's protocol. Once rescaled to these fixed quantiles, the answers obtained using the 4-step format can be mathematically aggregated using the mathematical apparatus of Cooke's protocol.

## 5.3 Data Analysis

The IDEA protocol was refined and tested as part of a forecasting "tournament" that started in 2011 as an initiative of the US Intelligence Advanced Research Projects Activity(IARPA).[3] Five university-based research teams were involved in predicting hundreds of geopolitical, economic and military events, with the goal of finding the key characteristics of efficient protocols for eliciting and aggregating accurate probabilistic judgements. The project used real events that resolved in the near-future to test the accuracy of forecasts. Thousands of forecasters made over a million forecasts on hundreds of questions (Ungar et al. 2012; Mellers et al. 2015). The data elicited with the IDEA protocol represent the answers to a subset of the questions developed by IARPA. All questions considered correspond to Bernoulli variables of the following sort: "Will the Turkish government release imprisoned Kurdish

[3] http://www.iarpa.gov/index.php/research-programs/ace.

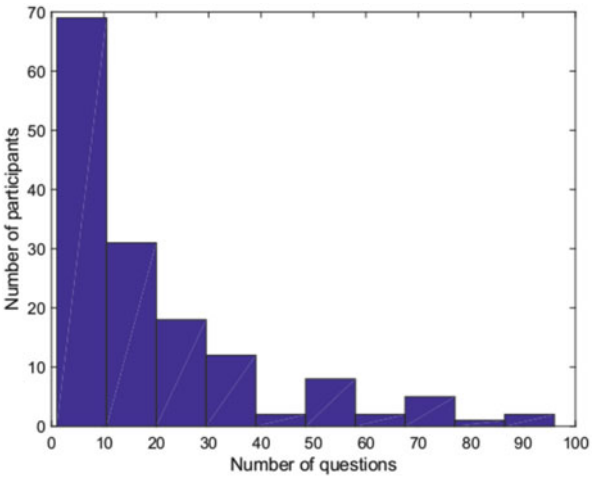


**Fig. 5.4** The number of questions answered by participants over 4 years

rebel leader Abdullah Ocalan before 1 April 2013?", which were answered using the 3-step format outlined above. All questions usually resolved within 12 months, hence they were suited for empirical validation studies. The elicitation took place remotely, initially via email, and from the second year of the tournament through a dedicated website[4] which was set up for the participants to answer the questions, discuss and upload/download necessary materials.

The tournament operated on a yearly basis, over the course of 4 years. Each year, new participants joined the IDEA group, and other participants dropped out. There were 150 participants (over the 4 years) who answered at least one question (both rounds). Eight of these participants returned each year. The level of participants' expertise covered a very wide range from self-taught individuals with specialist knowledge to intelligence analyst. A total of 155 questions were answered by at least one participant. However, no participant answered more than 96 questions. Figure 5.4 shows the distribution of the number of questions answered by the participants. The participants were divided into groups and the number of groups varied across years to keep the number of participants per group fairly constant (typically ten). Starting from the third year *Super-groups* were formed composed of the best performing participants from the previous year.[5] The number of participants composing the Super-group was equal to the number of participants from any other group.

Initial training of the participants took place before the game started. Some of the participants engaged in initial face-to-face training, where they learned about how the questions would be asked, why they were asked in this manner, and

[4] http://intelgame.acera.unimelb.edu.au/.

[5] Performance was measured using the average Brier score. This measure was imposed by the forecasting tournament rules and all participating team had to use it.

most importantly, the cognitive biases and group issues that can occur during an elicitation, and ways to mitigate them. Participants who did not receive face-to-face training, received online or telephone training. Training materials/documents that outlined and explained the issues above were also uploaded to the website for access and reference. Even though probabilistic training was not offered, many probabilistic concepts were introduced through practice questions that were part of the training.

### *5.3.1 Measures of Performance*

This section outlines some of the approaches to measuring expert performance and dependencies among experts' estimates that we have investigated for the dataset described above. Hence we restrict attention to evaluating assessments of binary variables. Experts are asked to represent their uncertainty as a subjective probability and their assessments may then be scored. Roughly speaking, a scoring rule is a numerical evaluation of the accuracy of expert assessments against actual outcomes (de Finetti 1962; Savage 1971; Winkler and Jose 2010). Despite the simplicity of this idea, there are many ways to score experts, deserving careful attention. Scoring rules are called *proper* if their expected pay-off is maximised when experts accurately express their true beliefs about the predicted event. Proper scoring rules encourage the experts to make careful and honest assessments (Winkler and Murphy 1968).

Along with evaluating individual experts' performances, we are also interested in experts' joint behaviour. Expert judgements are (in general) correlated with one another, if for no other reason, because people have access to similar information and have similar training and experiences (e.g. Booker and Meyer (1987)). This subject is discussed in Chap. 9 of this book (Wilson and Farrow 2018); here we only present the analysis of the dataset introduced above.

We are concerned with scoring as a way of rewarding those properties of expert subjective probability assessments that we value positively. We have investigated three of these properties: accuracy, calibration and informativeness.

#### 5.3.1.1 Accuracy

Accuracy measures how close an expert's best estimate is to the truth. One tool to measure accuracy is the Brier score (Brier (1950)), a proper scoring rule. The Brier score for events is twice the squared difference between an estimated probability (an expert's best estimate) and the actual outcome; hence it takes values between 0 and 2. Consider question/event $i$ with two possible outcomes $j$. The Brier score of expert $k$ assessing event $i$ is calculated as follows:

$$BrierScore_i^k = \sum_{j=1}^{2}(p_{ij}^k - x_{ij})^2,$$

where $p_{ij}^k$ is expert $k$'s probability for event $i$, output $j$, and $x_{ij}$ is 1 if output $j$ occurs and 0 otherwise. The above formula measures the accuracy of one estimate made by one expert for one question. Lower values are better and can be achieved if an expert assigns large probabilities to events that occur, or small probabilities to events that do not occur. An experts' accuracy can be then measured over many questions ($N$) and averaged to represent their overall accuracy:

$$BrierScore^k = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{2}(p_{ij}^k - x_{ij})^2$$

The number of questions and their overall sample distribution play an important role in interpreting such a score. By an overall sample distribution, we mean the inherent uncertainty of the events represented by the questions. This is also called the *base rate* and it is different for each different set of questions. However, its value contributes to the value of the average Brier score, even though it has nothing to do with the expert's accuracy. This challenges the comparison of experts' scores calculated from different sets of questions, with different base rates. Nevertheless, comparisons will be more meaningful when made on the same set of questions.

#### 5.3.1.2 Calibration

To deal with the *base rate* problem, Cooke discusses the benefits of using scores for average probabilities, rather than average scores for individual questions (variables) in Cooke (1991). He opts for calibration (which he calls *statistical accuracy*) rather than accuracy measures for evaluating experts' performance. A scoring rule is essentially a random variable and interpreting the scores' values requires knowledge about the score's distribution. An important justification for Cooke's proposal is that his (asymptotically proper) score has a known distribution, as opposed to (for example) the average Brier score, which does not. The average Brier score is a single number summary of the joint distribution of forecasts and observations. An empirical distribution of the average Brier score can be obtained for a given joint distribution of the forecasts and observations. However, this empirical distribution will differ for different joint distributions.

Before introducing Cooke's calibration score for events,[6] we need some notation. Assume the experts are asked to assign events to probability bins $b_i$. Let $p_i$ be the

[6]The calibration measure for events is based on similar concepts as the ones presented in Chap. 2 of this book (Quigley et al. 2018), when the calibration score is described for evaluating assessments about continuous variables.

probability of occurrence that corresponds to bin $b_i$. Each expert assigns events to bins. Let $n_i$ denote the number of events assigned (by an expert) to the bin $b_i$. Let $s_i$ denote the proportion of these events that actually occur; $s_i$ can be thought of as the empirical distribution of $b_i$, whose theoretical distribution is $p_i$. Ideally $s_i$ and $p_i$ should coincide. Nevertheless, in practice, they often do not. Cooke's calibration is essentially a comparison between the empirical and theoretical distributions, per bin, per expert. The discrepancy between the two is measured in terms of the relative information[7] $I(s_i, p_i)$ of $s_i$ with respect to $p_i$, defined in Chap. 2 of this book (Quigley et al. 2018). The relative information of one distribution with respect to another is a non-negative measure that equals zero iff $s_i = p_i$. Increasing values of $I(s_i, p_i)$ indicate increasing discrepancy. The relative information is calculated as follows:

$$I(s_i, p_i) = s_i \ln\left(\frac{s_i}{p_i}\right) + (1 - s_i) \ln\left(\frac{1 - s_i}{1 - p_i}\right)$$

A result in Hoel (1971) shows that for $n_i$ independent events whose probability of occurrence is $p_i$, $2n_iI(s_i, p_i)$ is asymptomatically Chi-squared distributed with one degree of freedom. Then, if ten bins are used and if all events are independent $\sum_{i=1}^{10} 2n_iI(s_i, p_i)$ is asymptomatically Chi-squared distributed with ten degrees of freedom. Under the (null) hypothesis that the experts estimate the theoretical distribution correctly, Cooke's calibration is defined as the probability of obtaining a result equal to or more extreme than the one observed. Hence, it corresponds to the p-value of a statistical test:

$$Cal(e) = 1 - \chi^2_{10}\left(\sum_{i=1}^{10} 2n_iI(s_i, p_i)\right),$$

where $\chi^2_{10}$ is the cumulative distribution function of a Chi-squared random variable with ten degrees of freedom.

For the Chi-square approximation to be reasonably close, the number of questions assessed by each expert should be quite large (hundreds). Since this is very rarely the case in practice, the empirical distribution of $\sum_{i=1}^{10} 2n_iI(s_i, p_i)$ (obtained via simulation) is used instead.

As for the average Brier score case, ideally expert performances should be compared on the same set of questions. When experts assess different questions, the power of the test used in measuring calibration should be adjusted to account for the different number of samples (the different number of questions) (Cooke 1991). Incorporating this adjustment into the simulated empirical distribution of the score is far from trivial. If a score has an exact distribution, rather than an asymptotic one, the power adjustment is not crucial.

[7]The relative information is usually known as the Kullback–Leibler divergence, or information divergence, or information gain, or relative entropy.

Using the same notation we could measure a different sort of calibration through the average Brier score discussed above. The average Brier score can be decomposed into two additive components called *calibration* and *refinement* (Murphy 1973). The calibration term for $N$ questions can be calculated as follows:

$$\sum_{i=1}^{10} \frac{n_i(p_i - s_i)^2}{N}$$

Very roughly, the refinement term is an aggregation of the resolution and the inherent uncertainty of the events assessed. The resolution term rewards expert estimates that are consistent with event probabilities. Other measures of resolution based on the notion of entropy associated with a probability mass function can be formulated. Entropy is a measure of the degree to which the mass is *spread out* and can be used in several ways to describe aspects of an expert's informativeness.

#### 5.3.1.3 Informativeness

Entropy is very often taken as a measure of lack of information in a distribution. The entropy of the distribution $(p_i, 1 - p_i)$, denoted $H(p_i)$ is calculated as follows:

$$H(p_i) = -p_i ln(p_i) - (1 - p_i) ln(1 - p_i)$$

The maximum value of $H(p_i)$ is $ln(2)$ and it is obtained when $p_i = 0.5$. Thus, the uniform distribution is the most entropic. The most informative distribution corresponds to the distributions with minimal entropy, 0. This is obtained only if $p_i = 0$ or $p_i = 1$. The entropy in the joint distribution of independent variables is the sum of entropies in the distributions of the individual variables. Two different entropy measures are defined in Cooke (1991), the average *response entropy* and the average *sample entropy*. The average response entropy in an expert's joint distribution on N events is defined as:

$$H_r = \frac{1}{N} \sum_{i=1}^{10} n_i H(p_i)$$

The response entropy measures the entropy in what the expert says. It does not depend on the actual occurrences of events. The average sample entropy, denoted $H_s$, is calculated as follows:

$$H_s = \frac{1}{N} \sum_{i=1}^{10} n_i H(s_i)$$

The sample entropy measures the entropy in the expert's performance, but it does not correspond to the distribution that the expert (or anyone else) believes. In contrast, response entropy corresponds to the distribution connected to the calibration hypothesis described above. If an expert is perfectly calibrated, then $H_s = H_r$. Unfortunately, $H_s = H_r$ does not imply perfect calibration.

An expert's informativeness may be also measured with respect to their choice of the probability bins. The choice (alone) of a more extreme probability bin (i.e. assigning a probability close to 0 or 1) can give yet another indication of the expert's informativeness. The average *response informativeness*, introduced in Hanea et al. (2016) is defined as follows:

$$I_r = \frac{1}{N} \sum_{i=1}^{10} n_i I(p_i, 0.5)$$

The response informativeness attains its minimum in 0, when all the variables are placed in the $(0.5, 0.5)$ bin. A higher informativeness score is preferred since it indicates that more variables were placed in more extreme bins.

All the formulations above assume that experts have placed events in probability bins. However IDEA asks experts to provide a best estimate and an uncertainty interval around their best estimate. In our analysis, the above measures are calculated by placing the best estimates into the bins and ignoring the upper and lower bounds. Nevertheless, the *interval' widths* can be considered as a measure of the experts' confidence, or lack thereof. A larger (smaller) interval may be interpreted as decreased (increased) confidence. Narrower bounds around a judgement are often interpreted as greater informativeness. Hence we can investigate the length of the uncertainty interval as a measure of confidence and the relationship between this measure and the measures of informativeness discussed above. These relationships are investigated in Hanea et al. (2016) for the dataset described above.

#### 5.3.1.4 Correlated Expert Judgements

Correlated expert judgements have been discussed occasionally in the literature but, to our knowledge, there has been little research on evaluating the extent to which this dependence is practically relevant. Cooke (1991) postulates that such correlation is:

> usually benign, and always unavoidable.

In contrast O'Hagan et al. (2006) worries that:

> groups of similar experts will receive too much weight and minority views will be under-represented.

Chapter 9 Wilson and Farrow (2018) of the book discusses this subject from a more general perspective. In contrast, we investigate only two particular conjectures about the dependence between the participants' answers elicited using the IDEA protocol (which permits and encourages interaction between the two

elicitation rounds). We conjecture that any additional dependence between judgements introduced through the discussion is justified by the increase in information resulting from discussion and by the reduction of misunderstandings or unintended dichotomies in responses. Moreover, this discussion takes place within groups, so our second conjecture is that the dependence structures within and between the groups are similar. If/when that is true, the expert data analysis can be (statistically) strengthened by pooling the estimates from all groups.

### *5.3.2 The Merits of Discussion*

Results on the benefits of the discussion between rounds, based on part of the 4 year dataset described earlier are presented in Hanea et al. (2016). The analysis was undertaken within groups and per year, hence the claimed benefits lack statistical power. However, the second conjecture formulated above is supported by the data analysis from Hanea et al. (2016), so we feel comfortable in pooling the expert data to form a larger dataset and hence permit more powerful statistical tests. This allows us to investigate how some of the performance scores detailed in Sect. 5.3.1 change per expert after discussion. Figure 5.5 shows pairs of four different scores (before and after discussion) corresponding to all participants who answered at least four questions. The crosses represent the average Brier scores, the diamonds represent the average confidence as measured by the length of the uncertainty intervals, and the x's represent the calibration terms of the Brier score. For all three measures low scores represent better performance. The dots represent the average response informativeness; better informativeness corresponds to larger values. The main diagonal is plotted for better visualisation. For the first three measures (Brier scores, confidence, and calibration), most of the points fall below the main diagonal, indicating better performance after discussion. For the fourth measure (informativeness), most of the points fall above the main diagonal, again indicating better performance in the second round.[8]

All the investigated measures of performance point to the value of facilitated conversations between experts in reconciling language based misunderstandings and interpretations of evidence. The relationship between these measures remains unclear in general. For this particular dataset, the authors of Hanea et al. (2017) found no, or little correlation between how accurate experts' estimates are, and how informative they are.

[8]Three quarters of the Brier scores and the average confidence scores are better in the second round, and two thirds of the calibration scores and the informativeness scores are better in the second round.

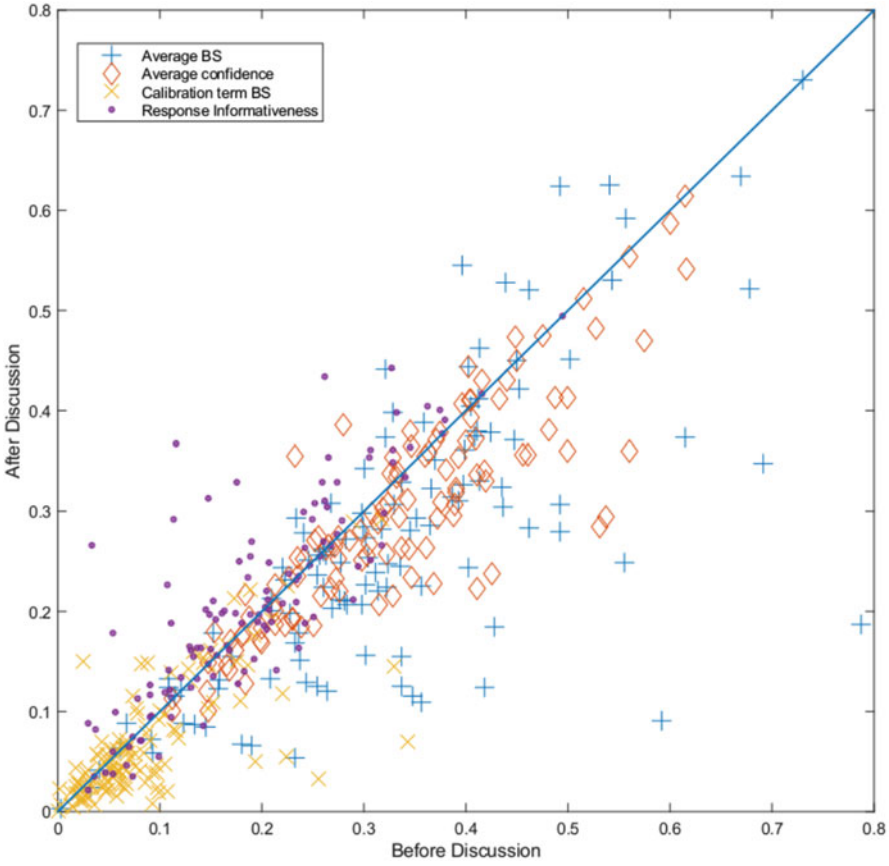


**Fig. 5.5** The average Brier scores, the average confidence,the calibration term of the Brier score and the average response informativeness of all participants, before and after discussion

### *5.3.3 Prior Performance as a Guide to Future Performance*

Each year of the tournament, we compared an equally weighted combination of all participants after the first round of opinions ("the wisdom of crowds") to the equally weighted opinions of the groups after discussion, using a within-subject design. In 1 year alone (2013–2014) we had sufficient data to calculate differential weights using Cooke's calibration score. Figure 5.6 shows the average Brier scores of the equally weighted combination of all participants' first round judgements (before discussion), compared with the equally weighted judgements of each group after discussion, together with their corresponding confidence intervals. An unequal, performance-based weighted combination of the super-group participants' judgements is shown in the same figure.

Although not statistically significant, the super-group (G1) outperformed the other groups of participants, suggesting that prior performance is a useful guide to

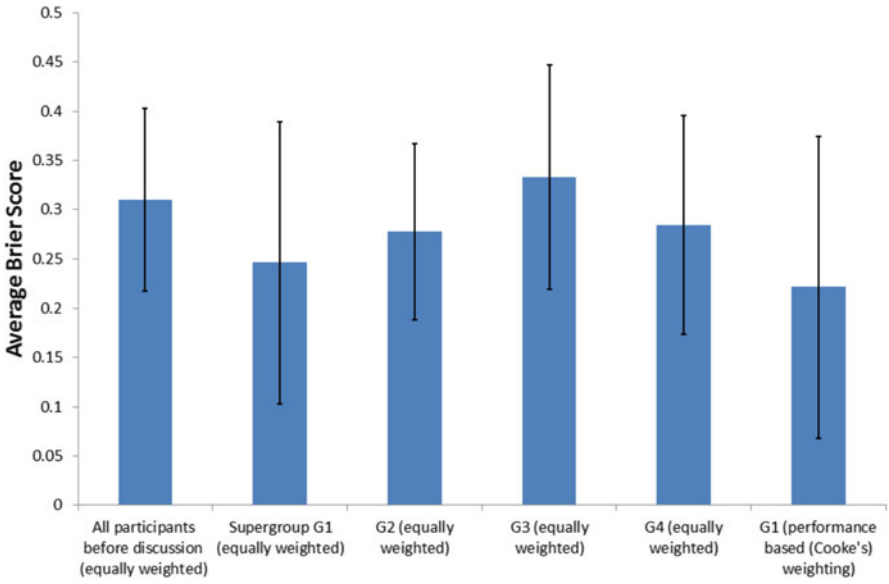


**Fig. 5.6** Forcasting tournament, year 2013–2014

future performance on similar estimation tasks. The same was observed for the forth year of the tournament. This finding is in agreement with the findings of Mellers et al. (2014). Using Cooke's calibration to derive performance based weights for an unequally weighted combination generates a slight improvement in performance.

These signals illustrate one of the most important lessons of empirical studies over the last decade: an expert's performance on technical questions may be predicted to some extent by the history of their performance on similar questions previously. Taking advantage of this phenomenon, Cooke's approach to differential weighting assimilates each expert's confidence and statistical accuracy into a single weight. The result is that group performance improves. Our results demonstrate that even in the relatively difficult conditions imposed in answering binary questions on the outcomes of geopolitical events, performance based differential weights calculated using Cooke's method improve the performance of groups, even those composed of relatively reliable forecasters.

## 5.4 A Guide to Facilitating the IDEA Elicitation Protocol

The purpose of this section is to present a summary guide for analysts and facilitators who intend to use the IDEA protocol in an uncertainty quantification exercise. Some of the recommended steps are similar to those needed when using other protocols, however, several are specific to IDEA. Much of this section has

been adapted from Hemming et al. (2017), and we suggest referring to this paper for more comprehensive advice and examples. In this section, we assume that the problem structuring, modelling, identification of data gaps and the requirements for expert input have been decided upon.

### *5.4.1 Preparing for an Elicitation*

Careful planning is necessary to ensure that experts are aware of time constraints, and that the deliverables of the elicitation become available in the time necessary. Below we briefly discuss a number of key elements to be taken into account prior to the elicitation.

#### 5.4.1.1 Key Documents

*Time-Line and Key Dates*
A list of tasks and a schedule of key dates for each of the steps of the elicitation before commencing the process is necessary. An elicitation using the IDEA protocol can take up to 6 weeks if using remote elicitation, or as little as 3 days if using a face-to-face elicitation. Additional time is required for the development of questions, recruitment of experts, approval of human research approvals, and the analysis of data. A sample timeline can be found in the supplementary material of Hemming et al. (2017).

*Human Subjects Research Ethics Approvals*
These approvals may be required, particularly if results are to be published, or to be used to inform decisions. If approval is necessary this may substantially delay the project.

*A Project Description*
This document outlines the purpose of the project, the relevant time-frames, the required expert input, and any payments. It also includes instructions on how the collected data will be used.

*A Consent Form*
A consent form should accompany the project description and be sent to experts to formalize their agreement to take part in the study and for the data to be retained and used for the specified purpose.

*Briefing Document*
The purpose of this document is to guide experts through the IDEA elicitation protocol. It should include instructions on how to answer the questions, reiterate that experts must make an initial private and anonymous estimate, whilst they are free

to talk to people outside of the elicitation group, they cannot discuss their estimates with anyone inside the group until the discussion round. Instructions should also explain the four-step or three-step format, and how their estimates will be interpreted or scored. The document should re-iterate the time-lines for the elicitation.

#### 5.4.1.2 The Questions

Even when the quantities to be elicited are identified, the elicitation questions should be framed such that the quantities to be elicited relate to potentially verifiable facts and have a clear operational meaning. Moreover, the questions should include details such as units, time-scales, and metrics. Vague, ambiguous or underspecified questions which could result in multiple interpretations should be avoided.

Ideally, one or two experts who will not participate in the elicitation should scrutinize the draft questions, ensuring (as far as possible) that the questions are fair and reasonable, within the domain of expertise of the participants, free from linguistic ambiguity or biases, and they can be completed within the allocated time-frame. The total number of questions that can be asked during an elicitation depends on the availability and the motivation of the experts. It also depends on the type (remote or face-to-face) and time-frame of the elicitation exercise. The authors of Hemming et al. (2017) suggest that no more than 20 questions should be asked within a single day of elicitation; many more can be asked if more time is available or through remote elicitation, but asking more questions may come at the cost of expert fatigue. Different settings will be detailed later in this section. When experts' judgements are aggregated using differential weighting schemes, calibration questions should be added to the set of questions.

#### 5.4.1.3 The Experts

Chapter 16 of this book (Bolger 2018) is dedicated to expert selection. We only very briefly touch upon this subject. The IDEA protocol relies on recruiting a diversity of experts. To generate a diverse group of experts, we recommend employing a range of techniques including professional network searches, peer-recommendations, on-line searches, and literature reviews. The techniques employed can have inherent biases and lead to the selection of older, well regarded individuals, or people whose ideas are in line with popular belief (often older males with a tertiary education). This may lead to a homogeneous and systematically biased group. Diversity should be reflected by variation within the group in age, gender, cultural background, life experience, education or specialisation, years of experience and position on the questions at hand.

#### 5.4.1.4 The Facilitator

A key requirement of a good facilitator regardless of the protocol they employ is that are neutral to the outcome of the elicitation, and capable of retaining objectivity. The facilitator must be competent in diplomatically handling a wide range of personalities, be able to encourage critical thinking within groups, and to pose counterfactuals.

When facilitating an elicitation using the IDEA protocol, the facilitator should be familiar with the aims and limitations of the IDEA protocol. This means they should be acutely aware of the various biases and heuristics common to expert judgement, and how elements of the IDEA protocol aim to counteract the expression of these biases. The facilitator should understand and be capable of explaining both the mathematical and the psychological theory behind the specific elicitation type and the aggregation method.

### *5.4.2 Implementing the IDEA Protocol*

#### 5.4.2.1 The Initial Meeting

The IDEA protocol commences with an initial meeting between the project team and the experts. The first project meeting is vitally important for establishing a rapport with the experts. A teleconference of approximately an hour is usually sufficient. During the meeting, the motivation for the project is introduced and the unavoidable frailties of expert judgement are explained. The motivation for a structured protocol is the desire to ensure the same level of scrutiny and neutrality is applied to expert judgement as is afforded to the collection of empirical data.

During this meeting the outline of the IDEA protocol, and the motivation behind its key steps are discussed. The format of the questions, the cognitive biases and group issues that can occur during an elicitation, and ways to mitigate them are explained. Probabilistic training may be included if experts do not have a minimum level of understanding of necessary probabilistic concepts. One rule is emphasised: the experts must not speak to one another prior to the discussion stage within the IDEA framework. However, they can and should speak to anyone else they like, and use any sources that may be relevant. We recommend going through one or two practice questions if time allows, as they help the experts familiarise themselves with the questions style and the overall process; otherwise practise questions can be incorporated subsequently. Finally, reiterate the time-lines and allow sufficient time for experts to ask questions. The supplementary material of Hemming et al. (2017) provides an example of how the project team might structure the teleconference.

#### 5.4.2.2 The Elicitation

The IDEA protocol provides a flexible approach to the elicitation of experts which enables on-line and remote elicitation, or to undertake the entire elicitation through a workshop (face-to-face). The choice of method will usually be a result of budget and time constraints, however, if the option is available then it is recommended that at the very least the discussion phase should be undertaken with use of a face-to-face elicitation.

*IDEA On-line*
The experts should be (individually) provided with the questions (including practice questions if they were not dealt with during the initial meeting), a briefing document to guide them through the elicitation process and to reiterate key steps, and training materials. The experts then create a unique codename/number which retains their anonymity in group discussions, but allows them to easily identify their own estimates. They should be sent a reminder about 3 days before the close of the first round to get their results in by the deadline. Ideally, allow 2 weeks for experts to complete the first round estimates.

Each expert sources information and consults colleagues independently, before answering the questions. Once all answers are collected, allow time for the expert data to be cleaned. If outliers or implausible values are revealed during this process, then it is best to clarify with experts whether these are true beliefs or mistakes before analysing the data.

After all the above steps are completed, a graphical output of the data should be collated and circulated among the experts. Compile the comments, rationales, re sources and links provided by the experts together with their estimates and distribute them together with the graphical output.

The discussion phase commences once experts have received the consolidated results of the first round estimates. This can be undertaken by email, a teleconference, or a web forum. The key aims of discussion are (1) to reduce linguistic uncertainty and (2) to make sure that experts have considered counter-factual explanations, contrary evidence and alternative models. The role of the facilitator is to guide and stimulate discussion but not dominate it. For example, the facilitator should pick some contrasting results and ask questions which help to determine the source of variation.

Following the discussion, facilitators should clarify meaning and/or better define the questions. If questions are reformulated or modified in any way, the new versions should be sent back to the experts, who now need to make second, anonymous and independent estimates for each. Another week or two should be allowed for the second round estimates. It is possible to ask many more questions, when elicitations run remotely (over the web or by email). People then have enough time to spread the tasks over several days.

*IDEA Face-to-Face*
Face-to-face workshops are time consuming and expensive, but they usually result in better buy-in and acceptance of the outcomes than do elicitations that are exclusively

remote. The duration of the workshop depends on the resources: it can range from 1 day to 3 days. If time allows the initial meeting can be part of the workshop, prior to training the experts, and discussing the questions to be elicited. Experts provide individual, anonymous initial estimates based on their prior knowledge and any information they can gather from the web or other immediately accessible sources.

A graphical output of the data is then collated and presented to the experts. The discussion stage starts and questions are analysed in turn. Typically, some questions are more problematic than others and require longer discussion. As above, the facilitator prompts the experts to think about alternative explanations and to reconcile different linguistic interpretations of the questions. The facilitator judges when the discussion has reached a point when no more useful contributions remain to be made and the questions are sufficiently clarified. The experts then make their second, anonymous and independent estimates for each question.

*Hybrid On-line and Face-to-Face IDEA*

Combining remote and face-to-face elicitation steps is also possible, and several options are available. A recommended combination (in case of restricted resources) is to elicit the first round estimates remotely, and then conduct face-to-face discussions and elicitation of the second round estimates during a 1 day workshop. Other combinations are nevertheless possible. Chapter 17 of this book Barons et al. (2018) presents an application of the IDEA protocol, where a 1 day face-to-face workshop was used to elicit the questions of interest, followed by a remote IDEA protocol for eliciting calibration questions.

## 5.5 Discussion

Expert judgements are part of the fabric through which scientists communicate with policy makers and decision makers. In most circumstances, the data we require for decisions are unavailable or incomplete. Expert judgements are an unavoidable part of every-day decision-making in all technical domains. Structured techniques such as those outlined here (and in the rest of the book) are perhaps surprisingly a relatively new initiative. A handful of publications in the early 1990s have been followed by a flowering of ideas, methods and empirical tests in the 2000s. Despite these developments, for the most part, scientists and decision makers alike have been satisfied with informal deliberation processes and ad-hoc methods for acquiring and combining opinions. Evidence accumulated since the 1950s in cognitive psychology especially has illuminated how subjective and unstructured deliberations are prey to a host of frailties that may substantially influence scientific estimates. Most worryingly, the scientist themselves will be unaware of these biases. Thus, these methods represent a critical advance in the place of science in decision making and policy development.

Here, we have outlined the IDEA protocol for structured expert judgement that takes several of the most promising elements of these emerging techniques,

combining them in a way that takes advantage of their strengths, and avoiding their potential weaknesses. The data presented here suggest that some of the potential flaws of this new combined approach are not serious impediments to its deployment. In particular, the potential for generating unwanted correlation structures seems to be outweighed by the improvement in the quality of individual estimates, and subsequently (aggregated) in group judgements.

We have also discussed some of the practical aspects of involving small groups in the process, face-to-face and remotely. This is especially important for the adoption of protocols by organisations such as regulatory agencies and businesses. Often, there is a need to acquire the *best possible* or *best available* expert opinion. Previously, this has been achieved by organisations going to the most highly regarded individual they can find, and using their opinion uncritically. Structured techniques outperform individuals of any status consistently and by a considerable margin. Thus, by using these techniques, organisations may discharge due diligence in decision making. The methodological details provided here ensure that their deployment can be practical and time-efficient.

## References

Aspinall W (2010) A route to more tractable expert advice. Nature 463:294–295

Aspinall W, Cooke R (2013) Quantifying scientific uncertainty from expert judgement elicitation. In: Rougier J, Sparks S, Hill L (eds) Risk and uncertainty assessment for natural hazards, chap 10. Cambridge University Press, Cambridge, pp 64–99

Barons M, Wright S, Smith J (2018) Eliciting probabilistic judgements for integrating decision support systems. In: Dias L, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgment, chap 17. Springer, New York

Bolger F (2018) The selection of experts for (probabilistic) expert knowledge elicitation. In: Dias L, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgment, chap 16. Springer, New York

Bolger F, Stranieri A, Wright G, Yearwood J (2011) Does the Delphi process lead to increased accuracy in group-based judgmental forecasts or does it simply induce consensus amongst judgmental forecasters? Technol Forecast Soc Chang 78(9):1671–1680

Bolger F, Hanea A, O´ Hagan A, Mosbach-Schulz O, Oakley J, Rowe G, Wenholt M (2014) Guidance on expert knowledge elicitation in food and feed safety risk assessment. EFSA J 12(6):3745

Booker JM, Meyer M (1987) Sources of correlation between experts: empirical results from two extremes. Los Alamos National Lab., NM (USA); Nuclear Regulatory Commission, Washington, DC (USA). Office of Nuclear Regulatory Research

Brier G (1950) Verification of forecasts expressed in terms of probability. Mon Weather Rev 78:1–3

Burgman M (2016) Trusting judgements: how to get the best out of experts. Cambridge University Press, Cambridge

Burgman M, Carr A, Godden L, Gregory R, McBride M, Flander L, Maguire L (2011) Redefining expertise and improving ecological judgment. Conserv Lett 4:81–87. doi:10.1111/j.1755-263X.2011.00165.x

Clemen R, Winkler R (1999) Combining probability distributions from experts in risk analysis. Risk Anal 19:187–203

Cooke R (1991) Experts in uncertainty: opinion and subjective probability in science. Environmental ethics and science policy series. Oxford University Press, Oxford

Cooke R, Goossens L (2008) TU Delft expert judgment data base. Reliab Eng Syst Saf 93(5): 657–674

Dalkey N (1969) An experimental study of group opinion: the Delphi method. Futures 1:408–426

de Finetti B (1962) Does it make sense to speak of 'good probability appraisers'? In: Good, J. (ed) The scientist speculates: an anthology of partly baked ideas. Basic Books, New York, pp 357–363

Gosling, J (2018) SHELF: the Sheffield elicitation framework. In: Dias L, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgment, chap 4. Springer, New York

Hanea A, Burgman M, McBride M, Wintle B (2017) The Value of performance weights and discussion in aggregated expert judgements. Risk Anal (re-submitted June 2017)

Hanea A, McBride M, Burgman M, Wintle B (2016) Classical meets modern in the IDEA protocol for structured expert judgement. J Risk Res doi:10.1080/13669877.2016.1215346 (Available online 9 Aug)

Hanea A, McBride M, Burgman M, Wintle B, Fidler F, Flander L, Manning B, Mascaro S (2016) Investigate discuss estimate aggregate for structured expert judgement. Int J Forecast doi:10.1080/13669877.2016.1215346 (Available online 8 June)

Hemming V, Burgman M, Hanea A, McBride M, Wintle B (2017) Preparing and implementing a structured expert elicitation using the IDEA protocol. Methods Ecol Evol, Accepted on 20.07.2017

Hoel P (1971) Introduction to mathematical statistics. Wiley, New York

Kerr N, Tindale R (2011) Group-based forecasting?: a social psychological analysis. Int J Forecast 27:14–40

McBride M, Garnett S, Szabo J, Burbidge A, Butchart S, Christidis L, Dutson G, Ford H, Loyn R, Watson DM, Burgman M (2012) Structured elicitation of expert judgments for threatened species assessment: a case study on a continental scale using email. Methods Ecol. Evol. 3: 906–920

Mellers B, Ungar L, Baron J, Ramos J, Gurcay B, Fincher K, Tetlock P (2014) Psychological strategies for winning a geopolitical forecasting tournament. Psychol Sci 25(4):1106–1115

Mellers B, Stone E, Atanasov P, Rohrbaugh N, Metz S, Ungar L, Bishop M, Horowitz M, Merkle E, Tetlock P (2015) The psychology of intelligence analysis: drivers of prediction accuracy in world politics. J Exp Psychol Appl 21:1–14

Morgan MG (2015) Our knowledge of the world is often not simple: policymakers should not duck that fact, but should deal with it. Risk Anal 35:19–20. doi:10.1111/risa.12306

Murphy A (1973) A new vector partition of the probability score. J Appl Meteorol 12(4):595–600

Murphy M, Black N, Lamping D, Mckee C, Sanderson C (1998) Consensus development methods and their use in clinical guideline development. Health Technol Assess 2(3):1–88

O'Hagan A, Buck C, Daneshkhah A, Eiser J, Garthwaite P, Jenkinson D, Oakley J, Rakow T (2006) Uncertain judgements: eliciting experts' probabilities. Wiley, London

Quigley J, Colson A, Aspinall W, Cooke R (2018) Elicitation in the classical method. In Dias L, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgment, chap 2. Springer, New York

Rowe G, Wright G (2001) Expert opinions in forecasting: the role of the Delphi technique. In: Principles of forecasting: a handbook for researchers and practitioners. Kluwer Academic Publishers, Norwell, pp. 125–144

Salomon Y (2013) Unimodal density estimation with applications in expert elicitation and decision making under uncertainty. Ph.D. thesis, Department of Mathematics and Statistics, The University of Melbourne

Savage L (1971) Elicitation of personal probabilities and expectations. J Am Stat Assoc 66: 783–801

Soll J, Klayman J (2004) Overconfidence in interval estimates. J Exp Psychol Learn Mem Cogn 30:299–314

Speirs-Bridge A, Fidler F, McBride M, Flander L, Cumming G, Burgman M (2010) Reducing overconfidence in the interval judgments of experts. Risk Anal 30:512–523

Ungar L, Mellers B, Satopaa V, Baron J, Tetlock P, Ramos J, Swift S (2012) The good judgment project: a large scale test of different methods of combining expert predictions. In: AAAI fall symposium series. (AAAI Technical Report FS-12-06)

Valverde L (2001) Expert judgment resolution in technically-intensive policy disputes. In: Assessment and management of environmental risks. Kluwer Academic Publishers, Norwell, pp 221–238

von der Gracht H (2012) Consensus measurement in Delphi studies: review and implications for future quality assurance. Tech Forcasting Soc Chang 79:1525–1536

Wilson K, Farrow M Combining judgements from correlated experts. In: Dias L, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgment, chap 9. Springer, New York (2018)

Winkler R, Jose V (2010) Scoring rules. In: Wiley encyclopedia of operations research and management science. Wiley, New York

Winkler R, Murphy A (1968) "Good" probability assessors. J Appl Meteorol 7:751–758

Wintle B, Mascaro M, Fidler F, McBride M, Burgman M, Flander L, Saw G, Twardy C, Lyon A, Manning B (2012) The intelligence game: assessing Delphi groups and structured question formats. In: Proceedings of the 5th Australian security and intelligence conference