Expert Knowledge Elicitation: Subjective but Scientific

Anthony O'Hagan

University of Sheffield

February 27, 2018

Abstract

Expert opinion and judgement enter into the practice of statistical inference and decision-making in numerous ways. Indeed, there is essentially no aspect of scientific investigation in which judgement is not required. Judgement is necessarily subjective, but should be made as carefully, as objectively, as scientifically as possible.

Elicitation of expert knowledge concerning an uncertain quantity expresses that knowledge in the form of a (subjective) probability distribution for the quantity. Such distributions play an important role in statistical inference (for example as prior distributions in a Bayesian analysis) and in evidence-based decision-making (for example as expressions of uncertainty regarding inputs to a decision model). This article sets out a number of practices through which elicitation can be made as rigorous and scientific as possible.

The principal focus is on the cognitive biases that experts are prone to when making probabilistic judgements, individually or in groups, and on how these can be addressed and minimised by a well-designed elicitation protocol. This is illustrated by the SHELF protocol.

1 Introduction

This article is based on a presentation I made to the October 2017 Symposium on Statistical Inference, organized and sponsored by the American Statistical Association, in a session devoted to the role of expert opinion and judgement in statistical inference and decision-making. From the presentations and discussions in that session, a position paper by the convenor and speakers in the session presents their view that

> ... we must accept that in practice there is subjectivity in every stage of a scientific enquiry, but objectivity is nevertheless the fundamental goal. At every stage, we should seek to be as objective as we can be, to base opinions and judgements on evidence and careful reasoning, and wherever possible to eradicate bias, prejudice, sloppy thinking, etc.
>
> We are firmly of the (subjective!) opinion that to recognise these facts openly can only be beneficial to the progress of science. Scientific activity encompasses many activities beyond the formal scientific method, but the rigour of the scientific method is an ideal towards which all such activity should strive. And scientific activity includes numerous components that involve subjective opinions and judgements, but objectivity is an ideal towards which all such judgements should strive. (Brownstein et al, 2018)

One way in which expert opinion and judgement enters into statistical inference and decision-making is through expert knowledge elicitation. Elicitation is the process of expressing expert knowledge in the form of probability distributions for uncertain quantities. It is increasingly used to quantify uncertainty about parameters in models that are applied in fields as diverse as medicine,

the environment, science and engineering. In the context of the Symposium on Statistical Inference, we can think of the elicited probability distributions as forming prior distributions for Bayesian statistical inference, or as directly informing decision-making under uncertainty.

In this article, I set out how I believe we can be as objective as possible when conducting expert knowledge elicitation. The goal is to formulate expert knowledge as accurately as possible in the form of a probability distribution, which in turn requires considerable expertise. An extensive review on elicitation can be found in O'Hagan et al (2006), while O'Hagan (2012) provides a more concise account of many important aspects of conducting an elicitation with scientific rigour. In this article, I focus on some ways to minimise cognitive biases in elicitation. Section 2 introduces the psychological literature on heuristics and biases, highlighting those that are of particular relevance in eliciting probability distributions. In Section 3, issues around the use of multiple experts are considered, including psychological hazards in elicitation from groups of experts. The SHELF protocol is presented in Section 4, with emphasis on how it is designed to address and minimise the biases identified in Sections 2 and 3. Finally, Section 5 reviews the principal ideas in this article and places them in the wider context of the use of expert opinion and judgement in scientific inference and evidence-based decision-making.

2 Biases in elicitation

Theories of subjective probability, in which probability is characterised as a personal degree of belief, date back at least to Ramsey (1926), with important contributions by de Finetti (1937, 1970) and Savage (1954). The idea of formally representing uncertainty using subjective probability judgements to inform decision-making began to be taken seriously in the 1960s and 1970s.

However, at about the same time psychologists were identifying problems in the ways that people make judgements. The ground-breaking research of Tversky and Kahneman (1974) set in motion the *heuristics and biases* research programme, the underlying principle of which is that people's judgements are often made on the basis of heuristics, which are quick, short-cut reasoning processes. These heuristics have served us and our ancestors well in countless everyday situations where we do not have time or need for prolonged thought, but the instinctive application of heuristics can lead to systematic errors of judgement (biases) in more complex, less everyday tasks. The conclusion is that for serious decision-making we need to consciously slow down, taking time to think. Kahneman (2011) expounds in detail this theory of fast and slow thinking.

Of the large number of heuristics that have been identified in the psychology literature, a few are of particular relevance to the judgement of probabilities, and hence to elicitation.

2.1 Anchoring

One of the original heuristics of Tversky and Kahneman (1974) was 'anchoring and adjustment': when asked to make a numerical judgement, people start with a readily available value and adjust from that point to assign a judgement. The initial value is called the anchor, because typically the adjustment is insufficient and so the resulting judgement is biased towards the anchor. The canonical case is when two successive judgements are made of related quantities; the first judgement acts as an anchor for the second. An example of this in the case of probability judgements is discussed in Section 2.4.

To illustrate a different form of anchoring in probability judgement, I have conducted an experiment on a sample of participants on courses I have given over the past four years. This is admittedly not a proper random sample from a

well defined population of subjects, but randomisation was applied within each course.

The participants were asked to make judgements about the number of Muslims in England and Wales according to the UK census of 2011. They were given the following information.

> In the 2011 United Kingdom Census, people were asked to specify their religion. Of the 52.0 million people in England and Wales who gave an answer (excluding 4 million non-respondents), 33.2 million stated that they were Christian (which includes Catholic and various Protestant denominations). Other possible answers (18.8 million people in total) were Buddhist, Hindu, Jewish, Muslim, Sikh, "Other religion", "No religion". In this exercise, you are asked to assess probabilities concerning the number, M, of Muslims amongst those 18.8 million.

They were asked to provide two probabilities, that M was more than 8 million and that it was more than 2 million. There were two versions of the exercise: in one version they were asked for $P(M > 8)$ first, and in the other version the first judgement was $P(M > 2)$. They did not see the second question until they had answered the first (and could not go back and revise their first answer). In each course, the allocation of versions to participants was randomised. I gave ten courses over four years, to participants from widely varying backgrounds. For instance, in some cases they were predominantly British, in two cases they were all from Nordic countries, and in two more mostly Italian. I have data from approximately 90 participants on each version of the exercise. Their average probabilities are given in Table 1.

	2 million first	8 million first
$P(M > 2 \text{ million})$	0.696	0.804
$P(M > 8 \text{ million})$	0.303	0.403

Table 1. Average probability judgements in the Muslims exercise

The effect of anchoring is seen clearly here. Participants who received the $P(M > 8)$ question first on average give higher probability judgements on *both* questions. Putting the number 8 million in their heads serves as an anchor, suggesting higher values for M than if the anchor is 2 million. Since the data are not from a proper random sample, any formal statistical analysis can only be indicative, but the differences are quite clear. In fact the same feature, of higher average probabilities on both questions for participants who received the version with $P(M > 8)$ first, was observed separately in every one of the ten courses, with the sole exception of just one inversion in one course.

2.2 Availability

According to the availability heuristic, also originally identified in Tversky and Kahneman (1974), an event is judged more probable if we can quickly bring to mind instances of it occurring. For instance, I will judge deaths from a particular disease to be more common if I know of people who suffer from, or have died from, this disease. Dramatic events are more memorable, and so are more easily brought to mind. Thus, aircraft and train crashes make headline news, and will influence a person's judgement of risk with these modes of transport. Car crashes are, in contrast, far more common but are rarely newsworthy, so their probability tends to be underestimated.

More generally, when experts are asked about an uncertain quantity, their

judgements will be more influenced by evidence that they can readily bring to mind. This will tend to be recent evidence and evidence that they have personally been involved in deriving. Less salient evidence is likely to be overlooked.

2.3 Range-frequency

If we divide the possible values of an uncertain quantity into a number of categories, and ask experts to assign a probability to each category, then they will tend to spread the total probability more or less evenly between the categories. Their stated probability judgements are then a compromise between this even spread and their actual beliefs about how probable each category is. This *range-frequency compromise* is related to a more general heuristic of the same name (Parducci, 1963), and leads to a bias that assigns less probability to the categories judged most likely, and more probability to the other categories.

To illustrate this bias, I carried out another experiment with participants in my most recent training course. They were asked to make judgements about the total number, C, of medals that China will win in the next summer Olympic Games. They were given a table of the number of medals won by China at all previous Summer Olympics, and they were asked to assign probabilities to C being 'more than 119', 'between 110 and 119', 'between 100 and 109', and so on. There were two versions of the exercises which again were randomly assigned to participants. One version had as its final category C being 'less than 60', while the other split this into two final categories, 'between 50 and 59', and 'less than 50'. Thus, one version had 8 categories, while the other had 9.

The range-frequency compromise suggests that the participants with the first version of the exercise will assign a smaller probability to the single category 'less than 60' than the total probability given by participants with the second version to their final two categories combined. This is indeed what I found,

with an average probability of 0.077 given to the single category 'less than 60', compared with an average total probability of 0.104 given to the split category. This was a small sample, and certainly does not prove that the range-frequency compromise is operating as expected in this example, but it is supportive of that principle and I hope to be able to establish it firmly in future courses.

2.4 Overconfidence

It is often said that experts are typically overconfident in their judgements. The evidence for this is primarily from studies in which subjects were asked to give an interval of values for an uncertain quantity with a specified probability, and where the frequency with which the true values fell within those intervals was less than the specified probability. For instance, in studies where subjects gave 95% probability intervals, fewer than 95%, and perhaps as few as 65% of those intervals were found to contain the corresponding true values. The intervals exhibited overconfidence.

A number of possible explanations can be advanced for this finding.

- Although not a heuristic in its own right, overconfidence may be related to anchoring. Subjects are often asked for an interval after they have first given an estimate, which can serve as an anchor. Then the interval being too narrow may be due to insufficient adjustment from the anchor.

- For experiments, the true values of quantities must be known to the researchers but not to the subjects. Some studies have asked about 'almanac' quantities, i.e. ones whose values might readily be found in a reference book or online, such as populations of cities, lengths of rivers, heights of buildings. In selecting quantities, the researchers may unconsciously choose ones which are 'more interesting', which could equate to them having unexpected values. If so, the low coverage frequency of the subjects'

intervals may be quite reasonable, and not evidence of overconfidence.

- When the subjects are presented as experts, they may feel under some pressure to demonstrate their expertise by giving narrower intervals than their knowledge would really justify. Equally, though, experts may give wider intervals when they fear there may be consequences to 'getting it wrong', thereby expressing underconfidence.

- It has been suggested that expertise is associated with experts developing their own short-cut heuristics to reach quick answers to questions commonly arising in their field. An instinctive reliance on such specialist heuristics may manifest as overconfidence, because in an elicitation exercise they are likely to be asked about less routine quantities.

- To assess a 95% interval, a subject must judge the probability of being outside that interval as only 5%, but it is difficult to judge small probabilities in this way.

The significance of the last point is that if an expert is asked to give a 95% interval they simply think in terms of giving a range of values in which they judge the quantity of interest is very likely to lie. They would give the same range if asked for a 99% interval, because 'very likely' is an imprecise term. To illustrate this, I conducted another experiment with my most recent course participants. In this exercise, they were given the following information.

> You are a taxi driver and you need to drive a customer from the town of Great Missenden to Terminal 5 of London's Heathrow Airport, to arrive no later than 17:45 on Friday evening. The journey comprises 11 miles on A roads, followed by 11 miles on motorways, including 5 miles on the M25. According to Google maps, the journey time in "usual traffic" should be 33 minutes, but you need to

arrange to pick the customer up early enough so that there is very little risk of not arriving by 17:45. Let the journey time on Friday be T (minutes).

Participants were then asked to specify an upper 95% bound (a value they were 95% confident that T would be below) and an upper 'plausible' bound (that they were 'almost sure' T would be below). As in the other exercises, there were two versions in which either the 95% or plausible bound was requested first, and they did not see the second question until they had answered the first. Table 2 shows the average values of their bounds.

	95% bound first	Plausible bound first
Upper 95% bound	93 mins	67 mins
Upper plausible bound	126 mins	84 mins

Table 2. Average bounds in the taxi journey exercise

Again, this is a single small sample, but it is striking that the participants gave an average of about 90 minutes for their first upper bound, whether they were asked first for a 95% bound or a plausible bound. If they had first given a 95% bound, to give a plausible bound they adjusted their first value upwards, whereas to go from a plausible to a 95% bound they adjusted downwards (adjustments which, because of anchoring, we may expect to have been inadequate!). I hope to be able to replicate this finding in future courses.

3 Multiple experts

When expert knowledge is sought, it is usual to seek judgements from more than one expert. Nevertheless, we generally require the outcome to be a single

probability distribution representing the combined knowledge of experts in the field. Resolving the experts' judgements into a single distribution is known as the problem of aggregation. There are two principal approaches.

- *Mathematical aggregation.* In this approach, also known as *pooling*, separate judgements are elicited from the experts and a probability distribution is fitted to each expert's judgements. These are then combined into the aggregate distribution using a mathematical formula (a pooling rule).
- *Behavioural aggregation.* In contrast, the behavioural approach asks the group of experts to discuss their knowledge and opinions, and to make group 'consensus' judgements, to which an aggregate distribution is fitted.

Neither approach is without disadvantages. Mathematical aggregation requires a choice of pooling rule, and numerous such rules have been proposed. In order to limit the choice, we might ask for the rule to have desirable properties, but French (1985) offers two reasonable consistency criteria and reports that no pooling rule can satisfy both.

The behavioural approach has the difficulty of persuading experts with differing opinions to reach 'consensus'. It is also open to various additional hazards that have been identified by psychologists. There are clearly problems associated with the personalities of those whose opinions may be sought as experts. A strong personality may dominate the group discussion and consensus judgements, without necessarily having a dominant position in terms of knowledge and expertise. Conversely, the judgements of a quieter and less extrovert expert may be ignored or overlooked. Even if one expert does not dominate the group, two or more like-minded experts may do so together.

Another feature of expert groups, which might be called a group heuristic, is a tendency for discussion to be restricted to ideas that will be broadly acceptable to all the group members, a behaviour known as *groupthink* (Janis, 1972). In the

context of behavioural aggregation in elicitation, this term describes a tendency for a consensus view to emerge that is overconfident, because of the very act of seeking consensus.

In a substantial guidance document on eliciting expert knowledge, EFSA (the European Food Safety Authority, 2014) recommends three protocols for elicitation with multiple experts. The Cooke protocol (Cooke, 1991) employs mathematical aggregation, but the pooling rule weights the experts according to their performance in judgements about a number of 'seed' variables whose true values are known to the elicitor. The Sheffield protocol in the EFSA guidance is a particular case of the SHELF protocol described here in Section 4, and employs behavioural aggregation. EFSA's third protocol is the classic Delphi method (Rowe and Wright, 1999), but adapted to elicit judgements of uncertainty rather than simply estimates. It is known as EFSA Delphi or probabilistic Delphi, and has features of both mathematical and behavioural aggregation. There is some interaction and sharing of knowledge between experts, as in behavioural aggregation, except that the interaction is strictly limited (with the intention of avoiding problems such as dominant personalities), while after this interaction it is necessary to apply a pooling rule to aggregate across the experts' final distributions.

A technical question that can be directed to the aggregate distributions derived from all of these methods is, "Whose probability distribution is this?" The point of the question is that the reason for seeking an aggregate distribution is to represent uncertainty about the quantity of interest, and for this purpose it should be interpretable as an expression of a person's subjective beliefs. It is clear that the result of mathematical aggregation is not the belief of any individual. It might be claimed that the result of a behavioural aggregation represents the beliefs of the group, but it is far from clear that the group has

any such beliefs when the distribution is more the outcome of compromise than consensus.

4 The SHELF protocol

SHELF is a package of materials developed by Jeremy Oakley and myself to guide and assist the conduct of an expert knowledge elicitation (Oakley and O'Hagan, 2016). Within the guidance, there are five elements (as set out in the "SHELF Overview" document in the SHELF package) which are prescriptive; provided that all these essential elements are properly followed, an elicitation is said to have been conducted according to the SHELF protocol. The following subsections describe these essential elements with particular reference to how they are designed to address the cognitive biases and challenges identified in Sections 2 and 3. Further information, including the many other aspects of an elicitation for which SHELF provides less prescriptive guidance, can be found in the SHELF package itself and in Gosling (2018).

4.1 SHELF templates

The templates in the SHELF package organise the process of the elicitation though a predefined series of steps, and the first essential element of the protocol is to follow the templates. In particular, elicitation of a probability distribution for a single quantity of interest (denoted here by X) from a single expert must follow a defined series of judgements, although there are a small number of alternative sequences, known as 'methods'. For instance, the tertile method has the following sequence.

1. *Plausible range.* The expert is asked to specify a lower plausible bound L and an upper plausible bound U for X, such that it would be very

surprising if the true value were found to lie outside the interval $[L, U]$. It is not important whether the expert interprets this interval as being a 95% interval, a 99% interval or an 'almost sure' interval. Its function is to establish that some values of X are judged by the expert to be simply not plausible. By beginning with this judgement, the expert is encouraged to think at the outset about the full range of possibility for X, which helps to counter overconfidence.

2. *Median.* The expert is next asked to specify their median value M, such that in their judgement X is equally likely to lie above or below M. Notice that the only numbers that have been suggested have come from the expert, rather than being given to the expert in a way that might create anchors. In judging M, the expert has anchors at L and U, and their anchoring effects should cancel out so as to have minimal impact on the expert's judgement of M.

3. *Tertiles.* The exert now specifies their tertiles $T1$ and $T2$, such that the expert judges it to be equally likely for X to be below $T1$, between $T1$ and $T2$, or above $T2$. Again, any anchoring effect on these judgements of the previously specified values L, U and M should be minimal.

The judgements of median and tertiles are not explicitly judgements of probability but simply ask the expert to identify equally likely ranges of possible values of X. Nevertheless, they are not easy for experts who have not encountered these ideas before. Training and careful explanation are essential. For every judgement that experts are asked to make within the SHELF protocol, the SHELF package includes a PowerPoint slide set to guide the expert in making, challenging and refining their judgements. Furthermore, an online e-learning course has recently been developed for experts to train in making these judgements, accessible from http://www.tonyohagan.co.uk/shelf/ecourse.html.

The SHELF quartile method simply replaces the judgement of tertiles with quartiles $Q1$ and $Q3$. With M, they divide the range of possible values into four equally likely parts. In practice, the expert is asked to specify $Q1$ so that they judge X to be equally likely to fall below $Q1$ or between $Q1$ and M, and to specify $Q3$ such that X is equally likely to be between M and $Q3$ or above $Q3$. This method addresses overconfidence and anchoring in the same way as the tertile method.

Another SHELF method that is of interest in the context of the present article is the roulette method. After a first step of specifying the plausible range, the expert is then presented with a tableau comprising a series of N 'bins', where N is typically about 10, each representing a range of possible values of X so that the bins partition the plausible range into N parts. The expert places counters in the bins to show the relative probability that X will fall in each bin. For instance, if the expert places 1 counter in the first bin and 3 counters in the next, he or she is making the judgement that X is three times as likely to lie in the second bin compared to the first. This method has the advantage that the experts find is simple to use, particularly if they are familiar with the concept of a probability density function. However, some bias due to the range-frequency compromise may be expected, as shown for instance in the example in Section 2.3.

In addition to prescribing a protocol to minimise cognitive biases, the SHELF templates serve another important purpose, which is to document the elicitation exercise. By completing the templates, the elicitor provides a traceable account of how the final consensus probability distributions were reached. Documentation like this represents good scientific practice, and allows any recipients of the elicited distributions to evaluate the extent to which they should adopt them for purposes of inference or decision-making.

4.2 The SHELF workshop

It is often important to elicit the knowledge of several experts, and then the SHELF method employs behavioural aggregation. One of its essential elements is simply bringing the experts together to discuss relevant evidence and share opinions in order to deliver some kind of agreed aggregate probability distribution. The nature of this agreed aggregate is another important element of the SHELF method, see Section 4.4. A SHELF elicitation workshop is ideally a physical meeting where the discussions can take place face-to-face, but this is not always achievable. The next best option is for some participants to be linked to the physical workshop by video-conference. Although it becomes harder to manage the discussion so that every expert is able to participate fully, video-conferencing provides a reasonable alternative.

4.3 The facilitator

SHELF is based on behavioural aggregation because we believe it makes better use of the combined knowledge and expertise of the experts, but Section 3 has identified a number of additional challenges associated with the behavioural approach. For this reason, the SHELF workshop must be led by a facilitator, who has expertise in the process of eliciting expert knowledge, and in particular is familiar with SHELF. The facilitator works with the experts to obtain accurate judgements of their knowledge and manages the group discussion. An experienced facilitator will be familiar with the possible sources of bias in behavioural aggregation, and needs to be constantly alert. A checklist for managing the discussion would include the following.

- Allow discussion all the while it seems to be developing ideas. Don't let the experts keep repeating the same arguments.

- Make sure all opinions are heard and properly considered. Keep bringing quieter members into the discussion, and curtail the more voluble experts if necessary.
- Don't allow arguments to be presented aggressively.
- Listen carefully. Try to get a sense of the strengths of competing arguments. Two people giving the same argument doesn't make it twice as valid.
- Highlight significant points of the discussion. It can be useful to write these on whiteboard or flipchart.

The facilitator's role may also be found in other elicitation protocols, but it is particularly important in SHELF.

4.4 The rational impartial observer (RIO)

Another essential element in the SHELF method is the way that a 'consensus', aggregate distribution is reached. Even after discussing and debating, experts will not reach complete agreement (such that they now have the same knowledge and beliefs about an uncertain quantity, represented by the same probability distribution). Their opinions may be modified by the discussion, but they will inevitably leave the workshop with differing beliefs about the quantity of interest. If, during the workshop, they are coerced into a consensus, this will be achieved by compromise and negotiation, rather than representing a true convergence of opinion.

In the SHELF method, the experts are asked to judge what a *rational impartial observer*, known informally as RIO, might reasonably believe, having seen their individual judgements and listened to their discussion. Experts are advised that RIO would not completely agree with any one expert, but would

see merit in the opinions of all the expert. Some arguments will have been more persuasive in the discussion, so that RIO would give more weight to some opinions than others, and it is the experts themselves who are best able to judge this.

Perhaps surprisingly, by taking the perspective of RIO, in practice experts can generally reach agreement on a distribution that represents a rational impartial view of their combined knowledge. Although RIO is a fictional person, an abstraction, the aggregate distribution does at least notionally represent the beliefs of a person. Furthermore, it is RIO's beliefs, taking into account the combined knowledge, expertise and reasoning of the group, that in effect are being sought when we conduct an elicitation with multiple experts.

4.5 Individual elicitation – discussion – group elicitation

Perhaps the most important of the essential elements of SHELF is what comes before the group discussion and elicitation of an aggregate distribution. A SHELF elicitation has two rounds of judgements. It begins by eliciting judgements and a distribution from each expert separately. This happens before any discussion. Once all the experts have made their personal judgements, all the resulting distributions are shown to the group. This stage provides the platform for discussion, because it is natural to focus the discussion on areas of disagreement: why does this expert believe that X could be much larger than the other experts? why is that expert so confident that X will lie in that narrow range?

The RIO-based group judgements form the second round which completes the elicitation. The individual judgements from the first round are now an important benchmark for the facilitator to assess the reasonableness of the final aggregate. The extent to which the final distribution gives more weight to some X values, and less weight to others, than the various experts' initial judgements

should be justified by the intervening discussion. If one expert's opinion appears to have been lost in the final distribution, and yet that expert's arguments were not undermined in the discussion, the facilitator might be concerned that this expert has simply given up. Anomalies of this kind can be a reason for the facilitator to challenge the group judgements.

4.6 The evidence dossier

Although not currently listed as an essential element of SHELF, much emphasis is placed in the guidance on preparing an evidence dossier. This is a summary of relevant evidence, and is made available to the experts before and during the workshop. It serves the important purpose of minimising availability bias, by ensuring that all the evidence is fresh in the experts' minds.

5 It isn't cheap

The SHELF protocol outlined above demands substantial input of resources, particularly the time of the experts, the facilitator and those organising the elicitation. Other widely-recommended protocols require similar resource commitments. Elicitation should not be seen as a cheap option. However, the result is information, elicited from experts in a rigorous and scientific manner. To obtain experimental data of the same information content and quality will almost invariably require much more resource.

The purpose for which the information is sought may not require such high standards, and in that case it would be possible to reduce and simplify the process. Judgements may be sought from fewer experts, possibly just one, and the elicitation may be done without the aid of a trained facilitator. For instance, if the purpose is Bayesian statistical inference, if the elicitation is of prior distributions and if the data will provide strong information, then formal,

rigorous elicitation may be needed for only a small number of parameters, and perhaps for none. Similarly, if the elicited distributions will be for inputs to a decision model, the decision may be insensitive to the uncertainty in some of those inputs, in which case it will not be necessary to elicit them so carefully. A process of 'minimal assessment' is described in the EFSA guidance (European Food Safety Authority, 2014).

References

Brownstein, N., Louis, T., O'Hagan, A. and Pendergast, J. (2018). The role of expert opinion and judgement in statistical inference and evidence-based decision-making. In submission.

Cooke, R. M. (1991). *Experts in Uncertainty: Opinion and subjective probability in science*. Oxford: Oxford University Press

de Finetti, B. (1937). Foresight: its logical laws, its subjective sources. Reprinted (and translated from the original French) in *Studies in Subjective Probability*, H. E. Kyburg, Jr. and H. E. Smokler (eds). New York: Wiley, 1964.

de Finetti, B. (1970). *Theory of Probability: a Critical introductory treatment* (translation by A. Machi and A. F. M. Smith, 1974–5), 2 volumes. Wiley.

European Food Safety Authority (2014). Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment. *EFSA Journal* 2014, 12(6):3734.

French, S. (1985). Group consensus probability distributions: a critical survey. In *Bayesian Statistics 2*, J. M. Bernardo et al (eds.), 183–202. Oxford: Oxford University Press.

Gosling, J. P. (2018). SHELF: the Sheffield elicitation framework. In *Elicitation: the science and art of structuring judgement*, L. C. Dias, A. Morton and J. Quigley (eds.), 61–93. Springer.

Janis, I. L. (1972). *Victims of Groupthink: a Psychological study of foreign-policy decisions and fiascoes*. Boston: Houghton Mifflin.

Kahneman, D. (2011). *Thinking, Fast and Slow.* Farrar, Straus and Giroux.

Oakley J. E. and O'Hagan, A. (2016). SHELF: the Sheffield Elicitation Framework (version 3.0). School of Mathematics and Statistics, University of Sheffield, UK. (http://tonyohagan.co.uk/shelf)

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E. and Rakow, T. (2006). *Uncertain Judgements: Eliciting expert probabilities*. John Wiley and Sons, Chichester.

O'Hagan, A. (2012). Probabilistic uncertainty specification: overview, elaboration techniques and their application to a mechanistic model of carbon flux. *Environmental Modelling and Software* **36**, 35–48.

Parducci, A. (1963). The range-frequency compromise in judgment. *Psychological Monographs* **77** (2, Whole No. 565).

Ramsey, F. P. (1926). Truth and probability. Reprinted in *Studies in Subjective Probability*, H. E. Kyburg, Jr. and H. E. Smokler (eds.), 2nd edition, 23–52. New York: R. E. Krieger Publishing Company, 1980.

Rowe, G. and Wright, G. (1999). The Delphi technique as a forecasting tool; issues and analysis. *International Journal of Forecasting* **15**, 353–375.

Savage, L. J. (1954). *The Foundations of Statistics*. New York: John Wiley.

Tversky, A. and Kahneman, D. (1974). Judgments under uncertainty: heuristics and biases. *Science* **185** (4157), 1124–1131.