Expert Judgment Elicitation using the Classical Model and EXCALIBUR

Briefing notes for Seventh Session of the Statistics and Risk Assessment Section's International Expert Advisory Group on Risk Modeling: Iterative Risk Assessment Processes for Policy Development Under Conditions of Uncertainty / Emerging Infectious Diseases: Round IV

Prepared by Willy Aspinall, 26 March 2008

1. THE ESSENTIALS

1.1 Introduction

Expert opinion is almost invariably sought when technical uncertainty impacts on an important decision process. Because such uncertainty is ubiquitous in scientific knowledge - if it were not, any decision related to the issue at hand would be obvious – there is the inescapable corollary that the experts themselves cannot be absolutely certain, and thus it is extremely unlikely 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 determination of an issue, soliciting expert advice is not new. Generally, however, this has been pursued on an informal basis, and such an unstructured approach is rarely, if ever, found entirely satisfying to all parties. Neither is it likely to be immune to legitimate criticism, from one side or another. To counteract these shortcomings, a s*tructured* expert judgment elicitation refers to the deliberate effort to subject the whole process to transparent methodological rules, with the goal of treating expert judgments as scientific data in a formal decision process.

Various methods for assessing and combining expert uncertainty are available in the literature. Some advocate a group decision-conferencing consensus approach for eliciting opinions, for instance, but other approaches exist for carrying out this process, and part of the motivation for the present pilot is to trial the EXCALIBUR structured expert judgment procedure, formulated by Cooke (1991) as the Classical Model, has been selected for scrutiny in application. The theoretical basis and principles of the Classical Model are described in Part 2 of these notes.

1.2 EXCALIBUR Procedure

The main steps in the procedure for applying the EXCALIBUR approach in practice can be summarised as follows:

- A group of experts are selected.
- Experts are elicited individually regarding their uncertainty over the results of possible measurements or observations within their domain of expertise.

- Experts also assess variables within their field, the true values of which are known or become known post hoc.

- Experts are treated as statistical hypotheses and are scored with regard to statistical likelihood (often called 'calibration') and informativeness.

- Scores are combined to form weights. These weights are constructed to be 'strictly proper scoring rules' in an appropriate asymptotic sense: experts receive their maximal expected long-run weight by, and only by, stating their true degrees of belief. With these weights, statistical accuracy strongly dominates informativeness – one cannot compensate poor statistical performance by very high information.

- Likelihood and informativeness scores are used to derive performance- based weighted combinations of the experts' uncertainty distributions.

The key feature of this method is the performance-based combination of expert uncertainty distributions. When it comes to attempting to resolve differences in expert judgments, people who seek to find a harmony of views by conciliation can be disconcerted by this approach, but extensive experience overwhelmingly confirms that experts grow to favour it because its performance measure are entirely objective and amenable to diagnostic examination.

1.3 Combining expert assessments to form a Decision Maker

A combination of expert assessments is often referred to as a "decision maker" (DM), in the sense of linear pooling. The steps in the process by which one can arrive at a decision maker outcome are summarised and illustrated schematically in Fig. 3. On the left hand side of this diagram, hypothetical examples of the responses of three different experts to three seed questions are depicted, showing how their calibration can vary, in relation to the true realization value for the seed item, and how their information can also vary, generally from expert-to-expert, rather than within experts. Note that each expert is required to provide a fixed number of quantiles (usually three) to express his or her degree of belief in their judgment of the seed item value and the credible interval within which it should fall in their opinion.

With a set of several seed items (usually about ten in number), a group experts can be ranked according to their individual calibration and information scores, and then according to the weights overall, as determined by the product of calibration and information scores. With these latter weights to hand, it is then possible to elicit from the same group of experts their quantile-based distributions for items of interest (i.e. for questions for which an expert consensus is sought), and these individual response distributions can be linearly pooled, applying the individual weights. It should be noted that a weighted combination distribution, obtained in this way, is seldom if ever identical to the distribution of any one contributing expert, but does represent a rational consensus of the information provided by members of the group as a whole, differentiated by their performance on the seed items.

The Classical Model is essentially a formal method for deriving the requisite weights for a linear pool in which, as just noted, these weights are expressed as the product of an individual's calibration and information scores. "Good expertise" corresponds to good calibration (high statistical likelihood the expert's distributions reflect true values) and

superior information. Strong weights reward good expertise, and pass these virtues on to the decision maker.

The reward aspect of weights is very important. An expert's influence on the decision maker should not appear haphazard, and he/she should be discouraged from attempting to game the system by tilting his/her assessments to achieve a desired outcome. Thus it is necessary to impose a strictly proper scoring rule constraint on the weighing scheme. Roughly speaking, this means that an expert achieves his maximal expected weight by, and only by, stating assessments in conformity with his/her true beliefs.

2 The Classical Model for Expert Judgment Elicitation

2.1 General remarks

The process by which experts come to agreement sensu stricto in science is the scientific method itself. With expert judgments regarded as scientific data, a structured expert elicitation formalism cannot pre-empt the scientific method, and therefore cannot have enforced agreement as a valid scientific goal.

Following, loosely, Cooke and Goossens (2008), there are three broadly different goals to which a structured judgment method may aspire in a decision-support role:

- To arrive at an administrative or political consensus (compromise) on scientific issues
- To provide a census of scientists' views
- To develop a rational evidence-based consensus on the particulars of the science of interest

Political consensus refers to a process in which experts are assigned weights according to the interests or stakeholders they represent. In practice, an equal number of experts from different stakeholder groups would be placed in an expert panel and given equal weight in the panel. In this way, the different groups are included equally in the resulting representation of uncertainty. This was the reasoning behind the selection of expert panels in the EU USNRC accident consequence studies with equal weighting (Goossens and Harper, 1998). In essence, the concept can be summed up as akin to "one man, one vote".

In contrast, a study aiming at furnishing a *scientific census* will try to survey the totality of views across an expert community, and express this as a distribution. The objective is to include extreme views and acute outliers, but at the same time seeking a proscription to limit their influence in some way. An illustration of an implementation of this type is found in the US Nuclear Regulatory Commission *Recommendations for Probabilistic Seismic Hazard Analysis: Guidance on Uncertainty and Use of Experts* (NUREG/CR-6372, p.36):

> "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"

The goal of representing *"the overall community*" may, in this view of the world of science, invoke differential weighting of experts' views, according to some judgment as to how representative they are thought to be of other experts. The philosophical underpinnings of the approach are elaborated in Budnitz et al. (1995; 1998); see also Winkler et al. (1995). However, the objectivity of the process for ascertaining the appropriate weights to assign to experts under such a scheme is open to challenge. Furthermore, the inadequacies of this approach in application have been roundly demonstrated recently in a major seismic hazard assessment for a nuclear power station in Switzerland; the attempt there to acquire a community consensus contributed to implausibly high hazard levels from the study, widespread criticism, multiple reviews and workshops, and a substantial discussion in the seismological literature (with too many references to cite here).

Expert agreement on the representation of the overall scientific community is the weakest, and most accessible, type of scientific consensus to which a study may aspire. Other types

of censual approach, in decreasing accessibility, are: agreement on a 'distribution to represent a group', agreement on a distribution, and agreement on a number.

Rational consensus refers to a group decision process, as opposed to a group census or consensus procedure. The group agrees on a method according to which a representation of uncertainty will be generated for the purposes for which the panel was convened, without knowing the result of this method. It is not required that each individual member adopt this result as his personal degree of belief. This is another form of "agreement on a distribution to represent a group". To be rational, this method must comply with necessary generic conditions devolving from the scientific method. Cooke (1991) formulates the necessary conditions or principles, which any method warranting the designation "scientific" should satisfy, as:

- **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.

Thus, a method is needed which satisfies these conditions and to which the parties commit, beforehand. Then, the method is applied and after the results of its application are obtained, parties wishing to withdraw from the consensus incur a burden of proof. They must demonstrate that some, hitherto unmentioned, necessary condition for rational consensus has been violated - without that, their dissent cannot be "rational". Of course, any party may withdraw from the consensus because the result is hostile to his interests – this is not rational dissent and does not negate rational consensus.

The second requirement, of empirical control, could strike some as peculiar in this context. How can there be empirical control with regard to so-called subjective probabilities? To answer this, the question to consider is: when is a problem an expert judgment problem? For instance, it would be bizarre to seek recourse to expert judgment to determine the speed of light in a vacuum, as this is physically measurable and has been determined sufficiently precisely, to everyone's satisfaction. Any expert queried on the speed of light would give the same answer as any other.

A scientific problem is amenable to expert judgment only if there is relevant scientific expertise. This entails that there are theories and measurements relevant to the issues at hand, but that the specific quantities of interest themselves cannot be measured in practice or, if they can, not within the timescale for a decision to be made. For example, toxicity of a substance for humans is measurable in principle, but is not measured for obvious ethical reasons. There are, however, toxicity measurements for other species that might be relevant to the question of toxicity in humans. If a problem is an expert judgment problem, then necessarily there will exist somewhere relevant experiments, observations or measurements.

Questions regarding such experiments can be used to implement empirical control. In a performance-based expert pooling scheme, these are usually referred to as "*seed*" questions. These need to be subject-matter specific: research indicates that performance on so-called almanac or general knowledge questions does not predict performance on variables in an

expert's field of expertise (Cooke et al., 1988). The key question regarding seed variables is this: is performance on seed variables judged relevant for performance on the variables of interest? For example, should an expert who gave very over-confident off-mark assessments on the variables for which the true values are known be allowed to be equally influential on the variables of interest as an expert who gave highly informative and statistically accurate assessments? This is a choice that often confronts a problem owner - after the results of an expert judgment study are in. If seed variables in this sense cannot be found, then rational consensus is not a feasible goal and the analyst should fall back on one of the other goals.

The above definition of "*rational consensus*" for group decision processes is evidently on a very high level of generality. Much work has gone into translating this into a workable procedure that gives good results in practice (Cooke and Goossens, 2008). This workable procedure is embodied in the "Classical Model" of Cooke (1991), described in subsequent paragraphs, and implemented as the EXCALIBUR software package (formerly EXCALIBR: Cooke and Solomatine, 1992).

Before going into detail, it is appropriate to say something about Bayesian approaches. Since expert uncertainty concerns experts' subjective probabilities, many people believe that expert judgment should be approached from the standpoint of the Bayesian paradigm - a model that is based on the representation of the preferences of a rational individual in terms of maximal expected utility. If a Bayesian is given experts' assessments on variables of interest and on relevant seed variables, then he may update his prior on the variables of interest by conditionalizing on the given information. This requires that the Bayesian formulates his joint distribution over:

- the variables of interest
- the seed variables
- the experts' distributions over the seed variables and
- the variables of interest.

Issues that arise in building such a model are discussed in Cooke (1991). Suffice it to say here that a group of rational individuals is not itself a rational individual, and group decision problems are notoriously resistant to a Bayesian treatment.

Here, it is assumed that uncertainty is represented as subjective probability and that the concern is with the results of possible – if inaccessible – observations (for further discussion of foundational issues, the reader is referred to Cooke, 2004). When expert opinion is expressible in a quantitative form it can be considered to be data, in just the same way as is empirical data (both represent an expression of belief about a particular variable value, and both should incorporate a statement of the associated uncertainty). In other words, expert opinion has essential characteristics in common with empirical data from experiments or physical observations: the elicitation method involves empirical control, but adduces what is sometimes referred to as "*subjective data*". This designation can be misleading: if the experts involved are truly expert, then their opinions must be objective to some degree, as are their assessments of uncertainty.

If the concept of subjective data is accepted, the question then is: how to combine a range of expert opinions in some optimal way? While the advantages and limitations of different expert weighting schemes are subjects of on-going active research (see Cooke, 2008), one particular formulation, the Classical Model (Cooke, 1991), has the necessary basis of proper

scoring rule implementation and the attribute of empirical control for deriving a rational consensus when eliciting the views of uncertain experts.

2.2 The Classical Model

The principles outlined above have been implemented for expert elicitation in the so-called "Classical Model", a performance based linear pooling or weighted averaging model (Cooke 1991). The weights are derived from experts' calibration and information scores, as measured on seed variables. Seed variables serve a threefold purpose:

- to quantify experts' performance as subjective probability assessors,
- to enable performance-optimized combinations of expert distributions, and
- to evaluate and hopefully validate the combination of expert judgments.

The name "Classical Model" derives from an analogy between expert calibration measurement and classical statistical hypothesis testing.

In the Classical Model, performance-based weights use two quantitative measures of competency: *calibration* and *information.* Loosely, *calibration* measures the statistical likelihood that a set of experimental results correspond, in a statistical sense, with the expert's assessments. *Information* measures the degree to which an expert's uncertainty distribution is concentrated.

These measures can be implemented for both discrete and 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 presented an uncertain quantity taking values in a continuous range, and they give pre-defined quantiles, or percentiles, of the subjective uncertainty distribution, typically 5%, 50% and 95%.

The quantile format has distinct advantages over the discrete format.

2.2.1 Calibration

For each quantity, each expert divides his belief range into four inter-quantile intervals for which the corresponding probabilities of concurrence are known, namely $p_1 = 0.05$ for a realization value less than or equal to the 5% value, $p_2 = 0.45$: realization value is greater than the 5% value and less than or equal to the 50% value, $p_3 = 0.45$,…and so on. (Other quantiles and interquantile ranges can be used in practice).

If N such quantities are assessed, each expert may be regarded as a statistical hypothesis, namely that each of the N realization values falls in one of the four inter-quantile intervals with probability vector:

$$p = \{0.05, 0.45, 0.45, 0.05\}$$

The sample distribution over the expert's inter-quantile intervals can then be formed by summing the number of realizations which fall in each interval, divided by total number N (see Fig. 1). Note that the sample distribution depends on the expert e.

If the realizations are indeed drawn independently from a distribution with three quantiles, as stated by the expert, then the quantity:

$$2N \cdot I(s(e) \mid p) = 2N \cdot \sum_{i=1..4} \{ s_i \cdot ln(s_i / p_i)\} \qquad [2.1]$$

is asymptotically distributed as a chi-squared variable with 3 degrees of freedom. This is the so-called likelihood ratio statistic, in which $I(s(e) \mid p)$ is the relative information or relative entropy (see e.g. Cover and Thomas, 1991) of distribution s with respect to p for expert e, and relative information is defined as follows. Let a discrete distribution have probability function s, and let a second discrete distribution have probability function p. Then the relative information of p with respect to s is: $s \cdot ln(s / p)$, which is also called the "Kullback information entropy" or the "Kullback-Leibler distance". If the leading term of the logarithm in equation [2.1] is extracted, the familiar chi-squared test statistic for goodness of fit is obtained; there are advantages in using this form (Cooke 1991).

In the Classical Model, the decision maker scores expert e as the statistical likelihood of the hypothesis:

H_e: "the inter-quantile interval containing the true value for each variable is drawn independently from probability vector p."

A simple test for this hypothesis uses the information likelihood ratio statistic from equation [2.1], and the likelihood, or probability value, of this hypothesis, to form the calibration score:

$$Calibration\ score(e) = Prob\ \{2N \cdot I(s(e) \mid p) \geq r \mid H_e\} \qquad [2.2]$$

where $Prob\{ \mid \}$ denotes the probability that information likelihood ratio is greater than or equal to r, *given* the hypothesis is true, where r is the relevant quantity value from the expert's sample distribution, outlined above.

Thus, the *Calibration score* is the probability under hypothesis H_e that a deviation at least as great as r could be observed on N realizations, if H_e were true. Although the calibration score uses the language of simple hypothesis testing, it must be emphasized that it is not used to reject expert hypotheses; rather, the terminology is used to measure the degree to which the data supports the hypothesis that the expert's probabilities are accurate. Low scores, near zero, mean that it is unlikely that the expert's probabilities are correct.

2.2.2 Information

The second scoring variable used in the Classical Model is *information* (alternatively, *entropy*). 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 some background measure. Being concentrated or "spread out" is measured relative to some other distribution. Commonly, the uniform and log-uniform background measures are used.

Measuring information requires associating a probability density with each quantile assessment of each expert. To do this, a unique density distribution is adopted that complies with the experts' quantiles and is minimally informative with respect to the background measure (a "minimally informative" distribution in this context means that distribution, out of all possible distributions, which matches the given quantiles but has least information or

deviation from the background distribution at other points between the elicited quantiles; sometimes referred to as a "vague" distribution, when used as a prior).

For a uniform background measure, the probability density is constant between the assessed quantiles, and is such that the total mass between the quantiles agrees with the probability vector p (identified above). The background measure is not elicited from the experts as it must be the same for all experts; instead it is chosen by the analyst.

Both the uniform and log-uniform background measures require an *intrinsic range* on which these measures are concentrated.

For this, the Classical Model implements the so-called "k% overshoot rule": for each item. First the smallest interval $I = [q_5 , q_{95}]$ is determined that contains all the assessed quantiles of all experts (i.e. where the lowest valued 5%ile quantile of all is q_5 , and the highest valued 95%ile is q_{95}) and also contains the realization for that item, if known. This interval is extended to a new, wider interval:

$$I^* = [q_L , q_H] \qquad [2.3]$$

where:

$$q_L = q_5 - k \cdot (q_{95} - q_5)/100$$

$$q_H = q_{95} + k \cdot (q_{95} - q_5)/100$$

See Fig. 2 for a diagrammatic representation of the intrinsic range.

The value of k, which determines the amount by which the range is extended, is chosen by the analyst. A large value of k tends to make all experts look more informative, and tends to suppress the relative differences in experts' information scores. Typically, $k = 10$ is chosen to produce a 10% overshoot.

With the intrinsic range so defined, the information score of expert e on assessments for N uncertain quantities is:

Information Score(e) =Average Relative information wrt Background

$$= (1/N) \cdot \sum_{i=1..N} \{I(f_{e,i} \mid g_i)\} \qquad [2.4]$$

where g_i is the background probability density for variable I over the extended intrinsic range, and $f_{e,i}$ is expert e's probability or variable density function for item i. The relative informations for all variables are summed and normalized over the N quantities involved. This normalized sum is proportional to the relative information of the expert's joint distribution with respect to the background distribution, under the assumption that the variables are independent.

Although the above choice for interpolating the experts' quantiles is, to some extent, arbitrary, it generally makes relatively little difference to the weights given to the experts. This is because the calibration scores, which usually drive the weighting, depend only on the quantiles, and not on the interpolation. The information score depends only on the quantiles and the choice of the intrinsic range to use. See Fig. 2 for an example of a single expert's

interpolated distribution across three quantile values, with the extended intrinsic range added at each end.

This way of interpolating affects the estimate of the combined Decision Maker distribution, however, and hence influences the selection of cut-off level for the weights. As with calibration, the assumption of independence here reflects a desideratum of the decision maker, and not an elicited feature of the expert's joint distribution. The information score does not depend on the seed item realization values. An expert can give himself a high information score by choosing his quantiles to lie very close together, but then his calibration score may suffer.

Evidently, the information score of expert *e* depends both on the intrinsic range chosen by the analyst and on the assessments of the other experts. Hence, information scores cannot be compared across studies. This particular information score is chosen for the Classical Model because it is:

- familiar
- tail insensitive
- scale invariant
- "slow"

The latter property means large changes in an expert's assessments produce only modest changes in his information score. This contrasts with the likelihood function in the expert's calibration score, which is a "fast" function. Taken together, the product of calibration and information is driven mainly by the faster function, i.e. the calibration score.

2.2.3 Pooling expert assessments to form a Decision Maker

A combination of expert assessments is often referred to as a "decision maker" (DM), in the sense of linear pooling. The steps in the process by which one can arrive at a decision maker outcome are summarised and illustrated schematically in Fig. 3. On the left hand side of this chart, hypothetical examples of the responses of three different experts to three separate seed questions are depicted, showing how their calibration can vary in relation to the true realization value for the seed item, and how their information can also vary, generally from expert-to-expert, rather than within experts. Note that each expert is required to provide a fixed number of quantiles (usually three) to express his degree of belief in his judgment of the seed item value and the credible interval within which it should fall, in his opinion.

With a set of several seed items (usually about ten in number), a group of experts can be individually ranked according to their individual calibration and information scores, and then according to the weights overall, as determined by the product of calibration and information scores. With these latter weights to hand, it is then possible to elicit from the same group of experts their quantile-based distributions for items of interest (i.e. for questions for which an expert consensus is sought), and these individual response distributions can be linearly pooled, by applying the individual weights. It should be noted that a weighted combination distribution, obtained in this way, is seldom if ever identical to the distribution of any one contributing expert, but does represent a rational consensus of the information provided by members of the group as a whole, differentiated by their performance on the seed items.

The Classical Model is a formal method for deriving the requisite weights for a linear pooling in which, as described above, these weights are expressed as the product of an individual's

calibration and information scores. "Good expertise" corresponds to good calibration (i.e. high statistical likelihood of acceptance of the hypothesis) and superior information. Strong weights reward good expertise, and pass these virtues on to the pooled outcome, the so-called decision maker.

The reward aspect of such weights is very important. An expert's influence on the decision maker should not appear haphazard, and the weighting scheme should be such that he is discouraged from attempting to game the system by tilting his assessments with the intention of achieving a desired outcome. Thus it is necessary to impose a strictly proper scoring rule constraint on the weighing scheme. Roughly speaking, this means that an expert achieves his maximal expected weight by, and only by, stating assessments in conformity with his/her true beliefs.

Consider the following scoring weight for expert *e:*

$$w_\alpha(e) = Ind_\alpha(\text{calibration score}(e)) \times \text{calibration score}(e) \times \text{information score}(e) \qquad [2.5]$$

where $Ind_\alpha()$ denotes an indicator function with $Ind_\alpha(x) = 0$ if $x < \alpha$ and $Ind_\alpha(x) = 1$ otherwise.

In this case, $Ind_\alpha()$ is based on the expert's calibration score, and only allows expert *e* to gain a non-zero weight $w_\alpha(e)$ if his score exceeds a threshold level defined by some value, α. Cooke (1991) shows that the expert's score $w_\alpha(e)$ is an asymptotically strictly proper scoring rule for average probabilities. The scoring rule constraint requires the term $Ind_\alpha(\text{calibration score}(e))$ to be applied to the expert's score, but does not say what value of α should be. Therefore, α can be chosen so as to maximize the combined score of the resulting decision maker when all the experts' distributions are pooled together.

Let $DM_\alpha(i)$ be the result of linear pooling for item *i* for all experts, with the total number of experts *E*, and with their individual weights proportional to $w_\alpha(e)$, as per equation [2.5]. Thus, summing over all *E*, and normalizing for the sum of individual weights:

$$DM_\alpha(i) = \sum_E \{w_\alpha(e) \cdot f_{e,i}\} \; / \sum_E \{w_\alpha(e)\} \qquad [2.6]$$

where $f_{e,i}$ is expert *e*'s probability or variable density function for item *i*

Next, define the "*global weight DM" a*s DM_{α^*} where α* maximizes the product:

$$\text{calibration score}(DM_\alpha) \times \text{information score}(DM_\alpha). \qquad [2.7]$$

This maximal weight is termed "global" because the information score is based on all the assessed seed items, not just the seed items.

Over the long run, an expert maximizes his expected weight by stating his true opinion. The conditions require that a minimum significance level α* be maintained, such that if the expert's calibration score falls beneath α*, he receives no weight. The requirement of being 'strictly proper' largely determines the form of the calibration term in the expert score, whereas the entropy term serves to represent information (or lack of it).

The significance level α* can be chosen to optimize the decision maker's distribution in the following sense. For a given significance level, the experts' weights and hence the decision maker's distributions for each variable are determined. Extracting the 5%, 50% and 95%

quantiles from this pooled distribution, the decision maker can then be treated as a 'virtual expert', and scored on the seed variables. Hence, for any significance level, calibration and entropy scores for the decision maker can be derived, as well as the 'virtual weight' that the decision maker would receive if he were scored along with the real experts.

Thus, the calibration and the information of *any* proposed decision maker can be computed with the expectation that the "optimal decision maker" should perform better than the result of simple averaging (i.e the *equal weights decision-maker DM*). Also, it would be hoped that the optimal DM is not significantly worse than the best expert in the panel.

In actual applications, decision maker optimization is achieved typically at a hypothesis rejection significance level of about 5%. In practice, some members of a group of experts are likely to receive negligible or even zero weight at significance levels of this order; however, the decision maker is then generally – but not invariably - substantially 'heavier' than all the remaining real experts, as is desirable. Reducing the significance level to lower and lower values enables all experts to receive some positive weight but, inevitably, this substantially degrades the decision maker's own calibration and entropy scores.

2.2.4 Calculation of the Decision Maker distribution

With the calibration and information scores determined for each expert, as described above, all the elements needed to determine the output distribution for a given query variable are now assembled. Given this set of weights, target variable quantiles for each query variable can be computed for the DM (usually 5%ile, 50%ile and 95%ile, if these are the calibration quantiles used). When the resulting weights for each expert at the selected significance cut-off level have been ascertained, the pooled distribution function is now simply the sum of the products of each expert's weight with his item distribution function. Fig. 5 illustrates the application of this procedure for two experts' distributions combined with unequal weights.

2.3 Variations on the theme in application

The EXCALIBUR program has been developed to offer various options for problem analysis and additional facilities for the analyst to understand the data. Options that may be used in practice, but not all the alternatives available, are summarised in this section.

3.1 Uniform and logarithmic scaling
In the implementation of the Classical Model, the experts are asked for a limited number of quantiles – typically 5%, 50%, and 95% quantiles for each target (and seed) variable, although more quantiles can be used if circumstances allow or call for it.

The analyst has to make a choice of scale for each query variable (logarithmic or uniform). As a rule-of-thumb, logarithmic scaling would be chosen when the range of credible values for the item or variable being considered spans over three orders of magnitude, or more – less than this and the uniform scale can cope quite adequately. If logarithmic scaling is chosen then the expert's corresponding quantile values are converted to logs and the background distribution for information scoring is taken to be log-uniform, before applying the same scoring analysis procedure as for uniform query variables.

The rest of the calculational procedure is explained here on the basis of uniform scaling, but the same principles apply to log scaling.

2.3.2 Alternative weighting schemes

The possibilities for scoring weights do not end with the global weights system, however. A variation on this scheme allows a different set of weights to be used for each different target item of interest. This is accomplished by using expert information scores for each item, rather than the average information score over all items and, when applied, is denoted by the sobriquet *item weights*.

The EXCALIBUR program provides the analyst with a facility to compute a decision maker based on giving all experts *equal weights*, mainly to allow comparison of the optimal decision maker with the results that would be obtained by simple averaging of expert views. Even greater degradation of the DM's calibration and entropy score results from assigning all experts equal weights. Such an uncritical combination of expert assessments generally results in inordinately large confidence bounds (credible intervals) in the pooled outputs. Thus, a primary virtue of the Classical Model is its power to reduce the 'noise' of divergent expert opinions, generally improving calibration synthesis at the same time. An example of a target item range graph (Fig. 4) illustrates typical quantile judgments from a group of experts for one variable, and the corresponding pooled decision maker results when optimal and equal weights are used.

In addition, EXCALIBUR has a facility for importing *user weights* from an external source. It may be that optimal decision maker weights have already been computed separately and it is desired to apply them to a new set of target questions. Or, user weights may be derived in some other way, for instance by mutual self-weighting, in which members of a group ascribe weights to each other member of the group and these self-inflicted weights are combined numerically in some way.

2.3.3 Item weights and expert learning

Taking the discussion of the method's strengths one step further, item weights are potentially more attractive than global weights as they allow an expert to up- or down-weight his responses for individual items according to how much he feels he knows about that item in particular. "Knowing less" implies choosing quantiles that are spread further apart and consequently lowering the information score for that item. Of course, good performance of item weights requires that experts can perform this differential judging successfully. Anecdotal evidence suggests that item weights analyses improve over the global weights counterpart as the experts receive more training in probabilistic assessment. Both item and global weights can be briefly described as optimal weights under a strictly proper scoring rule constraint. In both cases, calibration dominates over information, and information serves to modulate between more or less equally well-calibrated experts.

In some circumstances, a staged or iterative approach may be taken to the elicitation of expert opinion. If, after a few questions, an expert were to see that all seed question realizations fell outside his 90% credible interval bounds, he might conclude that these intervals were too narrow and might broaden them on subsequent assessments. This means that for this expert the uncertainty distributions are not independent, and he learns from the realizations. However, *expert learning* is not a goal of an expert judgment study and his joint distribution would not be retained. Rather, the decision maker wants experts who do not need to learn during a formal elicitation – such training should take place elsewhere.

2.3.4 Constrained optimization and selective weighting

As noted above, the significance level threshold value for un-weighting experts is determined either by optimizing numerically the calibration and information performance of the DM.

Alternatively, this threshold can be fixed by the analyst on the basis of some constraining criterion, influenced by other considerations. In a dam safety study in Britain (Brown and Aspinall, 2004), for instance, in addition to the optimised DM, other variant EXCALIBUR results were obtained by fixing the calibration power and significance level parameters of the hypothesis test so as to (a) ensure that all experts obtain some positive, non-zero weight, and (b) that the ratio between the highest and lowest weights was not too extreme. After discussion with the owners of this particular survey, the span between the best and poorest performances was fixed, pragmatically, to be no more than two orders of magnitude (i.e. the highest individual weighting being no more than a factor of 100 times that of the lowest). This approach, in which the weights of individuals are factored before pooling the responses from the whole group, quite strongly moderates the DM's performance, and hence curtails the potential for determining the optimal outcome in a decision theoretic sense
.
Thus, in the dam study case, additional analyses were conducted for the purpose of enhancing the DM's performance in some pragmatic way – without actually maximizing it absolutely – such that the severity with which low-weighted real experts were rejected was limited. This was achieved by tuning both the statistical power of the hypothesis test (effectively, by reducing the granularity of differentiation provided by the set of seed items) and the related significance level setting, which together determine the confidence level for hypothesis rejection upon which the calibration score is based. There is a wide range of possible combinations of settings for these two model parameters and, in the case of the dam study, it was decided that, whatever selections were made, a majority of the group (i.e. for no less than six of the eleven experts) must retain non-zero weights. Supplementary analysis runs were undertaken, therefore, to examine how the elicitation results might change if this position was adopted. The calibration power and significance level were each increased incrementally to allow the analysis to give more and more weight to the DM, until the minimum size of a majority quorum, mentioned above, was reached.

The results produced by this "artificial" pooling configuration were not dramatically different from those obtained with full optimization, although there are notable changes in the results for a few items, and hints of systematic shifts in the central value outcomes in several others. This said, the observation that differences in outcomes were generally modest is not surprising, however, if it is pointed out that each of the experts discounted in this way had low individual performance scores, and were not exerting much influence on the joint pooling, anyway. What is significant, however, is that, as a result, greater authority is given to the DM than would have been the case in a situation where all experts were allowed non-zero scores or given equal weights.

This *selective weighting* approach represents a shift towards a more homogeneous collective combination of the views of the most influential experts, and a position where the DM can then out-score most, if not all, of the individual experts. On this basis, it could be argued that results obtained under this *constrained optimization* scheme represent a more robust, and more rational, union of opinions than would be provided by making sure the views of the whole group were utilized with equal weights but, it should be remembered, they remain sub-optimal and hence less desirable from a decision theoretic perspective.

2.3.5 Discrepancy analysis

In an EXCALIBUR discrepancy analysis, the relative information of each expert's assessment, per item, is compared with the assessment of the DM (pooled decision-maker) for that item, and the relative information of the expert with respect to the DM is computed. These measures are averaged over all items, and are proportional to the relative information of the respective joint distributions if all items independent.

These numbers, which can be provided as output by the program, are greater or equal to zero, and get larger as the expert's assessment differs more and more from the GM's assessment for the given item. This enables the facilitator to see which experts agree or disagree most with the decision maker (agreement, or disagreement, is not well predicted by an expert's un-normalized weight).

2.3.6 Robustness tests

The question may be asked, how stable is the Classical Model decision maker outcome to the seed items used or the experts consulted? The EXCALIBUR program provides facilities for exploring these effects, under the control of the analyst.

To perform a robustness analysis on seed items used for calibration, new DMs are computed in EXCALIBUR by successively deleting one seed item at a time, and scoring the DM with the remaining seed items. The total relative information with respect to the background measure, the calibration and total relative information with respect the original (in this case, the optimised global weights) DM are tallied to explore which, if any, of the seed items exerts a strong influence on the results. If undue influence by one or more seed items is detected, the analyst and problem ownwer may wish to consider re-balancing the set of seed items by finding alternative questions that are more representative of the problem.

A similar process is followed for expert robustness testing: individual experts are removed from the computation of the DM, one at a time, in order to check which, if any, have a significant influence on the properties of the optimal DM. Of course, a single well-calibrated very informative expert in a group of several average performers will show up well in such a robustness test, which is right and proper, but if someone appears to score well beyond their apparent competency, then the analyst might wish to examine how they achieved such prominence. Such a situation is extremely rare in practice, however, and would be very unlikely to arise in a properly explained and well-managed group elicitation, when the temptation to attempt to game the procedure is rationally discouraged.

Thus, in optimizing the DM, the aim is not to secure robustness but to achieve genuine high performance against a proper scoring rule. Checking robustness is worthwhile for building confidence in the outcome, but it is unlikely that a facilitator – or problem owner - would opt for a lower performance DM simply because it appeared more robust.

As a rule of thumb, if the removal of any single seed item or loss of a single expert doesn't perturb the derived DM by more than mutual differences between experts, then the DM is responding to genuine variations in expert opinion and robustness is not a worry.

2.4 Summing up

In the Classical Model, calibration and information are combined to yield an overall or combined score with the following attributes:

1. Individual expert assessments, realizations and scores can be recorded. This enables any reviewer to check the application of the method, in compliance with the principle of **accountability / scrutability.**
2. Performance is measured and hopefully validated, in compliance with the principle of **empirical control.** An expert's weight is determined by performance on seed items.
3. The score is a long run proper scoring rule for average probabilities, in compliance with the principle of **neutrality.**
4. Experts are treated equally, prior to the performance measurement, in compliance with the principle of **fairness.**

Whilst expert names and qualifications should be part of the documentation of every expert judgment study, they are not usually associated directly with identifiable individual assessments in the open literature. The experts' reasoning is always recorded and that is sometimes published as expert rationales.

There is no mathematical theorem which states that either item weights or global weights will out-perform equal weights or out-perform the best expert. Indeed, it is not difficult to construct artificial examples where this is not the case. Selecting which of these weighting schemes to use is a matter of experience. In practice, global weights are used unless item weights perform markedly better.

Of course, there may be other ways of defining expert weights that perform better, and indeed there might be better performance measures. But, good performance on a one-off basis for a single individual data set is not convincing. What is convincing is good performance on a large diverse data set, such as the TU Delft expert judgment database (Cooke and Goossens, 2008). In practice a method should be easy to apply, easy to explain, should do better than equal weighting and should never do something ridiculous.

Forty-five different expert elicitations involving seed variables have been performed to date (Cooke and Goossens, 2008). These are all studies performed under contract for a problem owner, and reviewed and accepted by the contracting party. In most cases the results have been published. Given the body of experience with structured expert judgment that has now accumulated, the performance-based Classical Model approach is well established: as mentioned earlier, simply using equal weights for scientific uncertainty quantification no longer seems to be a convincing alternative.

This experience also shows that in the great majority of cases, the performance-based combination of expert judgment gives more informative and statistically more accurate results than either the best expert or the 'equal weight' combination of expert distributions (Cooke, 2004; Cooke and Goossens, 2000; Goossens et al., 1998). Upon reflection, it is evident that equal weighting has a very serious drawback. As the number of experts increases, the equal weight combination typically becomes increasingly diffuse, until it represents no one's belief and is useless for decision support. This is frequently seen as the number of experts exceeds, say, eight. The viability of equal weighting is maintained only by sharply restricting the number of experts who will be treated equally, leaving others outside

the process. It appeals to a sort of one-man-one-vote consensus ideal. Progress in science, however, is driven by rational consensus.

Ultimately, consensus is an equilibration of power; in science, it is not the power of the ballot but the power of arguments that counts (Kurowicka and Cooke, 2006), and this can be made manifest through the EXCALIBUR structured elicitation procedure.

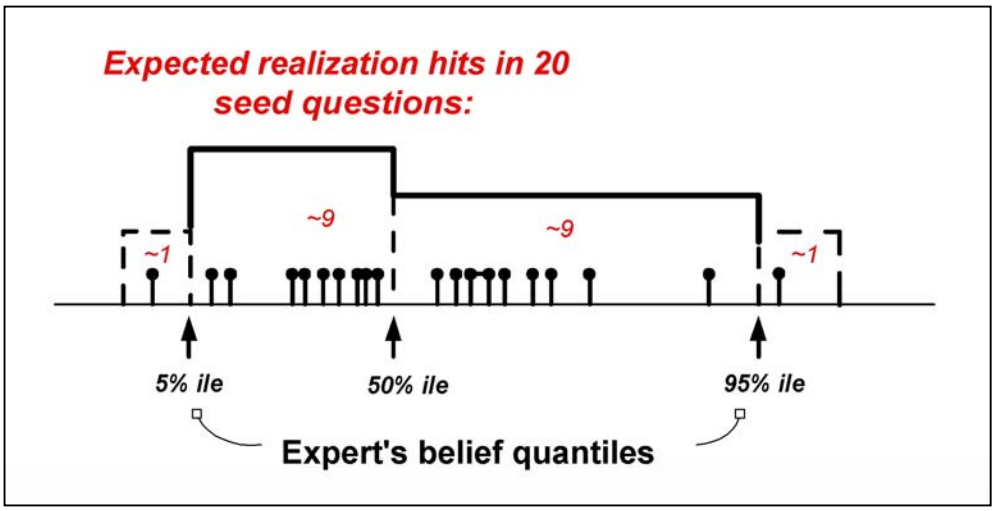

Fig. 1 Schematic depiction of seed item realizations in relation to the inter-quantile ranges of a well-calibrated expert: the realization values should be distributed within the inter-quantile ranges in close agreement to the proportions {0.05, 0.45, 0.45, 0.05}.

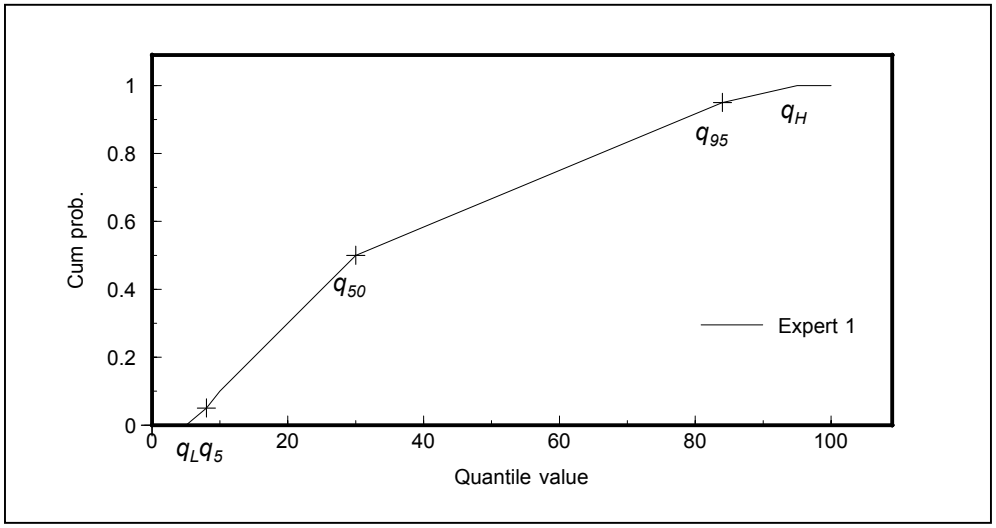

Fig. 2 Simple representation of an interpolated distribution of quantiles for one expert. With suitable overshoot adjustment, q_L , q_H define the *intrinsic range* (from the range of extreme quantile values provided by all experts by – see text). The distribution of Expert 1 is then approximated by linear interpolation over the quantile information (q_L , 0), (q_5 , 0.05), (q_{50} , 0.5), (q_{95} , 0.95), and (q_H , 1). This is the distribution with minimum information with respect to the uniform distribution on the intrinsic range, which satisfies this expert's quantiles.

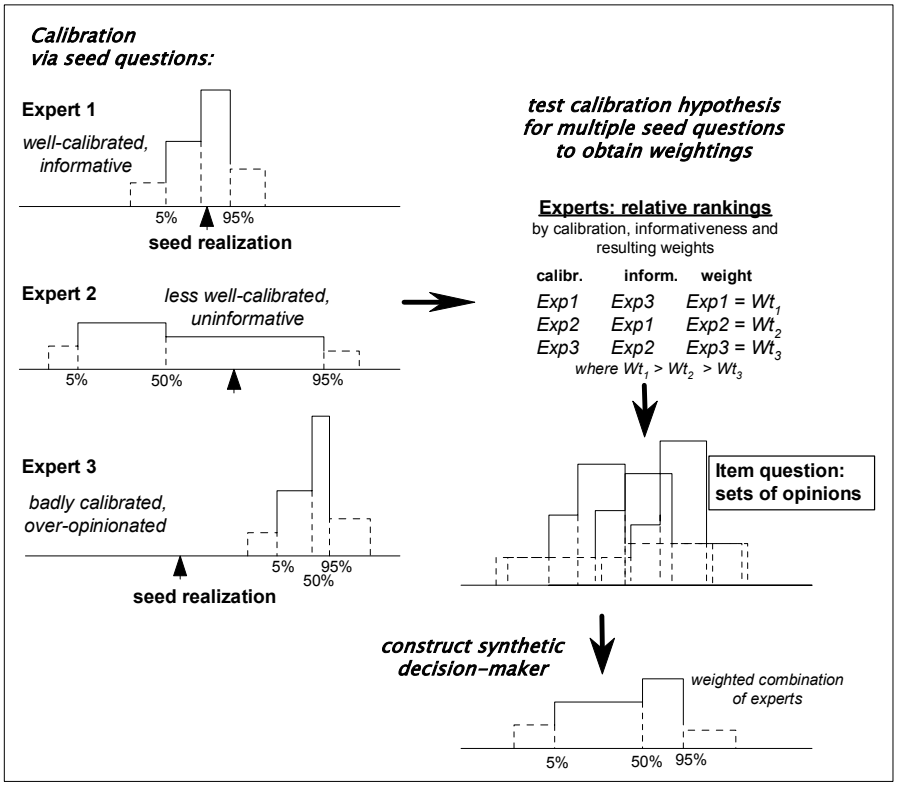

Fig. 3 Schematic chart showing how experts responses are calibrated against (multiple) seed questions at given quantiles to produce performance-based weights, which are then used to pool the experts' opinions for the corresponding quantiles of target items.

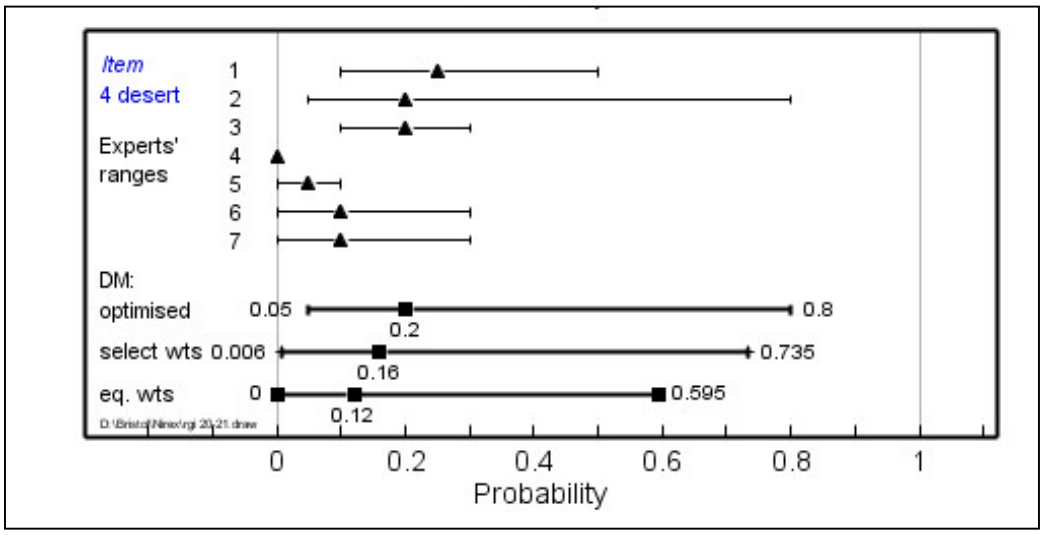

Fig. 4 Example of a range graph for a target item (in this case, a scenario probability relating to desert conditions) from an elicitation of a group of seven experts, showing the variability of individual opinions (bars 1 – 7). The weighted, pooled outcomes are shown as global (optimised)-, power-constrained (select wts)- and equal weights (eq. wts) decision-maker DMs in the three lower bars, illustrating the influence of different performance-based measure schemes on the pooling of opinions; the global (optimised) DM is normally the preferred solution, the others serving to indicate sensitivity to alternative assumptions about pooling strategies. For each row, the median estimate is marked by a symbol, and the 90% credible interval by the bars (note these need not be symmetrical about the median).

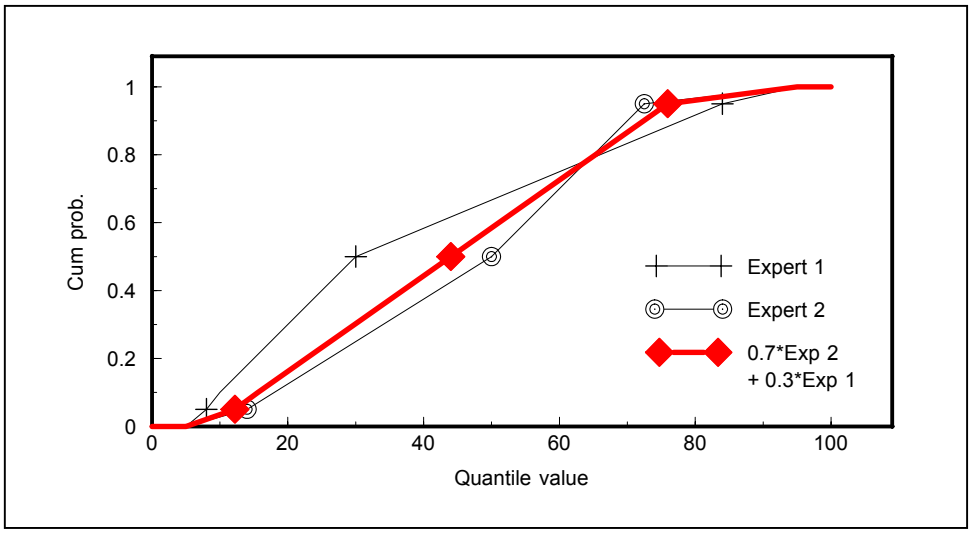

Fig. 5 Weighted combination of two experts' minimum information distributions, in which Expert 1 has weight 0.3 while Expert 2 has weight 0.7. This illustrates the process by which the Decision Maker's interpolated distribution is derived from experts' distributions and their weights ascribed in the Classical Model.

References

Aspinall, W.P., 2006. Structured elicitation of expert judgement for probabilistic hazard and risk assessment in volcanic eruptions. In: Statistics in Volcanology (eds. H.M. Mader, S.G. Coles, C.B. Connor and L.J. Connor) - Special Publications of IAVCEI No. 1; London, The Geological Society for IAVCEI: 15-30.

Brown, A.J. and Aspinall, W.P., 2004. Use of expert elicitation to quantify the internal erosion processes in dams. Proceedings of the British Dam Society Conference, Thomas Telford, pp 282-297

Budnitz R.J., Boore D.M., Apostolakis G., Cluff L.S., Coppersmith K.J., Cornell C.A. and Morris P.A., 1995. Recommendations for probabilistic seismic hazard analysis: guidance on uncertainty and use of experts, NUREG CR-6372, U.S. Nuclear Regulatory Commission.

Budnitz R.J., Apostolakis G., Boore D.M., Cluff L.S., Coppersmith K.J., Cornell C.A. and Morris P.A., 1998. Use of technical expert panels: applications to probabilistic seismic hazard analysis. Risk Analysis 1998;18:463–9.

Cooke, R. M., 1991. Experts in Uncertainty - Opinion and Subjective Probability in Science. Environmental Ethics and Science Policy Series. Oxford University Press, ISBN 0195064658.

Cooke, R.M., 2004. The anatomy of the Squizzle - the role of operational definitions in science. Reliability Engineering & System Safety, 85, 313-319.

Cooke, R.M., 2008. Guest Editorial, Special Issue on Expert Judgment. Reliability Engineering & System Safety. In press, corrected proof. doi:10.1016/j.ress.2007.03.001

Cooke R. and Goossens L., 2000 Procedures guide for structured expert judgment in accident consequence modelling. Radiation Protection Dosimetry, 90(3), 303-309.

Cooke, R.M. and Goossens, L.L.H.J., 2008. TU Delft expert judgment data base. Reliability Engineering & System Safety. In press, corrected proof. doi:10.1016/j.ress.2007.03.005.

Cooke, R.M., Mendel, M. and Thijs, W., 1988. Calibration and information in expert resolution. Automatica, 24, 87-94.

Cooke, R. and Solomatine, D. 1992. EXCALIBR User's Manual. Delft, Delft University of Technology/SoLogic Delft: 33 pp.

Cover, T.M. and Thomas, J.A. 1991. Elements of Information Theory. New York: Wiley.

Goossens L., Cooke R. and Kraan B. (1998) Evaluation of weighting schemes for expert judgment studies. PSAM4 Proceedings, eds. A. Mosleh and R.A. Bari. Vol. 4. Springer, 1937-1942.

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

Kurowicka, D. and Cooke, R., 2006. *Uncertainty Analysis with High Dimensional Dependence Modelling*. Wiley, Series in Probability and Statistics, Chichester: 284pp.

Winkler, R.L., Wallsten, T.S., Whitfield, R.G. Richmond, H.M. Hayes, S.R. and Rosenbaum, A.S.,1995. An assessment of the risk of chronic lung injury attributable to long-term ozone exposure. Operations Research, 43, 19 - 27.

Responding to typical comments

From an expert: *I don't know that*
Response: *No one knows, if someone knew we would not need to do an expert judgment exercise. We are tying to capture your uncertainty about this variable. If you are very uncertain then you should choose very wide confidence bounds.*

From an expert: *I can't assess that unless you give me more information.*
Response: *The information given corresponds with the assumptions of the study. We are trying to get your uncertainty* conditional *on the assumptions of the study. If you prefer to think of uncertainty conditional on other factors, then you must try to unconditionalize and fold the uncertainty over these other factors into your assessment.*

From an expert: *I am not the best expert for that.*
Response: *We don't know who are the best experts. Sometimes the people with the most detailed knowledge are not the best at quantifying their uncertainty.*

From an expert: *Does that answer look OK?*
Response: *You are the expert, not me.*

From the problem owner: *So you are going to score these experts like school children?*
Response: *If this is not a serious matter for you, then forget it. If it is serious, then we must take the quantification of uncertainty seriously. Without scoring we can never validate our experts or the combination of their assessments.*

From the problem owner: *The experts will never stand for it.*
Response: *We've done it many times, the experts actually like it.*

From the problem owner: *Expert number 4 gave crazy assessments, who was that guy?*
Response: *You are paying for the study, you own the data, and if you really want to know I will tell you. But you don't need to know, and knowing will not make things easier for you. Reflect first whether you really want to know this.*

From the problem owner: *How can I give an expert weight zero?*
Response: *Zero weight does not mean zero value. It simply means that this expert's knowledge was already contributed by other experts and adding this expert would only add a bit of noise. The value of unweighted experts is seen in the robustness of our answers against loss of experts. Everyone understands this when it is properly explained.*

From the problem owner: *How can I give weight one to a single expert?*
Response: *By giving all the others weight zero, see previous response.*

From the problem owner: *I prefer to use the equal weight combination.*
Response: *So long as the calibration of the equal weight combination is acceptable, there is no scientific objection to doing this. Our job as analyst is to indicate the best combination, according to the performance criteria, and to say what other combinations are scientifically acceptable.*