4

Quantifying scientific uncertainty from expert judgement elicitation

W. P. ASPINALL AND R. M. COOKE

> Now the geologists Thompson, Johnson, Jones and Ferguson state that our own layer has been ten thousand years forming. The geologists Herkimer, Hildebrand, Boggs and Walker all claim that our layer has been four hundred thousand years forming. Other geologists, just as reliable, maintain that our layer has been from one to two million years forming. Thus we have a concise and satisfactory idea of how long our layer has been growing and accumulating.
>
> *Mark Twain,* A Brace of Brief Lectures on Science*, 1871*

4.1 Introduction

Most scientists would like to see scientific advice used more in government decision-making and in all areas of public policy where science is salient, and many would welcome the opportunity to sit on expert review panels or scientific advisory committees. When it comes to taking such decisions in many areas of hazard and risk assessment, the traditional committee approach still holds sway. Often in a committee setting, however, the role of scientific uncertainty is not an item on the agenda, and seldom a prominent component of the discussion. But misunderstanding its importance or misstating its extent will contribute to poor decisions.

The slow, deliberative committee process, seeking a wide range of opinions with majority voting on outcomes, offers some parallels with the scientific process itself, but only in as much as a show of hands can equate to strength of argument. But as a means of gathering expert opinion it is inadequate under many conditions, such as an urgent civil emergency arising from an incipient natural disaster such as a hurricane or volcanic eruption – situations demanding prompt scientific advice. Usually, such advice would be the prerogative of a chief scientist, with all the contingent stresses and personalisation issues involved, including the pressure to be extremely cautious. When people's lives are at risk, the responsibilities are massive.

But, as with climate change modelling or forecasting severe weather events, in such cases there is always inherent randomness in the complex dynamic processes involved – usually

Risk and Uncertainty Assessment for Natural Hazards, ed. Jonathan Rougier, Steve Sparks and Lisa Hill. Published by Cambridge University Press. © Cambridge University Press 2013.

denoted as 'aleatory uncertainty' (see Chapter 2). Even the most sophisticated scientific models do not capture fully the range and extents of these stochastic variations – only Nature herself can do that! Understanding the potential scale and timing of natural hazards for the purposes of taking actions to protect the public inevitably requires simplifications and assumptions which, in turn, necessitate expert judgements, some of them intellectually brave.

How can this be done in a balanced, neutral but, in some sense, rationally optimised way? A formal procedure for eliciting and pooling judgements from a group of experts can help to quantify scientific uncertainty for high-stakes decision-making. There are several methods for doing this, each with their own flaws and biases (Kahneman *et al*., 1982; Tversky and Kahneman, 2005; Kynn, 2008; Kahneman, 2011). In the 1960s, for example, methods such as Delphi surveying aimed to extract a consensus from a group of experts for informing decisions.

Contributing to the difficulties in obtaining expert views in the presence of scientific uncertainty is the vexed issue of the distinction – if such exists – between 'objective' frequency- or statistics-based probability and 'subjective' degree-of-belief probability (see, e.g. Vick, 2002) – each having its own philosophical underpinnings, protocols and research literature. For most extreme, and difficult, hazard assessment problems, the former is rarely tenable – the data simply do not exist. This being so, recourse to belief-based expert judgements is the only viable option, and the means of eliciting such views becomes the real challenge.

In such circumstances, it is the decision-makers who are responsible – and answerable – for taking decisions under uncertainty but, not infrequently, they try to offload the onus onto scientists. As a failure of leadership, this sometimes is done covertly. Former EPA administrator Christine Todd Whitman didn't see it that way and roundly admitted:[1] 'A big part of my frustrations was that scientists would give me a range. And I would ask, "please just tell me at which point you are safe, and we can do that".'

However, it is neither rational nor appropriate to expect total consensus among experts when they are asked to make judgements on ill-constrained complex problems. In fact, the scientific method is the only legitimate way to seek to get scientists to agree. If the scientific method is unable to resolve an active issue, then the scientists *should* disagree, and any committee or other 'consensual' process that attempts to square this circle will promote confusion between consensus and certainty.

Thus, a valid goal of structured elicitation is to quantify uncertainty, not remove it from the decision process. For safety-critical decisions, individual experts are as likely to be overcautious as overconfident in their opinions. Ultimately, however, decision-makers want the best advice, not the most cautious and not the overconfident either, and this is where uncertainty muddies the water. A formal procedure for pooling the judgements of a small group of experts can give each the opportunity to express their 'true' opinion. In high-stakes

[1] Quote is from *Environmental Science and Technology Online*, 20 April 2005.

situations, the scientists involved can also defend their advice as coming from a collective decision process, rather than a single personal view.

A fundamentally distinct approach for doing this has been formulated by Cooke (1991), which differs from earlier elicitation procedures in that it does not try to impose total consensus or absolute agreement on a group. Instead, it is a structured procedure for obtaining uncertainty judgements from experts, measuring their individual judgement capabilities with a performance-based metric, and then applying mathematical scoring rules to combine these individual judgements into a 'rational consensus' that can inform the deliberations of policy-makers.

This particular approach features prominently in the current chapter, and crops up most frequently in Sections 4.2 and 4.3, on expert judgement and expert elicitations in practice, respectively. The discussion then moves on to consider issues in communicating expert uncertainty, with a focus on policy setting (Section 4.4), followed by a brief examination of possible, or desirable, future directions in research on expert elicitation, theoretical and methodological (Section 4.5).

4.2 Expert judgement elicitation

While expert opinion is almost invariably sought when scientific or technical uncertainty impacts on an important decision process, such uncertainty is ubiquitous with scientific knowledge – if it were not, any decision related to the scientific issue at hand would be obvious. Thus, there is the inescapable corollary that the experts themselves cannot be absolutely certain, and thus it is almost impossible that they will ever be in total agreement with one another. This is especially true where such uncertainty is substantial, or where the consequences of the decision are particularly serious or onerous.

In circumstances where scientific uncertainty impinges on the resolution of an issue, soliciting expert advice is not new. However, the forms which this can take continue to evolve. The main landmarks in this progression are grouped below according to whether the operable methodology is structured or unstructured. Very roughly, 'structured' means that the questions which experts answer have clear operational meaning, and the process of arriving at a defined position is traceable and open to review.

Operational meaning is of paramount importance. Questions of substantial cognitive depth often involve ambiguities ('what is meant by 'severe weather'?). In order to focus on the underlying uncertainties about nature, opacity or ambiguity in the meaning of words must be resolved. Left to their own devices, different experts will do this in different ways, resulting in unnecessarily noisy answers. Ambiguity can be reduced to an acceptable level by casting the questions in terms of thought experiments, or 'thought measurements' that could in principle, albeit not in practice, be performed. Translating the original questions into questions with clear operational meaning is often a difficult task. On the other hand, some questions have little or no cognitive depth ('In what year will man land on Mars'?).

Some examples of unstructured and structured methods are:

Unstructured expert judgement:
- expert oracles: courtroom experts
- expert pundits: talking-head experts
- expert surveys: the wisdom of crowds
- 'Blue ribbon'[2] panels: greybeard experts

Structured expert judgement
- Delphi
- nominal group techniques
- equal weighting
- self-weighting
- peer weighting
- performance weighting
- technical integrator based
- citation based
- likelihood based
- Classical Model.

A number of good books approach expert judgement with a wide compass and discuss various of these approaches: Kahneman *et al*. (1982); Morgan and Henrion (1990); Cooke (1991); Woo (1999, 2011); O'Hagan *et al*. (2006). The influential thinking of Tversky and Kahneman (2005) about judgements under uncertainty is reprised and extended by Kahneman (2011), with more contemporary psychological and neuroscience insights.

Cooke and Goossens (2008) compiled an archive of case histories, termed the TU Delft database, comprising expert judgement studies in which experts were asked questions from their field for which true values were known post hoc. This has provided a resource for testing various expert combination schemes. We offer a brief sketch of the expert judgement approaches, and illustrate the strengths and weakness of the most relevant of these with case studies and anecdotal experience. The unstructured approaches are treated summarily.

The above synoptic summary may surprise those readers who expected to see listed something denoting a Bayesian approach. Bayesian approaches fall under likelihood-based methods and are briefly referenced in that context.

4.2.1 Unstructured elicitations

Expert oracles are experts who have a hotline to the truth, denied to the rest of us. One such expert is as good as any other, provided (s)he is really an expert. The archetype is the ballistics expert giving courtroom testimony on which bullet came from which gun. There is no second opinion, no dissenting view and no uncertainty with oracular experts.

[2] 'Blue Ribbon' is a term invented in the United States; if it had been coined in the UK one presumes it would have been Blue Riband. Reluctantly, we feel obliged to stick with contemporary American usage.

Expert pundits are found on TV and radio news shows and in newspapers where experts are conjured up (or volunteer themselves) to shed light on whatever is newsworthy. The word derives from the ancient Sanskrit word *pandita*, meaning learned or wise. Without structured constraint, oracular status gets lost in the cacophony of acrimonious interruption, and the term pundit has acquired an unfavourable aroma. Nonetheless, these 'experts' are widely perceived as knowing more than the rest of us.

Expert surveys may appear, superficially, to be structured, but usually have some promotional aspect, as in '99% of all doctors prescribe Damnital for pain relief'. Unlike other opinion surveys, however, credibility is not always enhanced by including ever more experts. What would we say to '99% of financial experts say a financial crisis is not imminent'? The crowd may often be wise, but the herd sometimes stampedes over a cliff.

Blue ribbon panels of greybeards draw credibility from their scientific status and, purportedly, from the disinterest associated with their age and station – a position where knowledge marries honesty. On questions impacting the stature and well-being of a field to which such a person has devoted his/her career, is it naive to think that the marriage of knowledge and honesty is really made in heaven? Convening committees of the 'great and good' are commonplace in government, whereas blue ribbon panels in science are infrequent and, by science's usual pecuniary standards, expensive undertakings. In early 2010, President Obama convened the Blue Ribbon Commission on America's Nuclear Future,[3] and Chief White House Energy Advisor Carol Browner said 'It is time to move forward with a new strategy based on the best science and the advice of a broad range of experts.' US Energy Secretary Steven Chu said the commission will have a free hand to examine a 'full range of scientific and technical options' on waste storage, reprocessing and disposal – with one exception: the once scientifically favoured Yucca Mountain underground repository. Thus, all too often, independent scientific thought becomes totally marginalised in highly politicised matters.

4.2.2 Early structured elicitations and advances

To better understand where things stand now, a brief reiteration of the evolution of structured expert judgement methodologies from primeval forms to present praxis may be helpful; some thematic variations are mentioned along the way, for completeness.

The *Delphi method* was pioneered in the 1950s at the RAND Corporation (Helmer, 1966, 1968) and is the method that most laypeople have heard of. It was originally crafted for technology assessment ('In what year will commercial rocket transport become available?'), but quickly branched into many fields. Consensus is achieved by repeatedly feeding assessments back to a set of expert respondents and asking outliers to bolster their eccentric views with arguments. Delphi studies are frequently conducted by mail, and questions typically do not involve a great deal of cognitive depth. Sackman's withering *Delphi Critique* (1975)

[3] The BRC Final Report was released on 26 January 2012 (available at http://www.brc.gov on 27 July 2012); a draft with tentative conclusions was discussed in a Congressional joint hearing in October 2011.

noted that, without reporting the drop-out rate, the suspicion cannot be allayed that 'consensus' results from simply pestering the outliers away. Delbecq *et al*. (1975), and Gough (1975) in separate studies found that the Delphi method performed worst of all techniques analysed. Fischer (1975) and Seaver (1977) found no significant differences among any methods. The latter study also found that group consensus tended to produce more extreme probability estimates. The scoring variables used in these studies are not always directly comparable, and there is no direct quantification of uncertainty. Woudenberg (1991) reviewed the literature concerning quantitative applications of the Delphi method and found no evidence to support the view that Delphi is more accurate than other judgement methods or that consensus in a Delphi study is achieved by dissemination of information to all participants. Data then available suggested that any consensus achieved is mainly due to group conformity pressure, mediated by feedback to participants of statistics characterising group response rather than information *per se*, corroborating Sackman (1975). For more background on the Delphi method, see Cooke (1991).

Nominal group techniques (Delbecq *et al*., 1975) involve bringing the experts together to debate the issues under the guidance of a skilled facilitator. Ideally, differences are fully aired during the first day, some bonding occurs after dinner and on the morrow a group consensus emerges. These were among the techniques tested by Fischer (1975) and Seaver (1977) with inconclusive results. The fact that experts must come together for a day or more constitutes a serious limitation of a practical nature. High-value experts have high-value agendas, and scheduling such sessions can easily cost six months of calendar time, and some will inevitably cancel at the last minute. While nominal group techniques were developed to obtain consensus summaries in specific circumstances, such methods were found to have significant limitations (Brockhoff, 1975; Seaver, 1978).

Equal weighting is the simplest of the mathematical aggregation methods. In a panel of N experts, each expert's probability, or probability distribution, is assigned a weight $1/N$, and the weighted probabilities or distributions are added to produce a 'decision-maker' (DM). Equal weighting of forecasts (as opposed to probabilities) emerges in Bayesian updating models under certain conditions (Cooke, 1991), with Clemen and Winkler (1987) putting forward arguments in favour of the approach. Its primary appeal is its simplicity. While other weighting schemes require justification, rightly or wrongly, equal weighting is usually accepted uncritically and without demur. Evidence (Cooke and Goossens, 2008) shows that the equal weight combination is usually statistically acceptable – usually, but not always. In seven out of the 45 studies examined by them, the hypothesis that the equal weight combination was statistically accurate would be rejected at the 5% level. Another feature of equal weighting is that the combined distributions can become quite diffuse as the number of experts increases. Figure 4.1 shows the 90% central confidence bands and median values assigned by 30 BA pilots for one particular target item: the wide spread on the credible interval for the equal weight (EQUAL) DM is by no means atypical (also shown is the corresponding Classical model performance weight (PERF) DM – see below). For instance, we see that the equal weight DM 90% confidence spread is much broader than that of any of the individual experts. The performance-based DM, which in this analysis distributes most

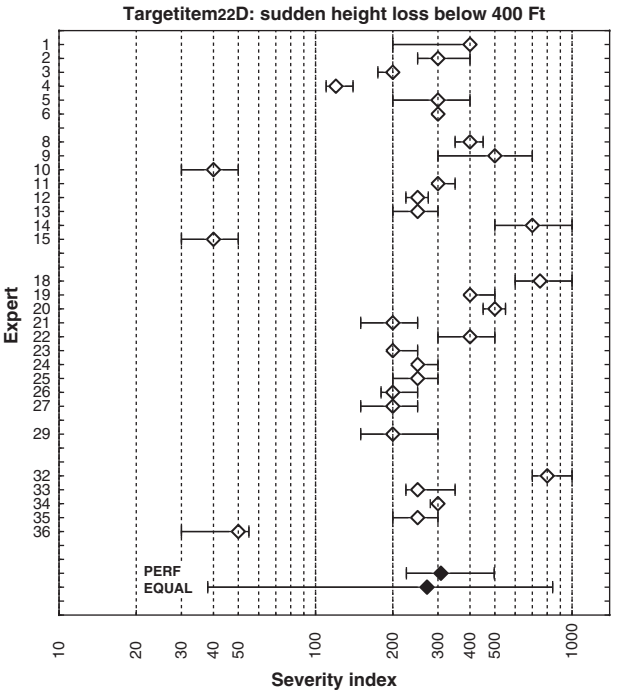

Figure 4.1 Range graph plot showing the median values and 90% confidence bands provided by 36 airline pilots asked to assign a relative severity index to sudden height loss below 400 ft altitude, a potential flight operations safety factor. At the bottom is the combined equal weights solution (EQUAL), with the corresponding classical model performance-based weights solution above (PERF). The equal weights credible interval is much wider than that of any individual expert. The performance-based DM distribution takes most support from the three top-ranked pilots, producing a much narrower uncertainty spread than equal weights. (Three other pilots, 10, 15 and 36, provided responses that deviated significantly from the majority, which would prompt a facilitator to explore their reasoning; six pilots provided null responses for this item.)

weight over the three top-ranked pilots, shows a distribution which is more 'humanoid'. (We note that three other pilots in this case provide responses that depart significantly from the majority, a diagnostic trigger for the facilitator to review, with them, their reasoning.)

Hora (2004) showed that equal weighting of statistically accurate experts always results in a statistically *inaccurate* combination. A procedure for uncertainty quantification based on equal weighting was developed for US nuclear risk applications (Hora and Iman, 1989). In the field of seismic risk, the SSHAC Guidelines (SSHAC, 1997) expound an elicitation approach that has the goal of assigning equal weights to experts (Hanks *et al*., 2009). The notion is that, as long as the evaluators have equal access to information and have fully participated in the workshops, 'equal weights are expected' (Hanks *et al*., 2009: 14). This said, within the higher-level SSHAC procedures the technical facilitator integrator (TFI) has the option to use different weights for different evaluator teams.

Self-weighting was originally used in some Delphi studies, in which experts were asked to rank their own expertise on a scale of 1 to 7 – the units were not given. Brockhoff (1975) found that women consistently ranked themselves lower than men, and that these rankings failed to predict performance. After Kahneman *et al*. (1982) highlighted the phenomenon of overconfidence, the inclination to use expert self-weights lost force. An early study of Cooke *et al*. (1988) found a strong negative correlation between informativeness and statistical accuracy. Roughly, that translates to: the tighter the 90% confidence bands, the more likely that the realisation falls outside these bands. The correlation is not strict, however, and there are often experts who are both statistically accurate and informative.

Peer weighting is an option that can be tried if the experts in a panel all know each other or each other's work. It was proposed independently by De Groot (1974) and by Lehrer and Wagner (1981). Real applications are sparse, but, in a study for dam erosion risk modelling (Brown and Aspinall, 2004), a comparison was made between performance-based expert weightings for an expert panel with mutual peer weighting by colleagues in the group. There were some major differences in ranking between the two: for example, some experts scored significantly less well on the performance-based measure than their colleagues might have anticipated, while others did much better (see Figure 4.2). If

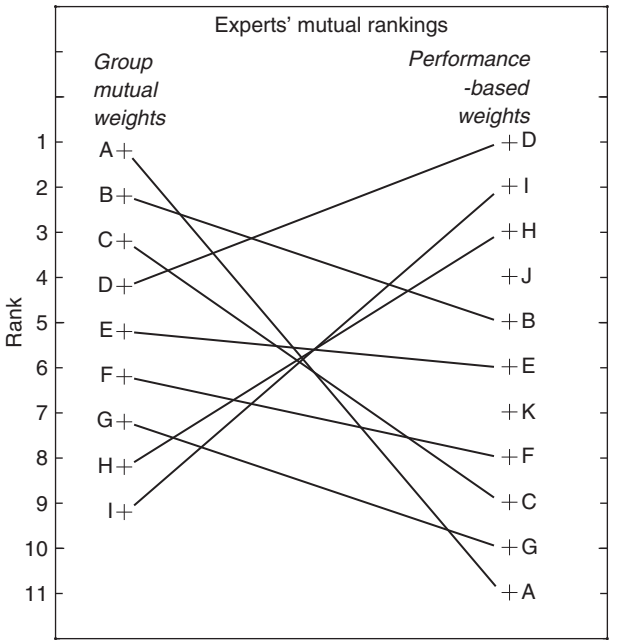

Figure 4.2 Dam erosion risk panel elicitation: comparison between experts' ordering by group-assigned mutual weights (nine initial experts: A–I) and the ranking of all 11 experts in the eventual elicited group (A–K) from performance-based scoring using Cooke's (1981) Classical Model (right-hand column). Note several apparent reversals of reputational authority are evident for the nine (see text).

anecdotal evidence can be relied upon, this sort of bias can be expected in peer weighting with almost any group of specialists, whatever the discipline: typically, some experts are well-regarded but tend to be strongly opinionated, while others – perhaps more reflective, and considered indecisive or diffident by their peers – are, in fact, better estimators of uncertainty. In the dam erosion study the quantification of model parameter uncertainties was the main objective, so it was decided appropriate by the problem owner that the better-calibrated experts should be rewarded with greater weight, even though this appeared to conflict with the peer appraisals.

Performance weighting denotes any system in which weights are assigned to experts' views according to their performance. These come in a number of variants, with two principal flavours dominating current praxis (see Section 4.2.3); other, less often used, variations include citation-based and likelihood-based weights.

Citation-based weights, which are based on the relative number of times that experts cite each other's work. Many complications arise in implementing this deceptively simple idea. Should self-citations be counted? How do we deal with joint authorship? Do we discount the citations of older experts, as these tend naturally to grow with age? The uncertainty analyses conducted in the Joint US Nuclear Regulatory Commission–European Union uncertainty analysis of accident consequence codes afforded an opportunity to test citation-based weighting, since all experts were well known and well published. The results analysed by Cooke *et al*. (2008) show that citation weighting performed about as well as equal weighting in terms of statistical accuracy and informativeness.

Likelihood-based weights can denote any of a variety of schemes. In any case, such methods require assessments of calibration variables whose true values are known post hoc. In the simplest case, an expert's weight is proportional to the probability which he assigned to the observed values of the calibration variables. Considered as a scoring variable, this is sometimes called the 'direct rule' in which an expert's reward is proportional to his/her probability of the event which occurred. The direct rule is notoriously improper: it encourages treating the most likely event as if it were certain (Cooke, 1991). Bayesian approaches can be seen as likelihood-based weighting methods, with admixture of the decision-maker's prior distribution. In the simplest models, a decision-maker's prior distribution is updated with predictions from experts, just as if these were physical measurements. The decision-maker must account for dispersion, correlation and bias in his 'expert-instruments' (Lindley, 1971; Winkler, 1981; Mosleh and Apostolakis, 1982). Bayesian methods based on expert probabilistic assessments were elaborated by Morris (1974, 1977) and Mendel and Sheridan (1989). A system proposed by Clemen and Reilly (1999) is similar to likelihood weighting, with a twist that accounts for expert dependence. A check, using studies from the TU Delft database, showed its performance was rather indifferent (Kallen and Cooke, 2002). In more recent work, Cooke *et al*. (2008) were able to test a version of likelihood weights using cases from the TU Delft database, and also found very uneven performance.

4.2.3 Current mainstream approaches to structured elicitation

In current practice, there are two major but different *performance weighting* schemes that are often utilised for structured elicitation, in which weights are assigned to experts' views. In one, weights are ascribed according to the appraisal of performance by a technical facilitator or facilitators, and, in the other, performance weights are determined by applying a scoring rule measure to an empirical test of judgement skill.

4.2.3.1 The SSHAC elicitation procedure

The pre-eminent approach to expert elicitation using a technical facilitator approach is found in the methodology espoused by the US Department of Energy (DoE) for estimating site-specific probabilistic seismic hazard curves (see also Chapter 8). In order to accommodate uncertainties in earthquake occurrence rates and in ground motion attenuation processes that influence effects at a site, the DoE favours a method[4] for expert elicitation that is based mainly on work for the US NRC, articulated by the Senior Seismic Hazard Analysis Committee (SSHAC, 1997; Budnitz *et al*., 1998).

The SSHAC committee report outlines a formalised expert elicitation procedure with four levels of intensity of application. The most extreme, SSHAC Level 4, utilises the concept of the 'Technical Facilitator Integrator' to integrate the information from groups of subject-matter specialists. The procedure also involves the use of participatory reviews, requiring additional specialist experts to follow progress in the different topics that contribute to the hazard assessment from the beginning of the study, rather than waiting until the end of the whole project to provide comments. It is held that this approach creates the proper conditions to provide for a defensible decision to be made, based on a commonality of information and data among all the experts involved.

The SSHAC thinking is intended to promote a methodological culture that aspires to direct incorporation of epistemic uncertainty into a 'total' risk assessment process, by 'finding the center, body, and range of the informed technical community' (Hanks *et al*., 2009). In pursuit of this, a TFI seeks to represent the whole community of experts by elicitation of the views of a sample group. The TFI may seek to correct the capriciousness of chance in the expert group selection: 'to represent the overall community, if we wish to treat the outlier's opinion as equally credible to the other panelists, we might properly assign a weight (in a panel of 5 experts) of 1/100 to his or her position, not 1/5'.[5] Thus differential weighting is based on the integrator's assessment of the 'representativeness' of an expert's views, and in that sense relates to the expert's 'performance' or, more accurately, the TFI's perception of it.

The SSHAC approach positively embraces the concept of equal weights, endeavouring to make a virtue of democratic principles in this way of thinking. However, this takes no account of the precept that not all experts are equally adept at making judgements, especially where this concerns uncertainty. It might be argued that a method that assesses the relative proficiency of experts in this regard will furnish a more robust approach for decision support.

[4] US DoE Standard: Natural phenomena hazards assessment criteria; DOE-STD-1023–95, April 2002.
[5] NUREG/CR-6372, at p. 36.

4.2.3.2 The Classical Model

One of the main goals of a formal elicitation of expert judgement is to remove as much subjectivity from decision-making as possible. In the context of scientific advice, misunderstanding can arise about objectivity and subjectivity in relation to expert judgement for two reasons. One is judgements are based on an individual expert's personal degree of scientific belief on an issue and, because this is 'personal', it is sometimes construed as being 'subjective', in some ill-defined sense. The second is that such judgements are often referred to as 'expert opinion' – while *anyone* can have an opinion, meaningful scientific judgement requires specialist knowledge, practised reasoning skills and real experience. An expert is someone who not only knows a great deal about a discipline, but also understands how that discipline is organised, its rules, semantics, culture and methodologies.

Thus, genuine expertise entails the key attribute of objectivity, and eliciting expert judgement calls for a structured, rational process that captures this objectivity in a neutral and impartial manner. In many elicitations, such as those undertaken in connection with the Montserrat volcanic activity (Section 4.3.5), emphasis is placed on ascertaining scientists' judgements about the extent of uncertainties on parameters, variables or probabilities, for the purpose of including these in probabilistic hazard and risk assessments.[6] For scientists assessing natural hazards, emergent diseases, climate change or any number of other challenges that may turn out to be major threats to public safety, concern ought to be focused on the systematic rational interpretation of all available evidence and associated uncertainties. Some of this evidence and many of the relevant uncertainties can only be characterised on the basis of expert judgement.

However, it is important to recognise that experts are not necessarily equally adept and proficient at judging scientific uncertainty so, for decision support, a constructive method is needed which can differentiate individual judgements. One way in which this can be done is to use a procedure which theoretically maximises objectivity in the face of uncertainty, utilising a method that can weigh experts' judgements on the basis of empirically validated performance metrics, measured within the specialism domain.

In what follows, a detailed description is given of the Cooke Classical Model (Cooke, 1991), the only structured group elicitation approach that has this empirical control, an important credential of the scientific method. In addition, the scheme offers reproducible and auditable tracking of how a particular scientific issue has been characterised quantitatively by a group of experts for input to a decision. In increasingly litigious times, this ought to be seen as beneficial by participating experts when they are called upon to give science-based advice in support of crucial, possibly life-or-death, decisions. It is in such circumstances that the concept and principles of the Classical Model come to the fore.

[6] Elsewhere, the apparently simpler, and hence more appealing, deterministic approach is sometimes followed. However, the extents of these multiple, and multiplied, uncertainties are almost impossible to accommodate rationally into decision-making under the deterministic prescription – sometimes the consequences of this shortcoming can be disastrous: for instance, deterministic underestimation of the upper bound magnitude for great earthquakes off Tohoku, northeast Japan, was a key factor in the tragic surprise caused by the earthquake and tsunami of 11 March 2011.

The Classical Model method relies on the use of proper scoring rules for weighting and combining expert judgements through statistical accuracy and information scores, measured on calibration variables (see Cooke, 1991), and operationalises principles for rational consensus via a performance-based linear pooling or weighted averaging model. The weights are derived from experts' calibration and information scores, as measured on seed item calibration variables. Calibration variables serve a threefold purpose:

(1) to quantify experts' performance as subjective probability assessors;
(2) to enable performance-optimised combinations of expert distributions; and
(3) to evaluate and hopefully validate the combination of expert judgements.

The name 'Classical Model' derives from an analogy between calibration measurement and classical statistical hypothesis testing. An expert's *calibration* score is the *p*-value (i.e. significance level) at which one would falsely reject the hypothesis that the expert's probability statements were correct. If an expert gives 90% confidence bands for several items which are later observed (or known otherwise), and if a substantial proportion of the observed values fall outside those confidence bands, then the calibration score is derived from the probability that such an outcome could have arisen purely by chance. The designation 'classical' denotes a contrast with various Bayesian models, which require prior assessments of expert and variables by the decision-maker.

The Classical Model performance-based weights also rely on a second quantitative measure of performance, *information*, which measures the degree to which a distribution is concentrated.

In practice, both of these measures can be implemented for either discrete or quantile elicitation formats. In the discrete format, experts are presented with uncertain events and perform their elicitation by assigning each event to one of several pre-defined probability bins, typically 10%, 20%, … 90%. In the quantile format, experts are challenged with an uncertain quantity taking a value in a continuous range, and they give pre-defined quantiles, or percentiles, for their corresponding subjective uncertainty distribution, typically 5%, 50% and 95%. In application, this format is found to have distinct advantages over the discrete format, not least in the effort demanded of experts to express their uncertainty judgements.

For *calibration* in the case of three quantiles, each expert divides the quantity range into four inter-quantile intervals within which they can judge the relative probabilities of success, namely: $p_1 = 0.05$: the quantity is less than or equal to the 5% value; $p_2 = 0.45$: greater than the 5% value and less than or equal to the 50% value, and so on.

If N such quantities are assessed, each expert may be regarded as a statistical hypothesis, namely that each realisation of the N falls in one of the four inter-quantile intervals with probability vector corresponding to those defined intervals, from which can be formed a sample distribution of the expert's inter-quantile intervals. If the realisations are indeed drawn independently from a distribution with quantiles as stated by the expert, then the likelihood ratio statistic is asymptotically distributed as a chi-square variable with 3° of freedom (see Cooke, 1991; Cooke and Goossens, 2008).

For such a distribution, it is straightforward to derive the familiar chi-square test statistic for goodness of fit, and this test scores the particular expert as the statistical likelihood of the hypothesis:

H_e: the inter-quantile interval containing the true value for each variable is drawn independently from the expert's defined probability vector.

Using the likelihood ratio statistic, the analyst computes a test p-value, the latter being the probability that a deviation at least as great as that observed could occur on N realisations if H_e were true; this p-value forms the basis for determining the expert's statistical accuracy or calibration score. Calibration scores are absolute and can be compared across studies. However, before doing so, it is appropriate to equalise the power of the different hypothesis tests by adjusting the effective number of realisations in each when these differ from one study to another.

Although the calibration score uses the language of simple hypothesis testing, it must be emphasised that expert-hypotheses are not being rejected; rather, this language is used to measure the degree to which the data support the hypothesis that the expert's probabilities are accurate. Low calibration scores, near zero, mean that it is unlikely that the expert's probabilities are correct.

Under the Classical Model, the second scoring variable is *information* or *informativeness*. Loosely, the information in a distribution is the degree to which the distribution is concentrated. Information cannot be measured absolutely, but only with respect to a background measure. Being concentrated or 'spread out' is measured relative to some other distribution – commonly, the uniform and log-uniform background measures are used (other possible background measures are discussed by Yunusov *et al*., 1999).

Measuring information requires associating a density with each quantile assessment of each expert. To do this, the unique density is found that complies with the experts' quantiles and is minimally informative with respect to the background measure. For a uniform background measure, the density is constant between the assessed quantiles, and is such that the total mass between quantiles i and $i+1$ agrees with the corresponding relative probability p_i. The background measure is not elicited from experts but must be the same for all experts; it is chosen by the analyst.

The *information score* for an expert is calculated from the average relative information of the expert's joint distribution given the selected background, under the assumption that the variables involved are independent. As with calibration, the assumption of independence reflects a desirable property for the decision-maker process, and is not an elicited feature of the expert's joint distribution. Furthermore, the information score does not depend on the realisations: an expert can give himself a high information score by choosing his quantiles very close together but, evidently, his or her information score also depends on the ranges and assessments of the other experts in the particular group. Hence, information scores cannot be compared across studies.

Of course, other measures of concentratedness could be contemplated. The above information score is chosen because it is tail-insensitive, scale invariant and 'slow', in the sense that

large changes in the expert assessments produce only modest changes in the information score. This contrasts with the likelihood function in the calibration score, which is a very fast function. This causes the product of calibration and information to be dominated by the calibration score.

To illustrate the ways in which the calibration and information traits of individual experts can vary in practice, a hypothetical quartet is presented schematically in Figure 4.3. In that example, four mythical experts respond to the same ten seed item calibration questions. While seed items are questions for which the true values are known to the facilitator, or will become known, experts are not expected to know these values precisely but, on the basis of their expertise, they should be able to define quantised credible ranges for each, within which the true value falls with some assignable probability.

Figure 4.3 illustrates some typical expert judgement traits which, in practice, can affect individual expert's uncertainty assessments to one degree or another; ideally, such propensities should be controlled for when eliciting opinions for decision support. However, it is very difficult to determine a priori from other criteria, such as publication record or reputation, which expert has what tendency in this sense – such insight only emerges from an empirical judgement calibration analysis.

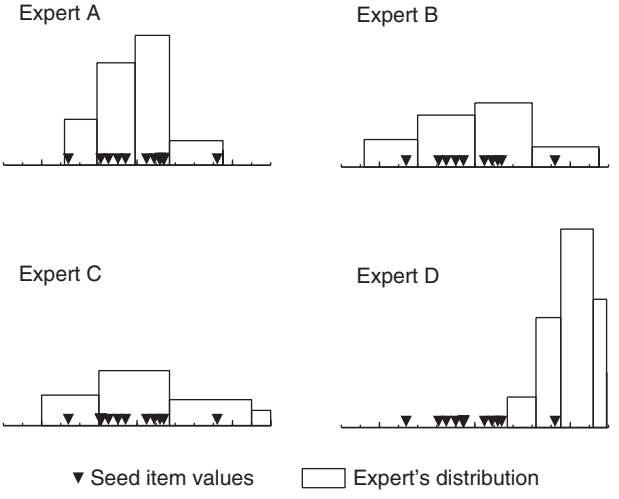

Figure 4.3 Schematic representation of four experts' uncertainty distributions in relation to a set of ten seed items used for calibration scoring with the Classical Model. Each expert defines an uncertainty range for each seed item by nominating three quantiles which should enclose the realisation value (small extensions are added in the tails to create four location probability bins). Over the ten items: Expert A is statistically coherent – his distributional spreads capture and take support from the realisation values – and he is also informative and hence scores well; Expert B is equally accurate, statistically, but is less informative than A, and would be ranked with a lower weight; Expert C is superficially similar to B, but his judgements are systematically shifted relative to realisation support and would be penalised for this bias; Expert D is very confident in his judgements yet extremely poorly calibrated, statistically – he should be surprised by the realisation values – and his weighting would be very low, even zero, as a consequence (see text for a more detailed description of scoring concepts).

Under the Classical Model concept, the combination of assessments by several experts is called a *decision-maker* (DM), and is an example of linear pooling.[7] The Classical Model is essentially a method for deriving weights in a linear pool. 'Good expertise' corresponds to good calibration (high statistical likelihood, high p-value) and high information. Weights are wanted which reward good expertise and which pass these virtues on, through the pooling DM.

The reward aspect of weights is very important. An expert's influence on the DM should not appear haphazard, and experts should be discouraged from gaming the system by tilting their assessments in an attempt to achieve a desired outcome. These goals are achieved by imposing a strictly proper scoring rule constraint on the weighting scheme. Broadly speaking, this means that an expert gains his or her maximal expected weight by, and only by, stating assessments in conformity with his or her true beliefs. Some might be tempted to attempt to game the process by deliberately expanding their credible ranges beyond genuine personal assessments, as underestimating uncertainty is a recognised trait. But being lured into this – even marginally – engenders the risk of being penalised via the information scoring: an expert cannot be sure a priori that he or she has a confirmed tendency to systematically understate uncertainties.

Within this model, two important variants of linear pooled DM are available: the *global weight DM* is termed global because the information score used for weighting is based on all assessed items. A variation on this scheme allows a different set of weights to be used for each item. This is accomplished by using information scores for each item rather than the overall average information score. Item weights are potentially more attractive as they allow an expert to up- or down-weight him or herself for individual items according to how much he or she feels they know about that item. 'Knowing less' means choosing quantiles further apart and lowering the information score for that item.

Of course, good performance of item weights requires that experts can perform this up/down weighting rationally and successfully. Anecdotal evidence suggests that, as experts gain more experience in probabilistic assessment, item weights pooling improves over the global weights option. Both item and global weights can be described as optimal weights under a strictly proper scoring rule constraint: in each, calibration dominates over information, while information serves to modulate between more or less equally well calibrated experts.

Since any combination of expert distributions yields assessments for the calibration variables themselves, any DM combination can be evaluated on the calibration variables. Specifically, the calibration and the information of any proposed DM can be computed with justifiable expectation that the best pooled DM would perform better than the result of

[7] Cooke (1991: ch. 8, 9) discusses asymptotically strictly proper scoring rules in formal mathematical terms, and also expounds on the properties of such rules for combining probabilities by linear pooling (ch. 11). Suffice it here to say that much work has been done on different ways of merging quantitative judgments and that there are strong reasons for forming (weighted) combinations of experts' density functions (or cumulative density functions) by linear pooling – rather than averaging variable values with the same probabilities (i.e. quantile functions). Counterfactual examples are easily constructed which demonstrate that alternatives, other than linear pooling, often result in combination averages that assign mass to forbidden probability intervals or otherwise produce irrational or impossible marginalisation or zero preservation exceptions (see Cooke, 1991, ch. 11).

simple averaging (i.e. an *equal weight DM*), and that the proposed DM is not significantly worse than the best expert in the panel – although this sometimes happens. Figure 4.1 shows a typical example of a performance-based weighted target item solution from an elicitation of 30 senior civil airline pilots concerning flight operation safety factors compared with the corresponding *equal weights* solution (see Section 4.3.4).

Thus, the Classical Model process is designed to enable rational consensus: participants pre-commit to a process bearing the essential hallmarks of the scientific method, before knowing the outcome. These hallmarks are:

Scrutability/accountability: all data, including experts' names and assessments, and all processing tools are available for peer-review and results must be open and reproducible by competent reviewers.

Empirical control: quantitative expert assessments are subjected to empirical quality controls.

Neutrality: the method for combining/evaluating expert opinion should encourage experts to state their true opinions, and must not bias results.

Fairness: experts' competencies are not pre-judged, prior to processing the results of their assessments.

The Classical Model approach is a bit more complicated than others, with the main precept that the weights reward good performance in probabilistic assessment. This is *not* the same as commanding a wealth of knowledge, or giving accurate predictions. Rather, a good probabilistic assessor is one who is able to quantify his or her uncertainty in a way which is informative and statistically accurate – considered as a statistical hypothesis, an expert would be credible.

The Classical Model system has been widely applied and documented (see Cooke and Goossens, 2008, and references therein). Independent espousals have come from practitioners across various fields (Woo, 1999, 2011; Aspinall, 2006, 2010; Klügel, 2007; French, 2008, 2011; Lin and Bier, 2008).

4.3 Expert elicitations in practice

There is a steadily burgeoning catalogue of case histories of expert-informed decisions in practice: in seismic hazard assessment, volcanic hazards assessment, emerging infectious disease risk modelling, flight safety and dam safety, finance, biosecurity and nuclear safety, all of which have utilised either structured elicitations (to different degrees of formalisation) or unstructured elicitations. A very informative generic discussion on how to conduct an expert elicitation workshop, targeted at facilitators mainly, is provided by the Australian Centre of Excellence for Risk Analysis (ACERA, 2010), a report which has a valuable literature review companion (ACERA, 2007). These two reports represent an excellent start point for novice facilitators and domain experts, alike.

Here we recount a few insights from examples of the formalised structured type, starting with two prominent SSHAC case histories, before adding to the narrative experiences with the Classical Model.

4.3.1 SSHAC and the PEGASOS Project

The SSHAC approach (see Section 4.2.3) was applied in the PEGASOS Project, the first European trial application of the procedure at its most complex level (level 4), with the goal of developing site-specific seismic hazard estimates for the sites of Swiss nuclear power plants (Abrahamson *et al*., 2004; Zuidema, 2006; Coppersmith *et al*., 2009). The power plant owner sponsored the study in response to a request from the Swiss regulator, following the latter holding discussions with US consultants and with US NRC officers. The PEGASOS Project was subdivided into four subprojects:

(1) SP1 – Seismic source characterisation (four groups of experts, each group consisting of three experts).
(2) SP2 – Ground motion characteristics (five experts).
(3) SP3 – Site-response characteristics (four experts).
(4) SP4 – Hazard quantification.

Therefore, the study followed the convolution approach, separating source, ground motion and site-response characteristics, and commissioned two TFIs to manage the elicitations, one for SP1 and another for SP2 and SP3. The study was based on the use of comprehensive logic trees to reflect the epistemic uncertainties related to source, ground motion attenuation and site characteristics relevant for the evaluation of seismic hazard models for the Swiss sites. Equal weights were used to combine the different expert opinions.

Seismologically, an interesting feature of the project was the attempt to constrain maximum ground motion levels by estimating an upper limit for this variable (see also Chapter 8). The final review of the project results by the sponsor (see Klügel, 2005a, 2007), as well as by the Swiss nuclear safety authority, identified the need for a further development of the probabilistic seismic hazard analysis, and a 'refinement' project is underway.

One of Klügel's (2005a, 2009) main criticisms (of many) of the original project was that the SSHAC expert opinion elicitation concept is philosophically suspect and logically flawed, and relies on inadequate elicitation and aggregation methods that are based on political consensus principles or on census principles, rather than on principles of rational consensus (Klügel, 2005d). His initial intervention (Klügel, 2005a) engendered a major debate in the literature (Budnitz *et al*., 2005; Klügel, 2005b–f, 2006, 2009; Krinitzsky, 2005; Lomnitz, 2005; Musson *et al*., 2005; Wang, 2005). Judging by the ensuing furore, not to mention the huge related costs and the sheer difficulty of coping with a logic tree that had at least 10^{15} retained branches to represent all the models and alternative parameterisations that the evaluator panels could envisage, it seems the SSHAC approach has not been an unqualified success in application in the PEGASOS Project. Nor, as discussed next, has the approach proved convincing in volcanological hazard assessments undertaken for the Yucca Mountain Project (YMP).

4.3.2 SSHAC and the Yucca Mountain Project

An important case history in terms of the use of expert judgement is provided by the proposed high-level radioactive waste repository at Yucca Mountain, Nevada, where studies of the geologic basis for volcanic hazard and risk assessment had been underway for nearly three decades. Probabilistic hazard estimates for the proposed repository depend on the recurrence rate and spatial distribution of past episodes of volcanism in the region, and are of particular concern because three episodes of small-volume, alkalic basaltic volcanism have occurred within 50 km of Yucca Mountain during the Quaternary.

The use of expert elicitations within the YMP was extended to cover more specific issues, such as the Near-Field/Altered Zone Coupled Effects Expert Elicitation Project (NFEE), and similar issues requiring expert judgement inputs. In addressing such technically complex problems, expert judgement had to be used because the data alone did not provide a sufficient basis for interpreting the processes or for the outputs needed for subsequent analyses. For example, experiments such as the 'single heater test' did not provide a direct estimate of the spatial distribution of fracture permeability changes: site-specific experimental data had to be interpreted and combined with other data and analyses before they could be used in performance assessments.

In such cases, experts have to integrate and evaluate data in order to arrive at conclusions that are meaningful for technical assessments of particular geophysical and geochemical effects, including quantitative and qualitative expressions of the uncertainties. In principle, the process which is followed should be the same, however abundant or scarce the data, the only difference being the level of uncertainty involved. In this sense, expert judgement is not a substitute for data; rather, it is the process by which data are evaluated and interpreted. Where data are scarce and uncertainties high, the uncertainties expressed by each expert and the range of judgements across multiple experts should reflect that high degree of uncertainty.

The YMP studies closely followed the procedural guidance first set out in SSHAC (1995), both in spirit (e.g. a belief in the importance of facilitated expert interactions) and, in many cases, in details of implementation (e.g. suggestions for conducting elicitation interviews). For example, the NFEE process was designed – in accordance with SSHAC guidance – to result in assessments that represent the 'range of technical interpretations that the larger informed technical community would have if they were to conduct the study'. However, in as much as SSHAC professes to be 'non-prescriptive' in specifying a single way of implementing the process, it would be wrong to suggest that all the YMP expert elicitations conformed precisely with the SSHAC procedures. In some cases, the processes were specifically tailored to be appropriate for relatively modest studies involving fewer experts.

The various guidance documents which sought to constrain work associated with the high-level waste programme signified that the goal was not to establish a rigid set of rules for eliciting expert judgement but, rather, to draw from experience – both successes and failures – criteria for identifying when expert judgement should be used, and to outline approaches for motivating, eliciting and documenting expert judgements. In this, there

was recognition that alternative approaches to the formal or informal use of expert judgement are available (e.g. Meyer and Booker, 1991, republished as Meyer and Booker, 2001).

It is outside the scope of the present chapter to go into all the detailed procedures that were used for the YMP studies, which are fully described in the literature (e.g. Kotra *et al*., 1996).

4.3.3 The Classical Model: a foundational example from meteorology

In October 1980, an experiment involving subjective probability forecasts was initiated at a weather station of the Royal Dutch Meteorological Institute. The primary objective of the experiment was to investigate the ability of the forecasters to formulate reliable and skilful forecasts of the probability that certain critical threshold values of important meteorological variables would (or would not) be exceeded. The trial involved prognostications of wind speed, visibility and rainfall, and went on for more than six years; the experts' judgements on future values of these variables were scored and combined with the *Classical Model* (Cooke, 1987, 1991).

The most important conclusions to be drawn from this investigation were that probabilistic expert assessments can provide a meaningful source of information, and that the quality of expert assessments drops off as the events being assessed extend further and further into the future. Also, the chosen model for combining expert judgement leads to a predictive performance that is markedly better than those of the individual experts themselves. Thus, it was found that there is every reason to consult more than one expert and, then, combine their judgements. Curiously, it was also found that the judgements of an individual expert with a good information foundation can be improved when they are combined with the assessments of an expert with an inferior information source. While this result seems counter-intuitive, it emerges overwhelmingly from the study, regardless of which framework was used for combining judgements.

Cooke (1991) emphasises the importance of defining appropriate scoring variables to measure performance: proper scoring rules[8] for individual assessments are really tools for eliciting probabilities. Generalisations of such scoring variables are applied in the Classical Model whose 'synthetic decision-maker' was seen to outperform results of simple arithmetic or geometric averaging. This result is significant in light of the work of Clemen and Winkler (1986), for instance, according to which simple models, such as the arithmetic and geometric averages, were considered to be more reliable on large datasets, including weather forecasting, than other more complicated combination models. From the very comprehensive Dutch meteorological study, however, it is apparent that it is possible to do better than simple averaging or equal weights.

[8] In decision theory a scoring rule is a measure of the performance of a person who makes repeated decisions under uncertainty. A rule is said to be 'proper' if it is optimised for well-calibrated probability assessments and 'strictly proper' if it is uniquely maximised at this point.

4.3.4 The Classical Model and airline pilots

Another extensive application of the Classical Model was undertaken with a large corps of senior airline pilots, in connection with the elicitation of opinions on civil aviation safety.[9] In this study, over 40 captains were calibrated using the scoring rules of the Classical Model. The basis of the calibration exercise was a suite of ten seed questions related to the topic of interest for which precise quantified realisations were available. The pilots were tested for their ability to estimate swiftly these quantities and their uncertainties on the basis of their experience, reasoning and beliefs. While there were wide scatters of opinions on questions (sometimes surprisingly so for professionals uniformly trained and working under identical controlled conditions), the application of the scoring rules to generate the weighted combination 'decision-maker' invariably produced central estimates that were within a few per cent of true values. This outcome demonstrated, to the satisfaction of the airline's flight operations management, that the technique provides decision guidance that is sufficiently accurate for most practical purposes.

A similar conclusion, regarding the utility of the classical model methodology in respect of provision of scientific advice during a volcanic eruption crisis (see Section 4.3.5), was reached by Aspinall and Cooke (1998).

4.3.5 Montserrat volcano and the Classical Model

Expert elicitation has been an important component in the provision of scientific advice during the volcanic eruption crisis (Aspinall, 2006) in Montserrat, West Indies, where a volcano has been erupting, occasionally violently, on a small, populated island for more than 15 years. Initially, when the volcano first became restless in 1995, about ten volcanologists became involved in advising the local government on what possible hazards might materialise, so the authorities could determine warning levels, travel restrictions and evacuations. As activity increased over subsequent years, when lethal hot block-and-ash flows spewed from the volcano at speeds of 150 km hr, destroying everything in their path, more and more scientists became involved in the challenge of assessing the volcano's possible behaviour (Aspinall *et al*., 2002).

This was the first time a formalised elicitation procedure was used in a live volcanic crisis. It allowed the disparate group of specialists to make rapid joint assessments of hazard levels as conditions at the volcano changed, sometimes dramatically, helping forecast conditions a few days or weeks in the future. In most cases, these turned out to reasonably match what happened (Aspinall and Cooke, 1998).

The relative speed with which such elicitations can be conducted, compared with traditional consensus processes, is one of its great advantages in such situations. Another is that it provides a route for experts uneasy about participating, or inexperienced with offering policy advice, to get involved in decision-making processes under the guidance of a neutral

[9] Unpublished report to British Airways by Aspinall & Associates, 1997 (WP Aspinall, pers. comm.).

facilitator, without feeling obliged to strive for complete agreement. Because no absolute consensus *sensu stricto* is imposed on the group by the Classical Model, an individual expert is not compelled to adopt the outcome as his or her personal belief. In practice, however, it is rare to meet substantial dissent; a more typical reaction is: 'Those aren't exactly the probability values I would have chosen but, for our present purposes, I can't object to any of them' (Aspinall and Cooke, 1998).

The Montserrat example evinces a spread of differential expert weights, typical of Classical Model results: using a set of seed items concerning facts about volcanic hazards, expert scores for 18 volcanologists are shown in Figure 4.4, together with a simple ordinal index of their experience of such crises. Characteristically (and comfortingly for grey-beards), veteran experts by and large achieve the highest scores, as might be hoped – over the years, they have been imbued with a functioning and reliable feel for uncertainty. But, there are clear exceptions to the principle – in Figure 4.4, experts 17 and 11 punch above

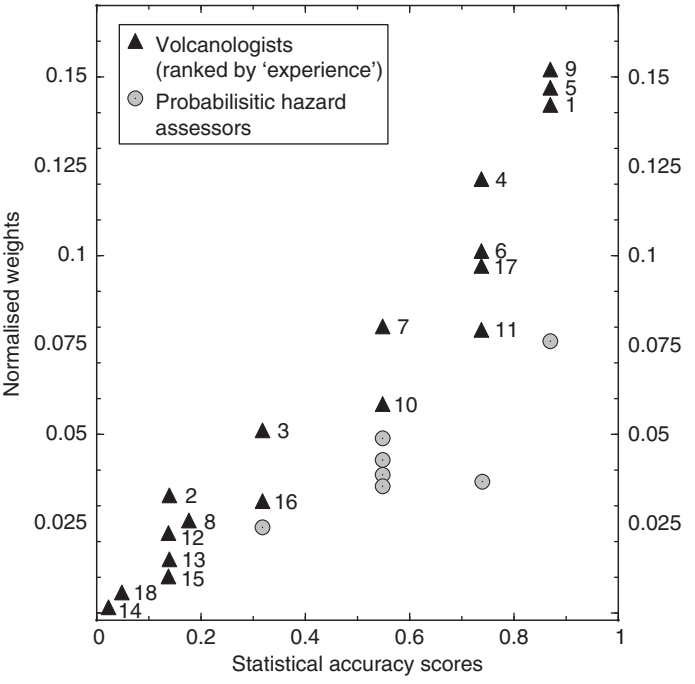

Figure 4.4 Profile of expert scores for 18 volcanologists (triangles) involved in the early stages of the Montserrat volcano crisis (see text). The horizontal axis shows individual statistical accuracy (calibration) scores obtained for a set of seed items on volcanic hazards, while the vertical axis (normalised weights) plots the resulting overall weight ascribed to each expert when individual statistical accuracy is factored by informativeness score; the weights are normalised to sum to unity across the whole group. A rudimentary ordinal ranking of individual volcanologist's experience is given (1–18). Circular markers show the corresponding scores achieved by a group of seven earth scientists, knowledgeable about probabilistic hazard assessment but not specialists in volcanology.

their expected weights, given their relative inexperience, while senior experts 2 and 3 fall down the rankings, being penalised for failing adequately to assess uncertainty ranges or for being statistically inaccurate – or both!

In a comparison exercise, the same elicitation seed items were presented to seven earth scientists who specialised in probabilistic seismic hazard assessment, but not volcanology *per se*. All seven of these non-specialists fared better than a few low-weight volcanologists (see Figure 4.4), but even the best is out-scored by the eight top-weighted volcanologists; the interpretation here is that the hazard analysts' deficit of subject-matter knowledge was offset by their understanding of how uncertainties play into probabilistic reasoning – but not to the extent that this made any of them into genuine experts on uncertainties relating to volcanic hazards.

This last case exemplifies a common denominator in any elicitation of expert judgement: there is no way of knowing for certain, a priori, who will achieve a worthwhile performance-based score until experts are objectively calibrated on some empirical basis, and whose opinions should be marked down for want of sound reasoning about uncertainty (for corroboration of this point, see Flandoli *et al*., 2011). Establishing what can, and should, be said collectively about scientific uncertainties can, in turn, have important potential impact on how such uncertainties are communicated beyond the confines of an expert group.

4.4 Communicating expert uncertainty

Communicating scientific uncertainty to policy-makers is an area where there is perhaps a significant obstacle to wider adoption of structured expert judgements in decision-support. Some feel that better information from an expert perspective – more complete in terms of quantified uncertainty, but more complex overall – may confound decision-makers rather than promote better decisions, so that improvements in capturing uncertainty in risk assessments have to be matched with improvements in communication.

For instance, Gigerenzer *et al*. (2005), discussing weather forecasts, point out that the quantitative statement 'there is a 30% chance of rain' is assumed to be unambiguous and to convey more information than does a qualitative statement like 'It might rain tomorrow'. Because the forecast is expressed as a single-event probability, however, it does not specify the class of events it refers to, and a bald numerical probability like this can be interpreted in multiple, mutually contradictory ways. To improve risk communication, Gigerenzer *et al*. suggest experts need to operationalise forecasts[10] by specifying, quite precisely, the reference class they are working with – that is, the class of events to which a single-event probability refers.

Other studies suggest most decision-makers welcome more information, not less. If technical information is presented in context, alongside training in uncertainty concepts, it can help decision-makers appraise the confidence they should have in their decisions. As discussed in Chapter 8 of this volume, given the depth and variety of uncertainties involved

[10] In agreement with the discussion in Section 4.2.3, *supra*.

in a probabilistic seismic hazard assessment, experts' views are an unavoidable component of the assessment process. Where a decision has to be made on the basis of very uncertain evidence – which is inevitably the case for low-probability scenarios – the only practical recourse is to make use of relevant expert judgement. Recognising this to be the case, the use of expert judgement is, nowadays, an accepted feature of such studies (see, for example, IAEA, 2010) – although how this is mobilised is not laid down, and remains a matter of debate, and some contentiousness.

This said, it is also important to distinguish expert judgement from engineering or other technical judgement, as the differences between them can be sufficient to affect system safety in a tangible way (Skipp and Woo, 1993).

4.4.1 Expert judgement and engineering judgement

Expert judgement, properly elicited, is not arbitrary but must satisfy various fundamental principles. In particular, the judgement must be coherent with the axioms of probability theory which provide systematic rules for the synthesis of subjective opinions.

By contrast, engineering judgement, in standard civil engineering practice for instance, is a less tangible and less well-defined concept, used in a more holistic fashion (Skipp and Woo, 1993). Thus, engineering judgement may be captured in colloquial statements of the form 'To do it that way will get you into difficulties', or, 'If you do it this way you will be alright.'

The recursive use of engineering judgement in estimating appropriate factors of safety defies quantification, except in the binary sense that inadequacy is ultimately exposed by failure. The main characteristics of engineering judgements are:

- not constrained by formal logic;
- cannot be easily calibrated;
- may incorporate conservatism;
- tend to be heuristic and holistic.

Differences of opinion over safety issues will always be present, but terminological confusions should be avoidable. Where an 'engineering judgement' is unwittingly offered in place of an 'expert judgement', or vice versa, problems can be expected. As probabilistic or Bayesian methods become more widely accepted as methods of weighing evidence, there will be an increasing need to clarify the nature of the judgement being offered (as note by Gigerenzer *et al*., 2005).

In the case of expert judgement in hazard assessments, more is required than some general awareness of the topic: the judgements need to be made by experts with extensive experience of doing hazard assessments in the relevant region. In principle, a rigorous mathematical check on the consistency of expert judgements with the fundamental laws of probability is afforded by comparative hazard assessments for a range of different sites across a region.

In practice, it often seems that the full ramifications of the use of expert judgement are not properly appreciated. Two salient comments can be made in this regard:

(1) the proper use of expert judgement means far more than simply adopting *ipso facto* a particular point of view which happens to have been expressed by an 'expert'; and
(2) even where formal procedures are used for making some decisions, other, similarly significant, decisions are very often afforded no special attention at all.

These points are both of considerable importance.

In the civil nuclear power industry, the way in which expert judgement is employed to come to decisions on all issues involving uncertainty is a matter which has long been a topic of considerable concern to the Nuclear Safety Directorate of HSE in the UK, and to their American counterparts, the United States Nuclear Regulatory Commission (US NRC).

Nowadays, regulatory authorities expect not only the process by which expert judgement is elicited and different opinions resolved to be satisfactory, they also require appropriate attention to be paid to the reporting of that process (in terms of an auditable 'trail') because legitimate divergence of opinion and lack of scientific consensus have repercussions on regulatory decision-making (Kelly *et al*., 1991).

Perhaps more than anywhere else, it is in regulatory areas that the search for formal methods of eliciting expert judgement and aggregating individual opinions has been motivated, with a need to find ways of handling divergent opinions within safety cases and for underpinning risk-informed policy setting.

4.4.2 Scientific uncertainty and policy setting

While it should be self-evident that uncertainty is an unavoidable ingredient in every instance of science-informed decision-making, this is not always recognised adequately, and can involve challenging complications for scientists and policy-makers alike. For instance, on one side there may be very knowledgeable scientists who are reluctant to offer opinions or expert judgements when the topic entails direct and important societal decisions which they may consider marginal to their expertise or inimical to their independence.

With scientists, too, there can be a reluctance that is best described as aversion to 'epistemic risk' – the fear of being proved 'wrong' (Parascandola, 2010). Empirical scientists, in particular, traditionally have taken a conservative approach towards epistemic risk by minimising the extent to which they push inferences beyond observations. Twentieth-century scientists and philosophers developed methods for managing epistemic risk within the scientific endeavour, by quantifying degrees of uncertainty and evidential support. In the biomedical sciences, for example, the principal strategy has been the avoidance of Type I error (i.e. rejecting the null hypothesis when it is true, or finding a false positive in other test situations), using the p-value and some statistical significance level as a criterion (usually $p < 0.05$).

Nearly all scientists (with some unscrupulous exceptions) wish to avoid a situation where fallacies or errors turn out to be present in their work. So, for some, there can be a reluctance to state something which might be crucial for a decision-taker because they fear it may be overturned eventually by new data or evidence. For tricky problems, academic scientists sometimes have the luxury of suspending judgement to mitigate epistemic risk but, once taken outside science into, say, wider society or a policy forum, this aversion can lead to bias, or even advice paralysis, when the issue concerns extremely low-probability events with significant consequences. In such circumstances, data are inevitably sparse and frequently the decision to be made is stark and dichotomous – evacuate the population or not – and scientific conservatism may be a mistaken strategy. The weakness of evidence linking special volcano-seismic events called tornillos to explosive eruptions at Galeras volcano, Colombia, prevented their significance being appreciated in 1993, with fatal results for scientists in the crater; the difficulty was one of weighing the possible implications of novel and meagre evidence, with large uncertainties, in a non-scientific context (Aspinall *et al*., 2003).

There is an almost endless list of life-and-death situations where such science-related concerns apply: approving a clinical trial for a new drug for human use; directing research in biomedicine, where ethical and controversial issues arise and deciding costly but only partially effective measures for the mitigation of natural hazards. It is, therefore, imperative in all such cases that scientific uncertainty is rationally evaluated, and not exaggerated or dismissed to minimise epistemic risk, nor left untreated for decisions to fall prey to the precautionary principle.

The counterpart to epistemic risk in the domain of the decision-maker is 'volitional risk'. Put crudely, this risk amounts to weighing the likelihood that a decision will appear foolish or stupid post hoc – an aversion surely endemic in politicians. In the face of this risk, a decision-taker will be strongly influenced to avoid putting himself or herself in a position where their judgement can be questioned. To do so, some decision-takers may choose to discount the implications of scientific uncertainty in a risk assessment or, as noted earlier, find a way to offload responsibility onto the hapless scientist. Others may reject uncertainty considerations because of fears they will undermine confidence in policy decisions or, where scientific doubt clearly exists, open up related regulations to legal challenge.

While volitional risk has been aired in the context of the recent banking crisis (where usual restraining effects were seemingly offset by 'moral hazard' – the prospect of being not allowed to fail), considerations of epistemic and volitional risk have received little attention for their role in, and influence on, the provision of scientific advice; this is a topic that might merit research.

Formalised expert elicitation, when it utilises a performance-based pooling algorithm, can help bridge these various positions and different perspectives, and can strengthen confidence in, and the credibility of, scientific advice in a policy arena. As just noted, politicians in particular are averse to uncertainty or ambiguity, so when expert elicitations indicate margins of uncertainty on scientific questions that are wider than hoped for, that information can be difficult to digest. For example, expert assessments of real-world problems with

strong political ramifications, such as SARS exposure risks to healthcare workers or vCJD infectivity from blood products (e.g. Tyshenko *et al*., 2011), often manifest more uncertainty about scientific knowledge than individual experts might express or problem owners might infer. This said, such exercises are found beneficial by problem owners; in the case of assessing risks due to vCJD infection via blood products, the organisers of a follow-on policy workshop indicated strong approbation of the method.[11]

In addition to the valuable attributes listed above, a structured elicitation habitually identifies certain key issues that can get overlooked with other decision-support procedures.

Another important trait of the performance-based Classical Model procedure is that the estimated overall uncertainty on a particular scientific judgement from a group of experts is often smaller than that derived from equal weights pooling (see, e.g. Figure 4.1), thereby increasing credibility with policy-makers. But providing a candid, unbiased presentation of the true spread of expert uncertainty remains essential for the decision process. Such uncertainty cannot be reduced to zero, even if individual experts in oracular mode (Section 4.2.1) may be inclined to play it down for self or professional aggrandisement.

If an elicitation has been geared purposely to quantifying scientific uncertainty as reliably as possible – given the nature of the problem and the current state of understanding – then communicating this to a problem owner as an authentic expression of the true extent of uncertainty can sometimes lead to acceptance issues: the scientific message, for his or her purposes, may be less clear-cut than hoped for!

4.5 Future directions

While different forms of structured expert elicitations have been tried in certain difficult regulatory decision domains – where scientific or engineering uncertainty is a fundamental element – such elicitations are also starting to take root in other areas, such as climate-change impact on fisheries (Rothlisberger *et al*., 2009), and invasive species threats (Neslo *et al*., 2009). Recently, the US EPA formed a task force (EPA, 2009) to recommend ways of using expert elicitations in regulatory issues whenever data are sparse, such as deciding dose criteria for air quality standards; some pioneering expert judgement work has been applied in this context (Tuomisto *et al*., 2008).

This expanding world of structured elicitation of experts is bringing to light a number of topics meriting direct research on the methodologies concerned.

[11] 'The primary benefit being that the [elicitation] exercise itself communicated several previously undetermined values with ranges for infectious disease uncertainty gaps. Elicitation results targeted iatrogenic transmission through surgical instruments reuse, genotype effects, incubation times, infectivity of blood and a ranking of various iatrogenic routes. The usefulness of the structured expert elicitation that uses a mathematical formalism was seen as a highly useful process that could be applied to other public health risk issues. During the open discussion the benefits identified by the workshop participants included its: 1) utility to help prioritize issues within and between different government departments; 2) application for emerging near-term threats such as pandemic influenza, swine flu and avian flu; 3) utility for future, longer-term foresight exercises; 4) value as a process in increasing overall transparency; 5) ability to reduce bias in decision making; 6) use as a tool to provide proxy values for risk assessors until evidence-based data become available; 7) capacity as a recorded exercise to help support policies decisions; 8) application to improve risk assessments and risk modelling for which data may be missing, inadequate or questionable; and 9) usefulness in delivering values in a shorter time for issues with high uncertainty.' Accessed January 2010 on the PRIONET website, http://www.prionetcanada.ca/

4.5.1 Research on expert judgement modelling

The following are just a few sample topics relating to expert elicitations that merit further research.

4.5.1.1 Scoring rules

Interest in single-variable scoring rules for combination (as opposed to the Classical Model, which is a grouped-variable scoring rule) is reviving after some of Hora's earlier publications. For instance, in 2004, Hora showed that an equal weight combination of well-calibrated experts is not well calibrated. Lin and Cheng (2008) have started using the Brier score, a forecast skill metric from meteorology. Wisse (2008) developed a linear Bayes moments method approach, advocating expectations as a basis, rather than probabilities; partly this was to reduce the computational burden that decision models involving continuous probability distributions carry, although this latter method has, thus far, failed to gain traction in practice.

4.5.1.2 Seed item choice

Further research would also be helpful into what properties make a set of seed questions suitable for calibration. Does an expert's performance on the seed questions faithfully predict their performance when it comes to the real problem under discussion? This is difficult to study with probabilistic predictions, such as in the volcano eruption problem, but appropriate seed questions need to enhance the credibility of the final result.

4.5.1.3 Empirical expert judgement studies

There have been a few, albeit limited, efforts to use the TU Delft database as a research resource, e.g. in the RESS special issue (Cooke, 2008). However, Lin and Cheng (2009) have been following this up. There is much more still to do here, especially in studying the properties of remove-one-at-a-time cross validation. Also needed are more published 'stories from the trenches' about how to configure an expert elicitation problem, how to structure the expert group – analyst interaction, how to provide feedback to the experts and problem owners, and finally, how to communicate uncertainty to the various stakeholders.

4.5.1.4 Fitting models with expert judgement

In nuclear safety, various authorities base radiation exposure guidelines on biological transport models, foreseen with numerous transport coefficients. However, the following problem presented itself: formal uncertainty analysis required a joint distribution over these transfer coefficients, but the experts involved were unable or unwilling to quantify their uncertainty over these parameters, claiming these were too far removed from their empirical knowledge. The solution was to quantify uncertainty just on certain observable quantities, which could be predicted by the models, and then pull this uncertainty back onto the model parameters in a process known as probabilistic inversion. The potential application of such techniques is huge, and there is much room for advancing the procedure where expert

judgements are involved. Kraan and Bedford (2005), and Kurowicka and Cooke (2006) treat probabilistic inversion in detail.

4.5.1.5 Dependence modelling/dependence elicitation

Another issue that emerged in the nuclear safety work is dependence elicitation. Prior to this, dependence in elicited uncertainties was largely ignored, as if all important dependencies were captured in functional relationships. Of course, this is not remotely true, as best illustrated with a few examples from the Joint USNCR, ERU Studies:

- the uncertainties in effectiveness of supportive treatment for high radiation exposure in people over 40 and in people under 40;
- the amount of radioactivity after one month in the muscle of beef and dairy cattle;
- the transport of radionuclides through different soil types.

The joint study had to break new ground in dependence modelling and dependence elicitation, and these subjects are treated extensively in the documentation (Goossens and Harper, 1998). The format for eliciting dependence was to ask about joint exceedence probabilities: 'Suppose the effectiveness of supportive treatment in people over 40 was observed to be above the median value, what is your probability that also the effectiveness of supportive treatment in people under 40 would be above its median value?' Experts became quickly familiar with this format. Dependent bivariate distributions were found by taking the minimally informative copula which reproduced these exceedence probabilities, and linking these together in a Markov tree structure. These techniques can and should be developed further.

4.5.1.6 Stakeholder preference modelling

A sharp distinction can be made between uncertainty quantification and preference modelling. The latter has become an active topic of late. Expert discrete choice data can be elicited in the form of rankings or pair-wise comparisons, and probabilistic inversion used to fit some model, typically a multicriteria decision model (MCDM). These models represent the value of an alternative as a weighted sum of scores on various criteria. With MCDM, a distribution over criteria weights is obtained that optimally recovers distributions of stakeholder rankings. The discrete choice data are split into a training set and a validation set, thereby enabling out-of-sample validation. There are many interesting research issues here, including how best to configure the discrete choice elicitation format, and how best to perform validation.

4.5.2 Research on elicitation methodology

There are several open research questions on expert elicitation methodological issues that will need to be answered if this fledgling field is to build its body of practice and become

recognised more widely as a robust methodology. French (2011) provides a valuable and timely review.

4.5.2.1 Expert panel selection

For example, the selection and number of experts to compose a group for elicitation needs careful thought. The problem owner may want the selection to be as impartial and independent as possible to prevent criticism of bias, or they may restrict the selection to company employees to ensure confidentiality and relevance.

To-date, experience with expert panels suggests that 8–15 experts is a reasonable, and viable, number for tackling a given problem, but this has not been rigorously tested. With more than 20 people involved, a situation of diminishing returns seems to ensue; a limited-size group, selected meticulously by the problem owner, can capture the wider collective view without compromising informativeness or calibration accuracy, and thus helps constrain investment of people's time and commitment in an elicitation exercise.

4.5.2.2 New protocols for distance elicitation

To-date, significant departures from established elicitation protocols have not been a big issue. This said, the existing procedures can be labour intensive and group elicitations, in particular, can be time-consuming and, in the case of a global discipline such as volcanology, often difficult to convene for reasonable cost. Eventually, opportunities to take advantage of new media technologies and to find ways and means of using them to conduct tele-elicitations – without loss of formalism and structure – will need to be explored. Then, a good understanding of, and perhaps training in, the principles and concepts will be even more essential – for facilitators and experts, alike.

4.5.2.3 Target item question framing

Sometimes an elicitation reveals two camps of opinion within a group, making it difficult to derive a rational consensus value. In such cases, the dichotomy usually arises either from an ambiguity in question framing, or because the two groups have different experience references. Detecting that this condition may exist within a group elicitation is a powerful diagnostic property of the structured elicitation process, and the difficulty can usually be resolved by facilitated discussion or by reframing the target question. It is very common to encounter the need for such clarification among a group of experts, however close-knit they are in terms of specific scientific interests (e.g. seismologists) or professional training (e.g. airline pilots).

4.5.2.4 Other methodology issues

Additional existing concerns that deserve research include:

(1) Could training experts improve their performance?
(2) What are the pros and cons of different structures for conducting an elicitation?

(3) Can more qualitative aspects of human behaviour (e.g. social vulnerability information) be factored into assessments of risk using expert judgement?
(4) Can expert elicitation be studied usefully in terms of group dynamics and psychology?

Lastly, are there alternatives to quantitative elicitation based on the use of a few defining quantiles – e.g. ranking, ordering or probabilistic inversion? And under what circumstances would such alternative formulations or frameworks be regarded as more appropriate than the quantitative distributional approach?

4.5.3 Teaching and expert learning

It is tempting to think that an expert might be able to improve his or her expert judgement performance in the Classical Model sense through participating in elicitations. Conventionally, expert learning is not a proper goal for an expert judgement elicitation; rather, the problem owner should want experts who are proficient in judging uncertainties, independent of the elicitation process itself. Nor, in principle, should they be learning the fundamentals of the subject-matter from the elicitation. However, in discussing and framing target questions in the preamble to elicitation, experts may develop a better sense of the relevant uncertainties from a facilitated exchange of new information, status reviews and data summaries.

If, after responding to seed questions, an expert were to learn that all, or many of the actual realisations fell outside his 90% confidence intervals, he might conclude that his intervals had been too narrow, and try to broaden them in subsequent assessments. Weak anecdotal evidence from repeated uncertainty assessments – as in the volcano case above – suggests that, with a few notable exceptions, self-directed training in this regard is difficult: expert leopards do not find it easy to change their uncertainty spots!

Perhaps this is right and proper, when experts are enjoined to express their true beliefs, but as applications of the Classical Model method proliferate, expert learning in this context should become a topic meriting detailed research. New findings could inform approaches to teaching and training.

Some recent pioneering efforts have been made towards teaching the principles of expert elicitation techniques at undergraduate and post-graduate level (Stone *et al.*, 2011). This limited experience suggests that introducing the concepts to young students, letting them actively apply the elicitation procedures and use performance analysis algorithms, such as the Classical Model, are a powerful way of schooling them in how demanding it can be to evaluate scientific uncertainty convincingly, reliably and defensibly.

It is clear that concepts like calibration, entropy and mutual information – and their roles in quantifying scientific uncertainty in a decision support context – are widely unfamiliar to the majority of scientists, young and old alike, so training and teaching initiatives are needed.

4.6 Summing up

Notwithstanding all the diversity of topics touched on above, and the many excursions that the discussion has taken, what can be said with confidence is that combining a structured elicitation of expert opinion with a formalised scoring process is an affirmative way of catalysing scientific discussion of uncertainty, and provides a rational approach for characterising and quantifying hazard and risk assessments in a robust way. Structured elicitation procedures can help identify, pose and respond to many natural hazards questions where scientific uncertainty is pervasive: our judgement is that the range and compass of the approach is almost unlimited.

Experience indicates also that the structured elicitation process that feeds into formalised judgement pooling with the Classical Model is a highly effective way of framing scientific debates, and that some mutual learning takes place among a group with consequent improved generic understanding of uncertainty. These are central and valuable features of this particular expert elicitation procedure, helping improve scientific advice for decision-making.

As mentioned earlier in Section 4.2.2, one impediment to the wider take-up of such procedures can be the cost of an elicitation in terms of human time and resources, and funding for participation, travel and meetings. The time of high-level specialists, especially those internationally recognised, is always at a premium. Thus cost and time frequently prove practical limitations, with problem owners often reluctant to furnish sufficient funding to do the job properly and comprehensively. Moreover, with many important science-based decision issues it can be a challenge both to constrain a problem owner's expectations to manageable proportions, and to rein in the experts' almost inevitable desires to tackle a wide scope and large number of target questions. Extensive experience indicates that a well-designed exercise with a practised facilitator can tackle 10–30 scientifically tricky target items over a two-day period; attempting more may entail 'elicitation fatigue'. This said, the approach has been used successfully in more than 50 studies and is rapidly gaining academic credibility, professional acceptance and even political standing.

Acknowledgements

Our thanks go to R. S. J. Sparks in particular, and to J. C. Rougier and J. Freer (University of Bristol) and G. Woo (Risk Management Solutions Ltd.) for discussions, suggestions, improvements and generous encouragement. As with any errors, the views expressed are those of the authors, and should not be construed as representing the opinions or positions of any group or organisation with whom they may be or have been associated. WPA was supported in part by a European Research Council grant to Prof. R. S. J. Sparks FRS: VOLDIES Dynamics of Volcanoes and their Impact on the Environment and Society.

References

Abrahamson, N. A., Coppersmith, K. J., Koller, M., *et al*. (2004) Probabilistic seismic hazard analysis for Swiss nuclear power plant sites, PEGASOS Project, Vols 1–6, NAGRA.

ACERA (2007) Eliciting expert judgments: literature review. http://www.acera.unimelb.edu.au/materials/endorsed/0611.pdf (accessed 28 December 2011).

ACERA (2010) Elicitation tool: process manual. http://www.acera.unimelb.edu.au/materials/software.html (accessed 28 December 2011).

Aspinall, W. P. (2006) Structured elicitation of expert judgment for probabilistic hazard and risk assessment in volcanic eruptions. In *Statistics in Volcanology*, ed. H. M. Mader, S. G. Coles, C. B. Connor and L. J. Connor, London: The Geological Society for IAVCEI, pp. 15–30.

Aspinall, W. (2010) Opinion: a route to more tractable expert advice. *Nature* **463**: 294–295.

Aspinall, W. and Cooke, R. M. (1998) Expert judgment and the Montserrat Volcano eruption. In *Proceedings, 4th International Conference on Probabilistic Safety Assessment and Management PSAM4*, ed. Ali Mosleh and Robert A. Bari, vol. 3, New York, NY: Springer, pp. 2113–2118.

Aspinall, W. P., Loughlin, S. C., Michael, A. D., *et al*. (2002) The Montserrat Volcano Observatory; its evolution, organization, role and activities. In *The Eruption of Soufrière Hills Volcano, Montserrat from 1995 to 1999*, ed. T. H. Druitt and B. P. Kokelaar, London: Geological Society of London, pp. 71–91.

Aspinall, W. P., Woo, G., Voight, B., *et al*. (2003) Evidence-based volcanology; application to eruption crises. *Journal of Volcanology and Geothermal Research: Putting Volcano Seismology in a Physical Context; In Memory of Bruno Martinelli* **128**: 273–285.

Brockhoff, K. (1975) The performance of forecasting groups in computer dialogue and face to face discussions. In *The Delphi Method, Techniques and Applications*, ed. H. A. Linstone and M. Turoff, Reading, MA: Addison Wesley, pp. 291–321.

Brown, A. J. and Aspinall, W. P. (2004) Use of expert elicitation to quantify the internal erosion processes in dams. In *Proceedings of the British Dam Society Conference: Long-term Benefits and performance of Dams*, Canterbury: Thomas Telford, pp. 282–297.

Budnitz, R., Apostolakis, G., Boore, D. M., *et al*. (1998) Use of technical expert panels: applications to probabilistic seismic hazard analysis. *Risk Analysis* **18**: 463–469.

Budnitz, R. J., C. A. Cornell and Morris, P. A. (2005) Comment on J.U. Klugel's 'Problems in the application of the SSHAC probability method for assessing earthquake hazards at Swiss nuclear power plants,' in Engineering Geology, vol. 78: 285–307. *Engineering Geology* **82**: 76–78.

Clemen, R. T. and Reilly, T. (1999) Correlations and copulas for decision and risk analysis. *Management Science* **45**: 208–224.

Clemen, R. T. and Winkler, R. L. (1986) Combining economic forecasts. *Journal of Business & Economic Statistics* **4** (1): 39–46.

Clemen, R. T. and Winkler, R. L. (1987) Calibrating and combining precipitation probability forecasts. In *Probability & Bayesian Statistics*, ed. Viertl, R.., New York, NY: Plenum Press, pp. 97–110.

Clemen, R. T. and Winkler, R. L. (1999) Combining probability distributions from experts in risk analysis. *Risk Analysis* **19**: 187–203.

Cooke, R. (1987) A theory of weights for combining expert opinions. Delft University of Technology, Department of Mathematics and Informatics, Report 87–25.

Cooke, R. M. (1991) *Experts in Uncertainty:Opinion and Subjective Probability in Science*. Oxford: Oxford University Press.

Cooke, R. M. (2008) Special issue on expert judgment. *Reliability Engineering & System Safety* **93**: 655–656.

Cooke, R. M. and Goossens, L. L. H. J. (2008) TU Delft expert judgment data base. *Reliability Engineering & System Safety* **93**: 657–674.

Cooke, R. M., Mendel, M. and Thijs, W. (1988) Calibration and information in expert resolution. *Automatica* **24**: 87–94.

Cooke, R. M., ElSaadany, S. and Huang, X. (2008) On the performance of social network and likelihood-based expert weighting schemes. *Reliability Engineering & System Safety* **93**: 745–756.

Coppersmith, K. J., Youngs, R. R. and Sprecher, C. (2009) Methodology and main results of seismic source characterization for the PEGASOS Project, Switzerland. *Swiss Journal of Geosciences* **102**: 91–105.

De Groot, M. (1974) Reaching a consensus. *Journal of the American Statistical Association* **69**: 118–121.

Delbecq, A. L., Van de Ven, A. and Gusstafson, D. (1975) *Group Techniques for Program Planning*. Glenview, IL: Scott, Foresman.

EPA (2009) Expert elicitation white paper: external review draft, January. http://www.epa.gov/spc/expertelicitation/index.htm (accessed 6 November 2009).

Fischer, G. (1975) An experimental study of four procedures for aggregating subjective probability assessments. Technical Report 75–7, Decisions and Designs Incl., McLean, Virginia.

Flandoli, F., Giorgi, E., Aspinall, W. P., *et al*. (2011) Comparison of a new expert elicitation model with the Classical Model, equal weights and single experts, using a cross-validation technique. *Reliability Engineering & System Safety* **96**: 1292–1310.

French, S. (2008) Comments by Prof. French. *Reliability Engineering & System Safety Expert Judgment* **93**: 766–768.

French, S. (2011) Aggregating expert judgment. *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas (RACSAM)* **105**: 181–206.

Gigerenzer, G., Hertwig, R., van den Broek, E., *et al*. (2005) 'A 30% chance of rain tomorrow': how does the public understand probabilistic weather forecasts? *Risk Analysis* **25**: 623–629.

Goossens, L. H. J. and Harper, F. T. (1998) Joint EC/USNRC expert judgment driven radiological protection uncertainty analysis. *Journal of Radiological Protection* **18**: 249–264.

Gough, R. (1975) The effect of group format on aggregate subjective probability distributions. In *Utility, Probability and Human Decision Making*, ed. D. Wendt and C. Vlek, Dordrecht, Reidel.

Hanks, T. C., Abrahamson, N. A., Boore, D. M., *et al*. (2009) Implementation of the SSHAC Guidelines for Level 3 and 4 PSHAs: experience gained from actual applications. US Geological Survey Open-File Report 2009–1093. http://pubs.er.usgs.gov/publication/ofr20091093 (accessed 27 July 2012).

Helmer, O. (1966) *Social Technology*, New York, NY: Basic Books.

Helmer, O. (1968) Analysis of the future: the Delphi method, and The Delphi Method–an illustration. In *Technological Forecasting for Industry and Government*, ed. J. Bright, Englewood Cliffs, NJ: Prentice Hall.

Hora, S. C. (2004) Probability judgments for continuous quantities: linear combinations and calibration. *Management Science* **50**: 597–604.

Hora, S. C. and Iman, R. L. (1989) Expert opinion in risk analysis: the NUREG-1150 experience. *Nuclear Science and Engineering* **102**: 323–331.

IAEA (2010) *Seismic Hazards in Site Evaluation for Nuclear Installations*. Vienna: International Atomic Energy Agency.

Kahneman, D. (2011) *Thinking, Fast and Slow*, London: Allen Lane.

Kahneman, D., Slovic, P. and Tversky, A. (eds) (1982) *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press.

Kallen, M. J. and Cooke, R. M. (2002) Expert aggregation with dependence. In *Probabilistic Safety Assessment and Management*, ed. E. J. Bonano, A. L. Camp, M. J. Majors, R. A. Thompson, London: Elsevier, pp. 1287–1294.

Kelly G. B., Kenneally R. M. and Chokshi N. C. (1991) Consideration of seismic events in severe accidents. In *Proceedings of Probabilistic Safety Assessment and Management* **2**, pp. 871–876.

Klügel, J.-U. (2005a) Problems in the application of the SSHAC probability method for assessing earthquake hazards at Swiss nuclear power plants. *Engineering Geology* **78**: 285–307.

Klügel, J.-U. (2005b) Reply to the comment of Krinitzsky on J.U. Klugel's 'Problems in the application of the SSHAC probability method for assessing earthquake hazards at Swiss nuclear power plants', in Engineering Geology, vol. 78: 285–307. *Engineering Geology* **82**: 69–70.

Klügel, J.-U. (2005c) Reply to the comment on J.U. Klugel's 'Problems in the application of the SSHAC probability method for assessing earthquake hazards at Swiss nuclear power plants', in Engineering Geology, Vol. 78: 285–307, by Budnitz, by, J.U, Klugel. *Engineering Geology* **82**: 79–85.

Klügel, J.-U. (2005d) Reply to the comment on J.U. Klugel's: 'Problems in the application of the SSHAC probability method for assessing earthquake hazards at Swiss nuclear power plants,' Eng. Geol. Vol. 78: 285–307, by Musson *et al*. *Engineering Geology* **82**: 56–65.

Klügel, J.-U. (2005e) Reply to the comment on J. U. Klugel's: 'Problems in the application of the SSHAC probability method for assessing earthquake hazards at Swiss nuclear power plants,' in Engineering Geology, vol. 78: 285–307, by Lomnitz, by J.U. Klugel. *Engineering Geology* **82**: 74–75.

Klügel, J.-U. (2005f) Reply to the comment on J. U. Klugel's: Problems in the application of the SSHAC probability method for assessing earthquake hazards at Swiss nuclear power plants, in Engineering Geology, Vol. 78: 285–307, by Wang, by J.U, Klugel. *Engineering Geology* **82**: 89–90.

Klügel, J.-U. (2007) Error inflation in probabilistic seismic hazard analysis. *Engineering Geology* **90**: 186–192.

Klügel, J.-U. (2009) Probabilistic seismic hazard analysis for nuclear power plants: current practice from a European perspective. *Nuclear Engineering and Technology* **41**: 1–12.

Klügel, J. U., Mualchin, L. and Panza, G. F. (2006) A scenario-based procedure for seismic risk analysis. *Engineering Geology* **88**: 22.

Kotra, J. P., Lee, M. P., Eisenberg, N. A., *et al*. (1996) *Branch Technical Position on the Use of Expert Elicitation in the High-Level Radioactive Waste Program*. NUREG-1563. Washington, DC: US Nuclear Regulatory Commission.

Kraan, B. C. P. and Bedford, T. J. (2005) Probabilistic inversion of expert judgments in the quantification of model uncertainty. *Management Science* **51** (6): 995–1006.

Krinitzsky, E. L. (2005) Comment on J.U. Klugel's 'Problems in the application of the SSHAC probability method for assessing earthquake hazards at Swiss nuclear power plants', in Engineering Geology, vol. 78: 285–307. *Engineering Geology* **82**: 66–68.

Kurowicka, D. and Cooke, R. (2006) *Uncertainty Analysis with High Dimensional Dependence Modelling*. Chichester: John Wiley and Sons.

Kynn, M. (2008) The 'heuristics and biases' in expert elicitation. *Journal of the Royal Statistical Society: Series A* **171**: 239–264.

Lehrer, K. and Wagner, C. G. (1981) *Rational Consensus in Science and Society*. Dordrecht: Reidel.

Lin, S.-W. and Bier, V. M. (2008) A study of expert overconfidence. *Reliability Engineering & System Safety* **93**: 775–777.

Lin, S.-W. and Cheng, C.-H. (2008) Can Cooke's Model sift out better experts and produce well-calibrated aggregated probabilities? 2008 *IEEE International Conference on Industrial Engineering and Engineering Management*, Singapore: IEEE Singapore Section, pp. 425–429.

Lin, S.-W. and Cheng, C.-H. (2009) The reliability of aggregated probability judgments obtained through Cooke's classical model. *Journal of Modelling in Management* **4**: 149–161.

Lindley, D. V. (1971) *Making Decisions*, London: John Wiley and Sons.

Lindley, D. (1983) Reconciliation of probability distributions. *Operations Research* **31**: 866–880.

Lomnitz, C. (2005) Comment on J.U. Klugel's 'Problems in the application of the SSHAC probability method for assessing earthquake hazards at Swiss nuclear power plants', in Engineering Geology, vol. 78: 285–307. *Engineering Geology* **82**: 71–73.

Mendel, M. and Sheridan, T. (1989) Filtering information from human experts. *IEEE Transactions on Systems, Man and Cybernetics* **19** (1): 6–16.

Meyer, M. A. and Booker, J. M. (2001) *Eliciting and Analyzing Expert Judgment: A Practical Guide*. Philadelphia, PA: ASA-SIAM.

Morgan, M. G. and Henrion, M. (1990) *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. New York, NY: Cambridge University Press.

Morris, P. (1974) Decision analysis expert use. *Management Science* **20**: 1233–1241.

Morris, P. (1977) Combining expert judgments: a Bayesian approach. *Management Science* **23**: 679–693.

Mosleh, A. and Apostolakis, G. (1982) Models for the use of expert opinions. Workshop on low probability high consequence risk analysis, Society for Risk Analysis, Arlington, VA, June.

Musson, R. M. W., Toro, G. R., Coppersmith, K. J., *et al*. (2005) Evaluating hazard results for Switzerland and how not to do it: A discussion of 'Problems in the application of the SSHAC probability method for assessing earthquake hazards at Swiss nuclear power plants' by J-U Klugel. *Engineering Geology* **82**: 43–55.

Neslo, R., Micheli, F., Kappel, C. V., *et al*. (2009) Modeling stakeholder preferences with probabilistic inversion: application to prioritizing marine ecosystem vulnerabilities. In *Real Time and Deliberative Decision Making: Application to Risk Assessment for Non-chemical Stressors*, ed. I. Linkov, E. A. Ferguson and V. S. Magar, Amsterdam: Springer: pp. 265–284.

O'Hagan, A., Buck, C. E., Daneshkhah, A., *et al*. (2006) *Uncertain Judgments: Eliciting Experts' Probabilities*. Chichester: John Wiley and Sons.

Parascandola, M. (2010) Epistemic risk: empirical science and the fear of being wrong. *Law, Probability and Risk* **9**: 201–214.

Rothlisberger, J. D., Lodge, D. M., Cooke, R. M., *et al*. (2009) Frontiers in ecology and the environment. *The Ecological Society of America*. www.frontiersinecology.org (accessed 27 July 2012).

Sackman, H. (1975) *Delphi Critique, Expert Opinion, Forecasting and Group Processes*, Lexington, MA: Lexington Books,.

Seaver, D. (1977) How groups can assess uncertainty. *Proceedings of the International Conference on Cybernetics and Society*, New York: Institute of Electrical and Electronics Engineers, pp. 185–190.

Seaver, D. A. (1978) Assessing probabilities with multiple individuals: group interaction versus mathematical aggregation. Technical Report SSRI-78–3, Social Science Research Institute, University of Southern California, Los Angeles.

Skipp, B. O. and Woo, G. (1993). A question of judgement: expert or engineering? In *Risk and Reliability in Ground Engineering*, London: Thomas Telford, pp. 29–39.

SSHAC (1995) *Recommendations for Probabilistic Seismic Hazard Analysis: Guidance on Uncertainty and Use of Experts; prepared by Senior Seismic Hazard Analysis Committee (SSHAC)*, Livermore, CA: Lawrence Livermore National Laboratory.

SSHAC (1997) [Senior Seismic Hazard Analysis Committee, R. J. Budnitz, Chairman, G. Apostolakis, D. M. Boore, L. S. Cluff, K. J. Coppersmith, C. A. Cornell and P. A. Morris] *Recommendations for Probabilistic Seismic Hazard Analysis: Guidance on Uncertainty and Use of Experts*, NUREG/CR-6372, Washington, DC: US Nuclear Regulatory Commission.

Stone, J., Aspinall, W. P. and Sparks, R. S. J. (2011) Poster presented at Soufriere Hills Volcano: 15 Years on Conference, Montserrat, West Indies, April.

Tuomisto, J. T., Wilson, A., Evans, J. S., *et al*. (2008) Uncertainty in mortality response to airborne fine particulate matter: combining European air pollution experts. *Reliability Engineering & System Safety Expert Judgment* **93**: 732–744.

Tversky, A. and Kahneman, D. (2005) Judgment under uncertainty: heuristics and biases. In *Social Cognition Key Readings*, ed. D. L. Hamilton, New York, NY: Psychology Press, chapter 10.

Tyshenko, G. M., ElSaadany, S., Oraby, T., *et al*. (2011) Expert elicitation for the judgment of Prion disease risk uncertainties. *Journal of Toxicology and Environmental Health, Part A: Current Issues* **74**: 261–285.

Vick, S. G. (2002) *Degrees of Belief: Subjective Probability and Engineering Judgment*, Reston, VA: ASCE Press.

Wang, Z. (2005) Comment on J.U. Klugel's: Problems in the application of the SSHAC probability method for assessing earthquake hazards at Swiss nuclear power plants, in Engineering Geology, vol. 78: 285–307. *Engineering Geology* **82**: 86–88.

Winkler, R. L. (1981) Combining probability distributions from dependent information sources. *Management Science* **27**: 479–488.

Wisse, B., Bedford, T. and Quigley, J. (2008) Expert judgement combination using moment methods. *Reliability Engineering & System Safety Expert Judgement* **93**: 675–686.

Woo, G. (1999) *The Mathematics of Natural Catastrophes*, River Edge, NJ: Imperial College Press.

Woo, G. (2011) *Calculating Catastrophe*, River Edge, NJ: Imperial College Press.

Woudenberg, F. (1991) An evaluation of Delphi. *Technological Forecasting and Social Change* **40**: 131–150.

Yunusov, A. R., Cooke, R. M. and Krymsky, V. G. (1999) Rexcalibr-integrated system for processing expert judgment. In *Proceedings 9th Annual Conference: Risk Analysis: Facing the New Millenium*. Rotterdam, 10–13 October, ed. L. H. J. Goossens, Netherlands: Delft University, pp. 587–589.

Zuidema, P. (2006) PEGASOS: A PSHA for Swiss nuclear power plants–some comments from a managerial point of view. OECD-IAEA Specialist meeting on the Seismic Probabilistic Safety Assessment of Nuclear Facilities, Seogwipo, Jeju Island, Korea.