Expert judgement – Calibration and combination

Louis H.J. Goossens and Roger M. Cooke
Delft University of Technology, Delft, the Netherlands

ABSTRACT: Governmental bodies and companies are confronted with the problem of achieving rational consensus in the face of substantial uncertainties. Decisions with regard to risks of technical systems must be taken on the basis of predictions of technical and organisational system behaviour. These predictions use mathematical models containing scores of uncertain parameters. Decision makers want to take, and want to be perceived to take, these decisions in a rational manner. The question is, how can this be accomplished in the face of large uncertainties? One available source is experts in the many fields of interest. This paper describes the use of structured expert judgement in a formal manner, in particular addressing the issues of calibrating experts and optimising the combined assessments of a panel of experts. The paper refers to the Procedures Guide of Structured Expert Judgement (Cooke and Goossens 2000) published by the European Union as EUR 18820. This Procedures Guide addresses two methods for using expert judgements developed at Delft University of Technology. The Classical model is able to provide rational probability assessments, because of its use of so-called seed variables. Assessments of seed variaables are asked from experts. The realizations of the seed variables are known by the analysts, but not by the experts. Examples will be referred to for further illustration of applications relevant in the field of risk assessment.

1 INTRODUCTION

The Governmental bodies are confronted with the problem of achieving rational consensus in the face of substantial uncertainties. The area of accident consequence management for nuclear power plants affords a good example. Decisions with regard to evacuation, decontamination, and food bans must be taken on the basis of predictions of environmental transport of radioactive material, contamination through the food chain, cancer induction, and the like. These predictions use mathematical models containing scores of uncertain parameters. Decision makers want to take, and want to be perceived to take, these decisions in a rational manner. The question is, how can this be accomplished in the face of large uncertainties? Indeed, the very presence of uncertainty poses a threat to rational consensus. Decision makers will necessarily base their actions on the judgments of experts. The experts, however, will not agree among themselves, as otherwise we would not speak of large uncertainties. Any given expert's viewpoint will be favorable to the interests of some stakeholders, and hostile to the interests of others. If a decision maker bases his/her actions on the views of one single expert, then (s)he is invariably open to charges of partiality toward the interests favored by this viewpoint.

An appeal to 'impartial' or 'disinterested' experts will fail for two reasons. First, experts have interests; they have jobs, mortgages and professional reputations. Second, even if expert interests could somehow be quarantined, even then the experts would disagree. Expert disagreement is *not* explained by diverging interests, and consensus cannot be reached by shielding the decision process from expert interests. If rational consensus requires expert agreement, then rational consensus is simply not possible in the face of uncertainty.

If rational consensus under uncertainty is to be achieved, then evidently the views of a diverse set of experts must be taken into account. The question is how? Simply choosing a maximally feasible pool of experts and combining their views by some method of equal representation might achieve a form of *political consensus* among the experts in-

volved, but will not achieve *rational* consensus. If expert viewpoints are related to the institutions at which the experts are employed, then numerical representation of viewpoints in the pool may be, and/or may be perceived to be influenced by the size of the interests funding the institutes.

Rational consensus is attainable in the face of large uncertainties if stakeholders commit in advance to the method by which expert views are selected and combined. Once committed to the method of selection and combination, a stakeholder cannot rationally reject the results post hoc without breaking his prior commitment. Such rejection would incur an additional burden of proof: explain *why* the method itself is not sufficient for rational consensus and why the prior commitment to the method should not have been made.

In general, rational decision making requires a quantification of the uncertainties. Therefore expert input to a rational decision process must take the form of quantified expert uncertainties. Expert 'best estimates' will not suffice, as these will not indicate how much the actual (unknown) values may plausibly differ from the 'best estimates'. In our view expert uncertainties should be quantified as subjective probability distributions.

This paper examines the properties which such a method must have. The method of selection of experts is discussed extensively in e.g (Cooke and Goossens 2000), and will not be discussed here. This paper focuses on the method of combination of experts' assessments. Background studies are summarized in section 2. What is undertainty? is briefly explained in section 3. Necessary conditions for rational consensus using expert judgment are discussed in section 4. Section 5 provides an overview of forms in which expert judgements may be cast. Section 6 discusses the issue of performance measures and section 7 describes the implementation of these principles in terms of seed questions for the experts. Section 8 summarizes 3 examples of expert judgement studies and finally section 9 gathers conclusions. Part of the texts in this document come from Cooke *et al* (1999) and Goossens and Cooke (2001).

2 BACKGROUND OF EXPERT JUDGEMENT

The first study which used expert judgements extensively, was WASH 1400 (USNRC 1975), to meet data requirements for the risk assessment of nuclear power plants. Data handbooks also used expert judgements (IEEE STD 500 1977, T-book 1994). The first extensive risk assessment study of chemical installations (Canvey Island 1978) made use of data mostly coming from expert judgements. The NUREG 1150 study (USNRC 1990) was the first attempt of a structured and well-thought procedure for the whole expert elicitation process. Almost a decade later *Guidance on uncertainty and use of experts* (USNRC 1997) was published, at the time the USNRC-CEC study on *Expert judgement and accident consequence uncertainty analysis* (Goossens and Kelly 2000) started. This latter study led to the publication of the *Procedures guide on structured expert judgement* (Cooke and Goossens 2000).

Behavioural and mathematical approaches are available for the elicitation and aggregation of individual experts' assessments (Clemen and Winkler 1999). Mathematical aggregation methods construct a single "combined" assessment per variable by applying procedures or analytical models that operate on the individual assessments. In contrast, behavioural aggregation methods involve interaction of the experts with a view to accomplishing homogeneity of information of relevance to the experts' assessments of the variables of interest. Through this interaction, some behavioural approaches, e.g., Kaplan's expert information approach (Kaplan 1992), aim at obtaining agreement among the experts on the final probability density function obtained per variable. In others, e.g., approaches discussed by Budnitz *et al* (1998) and by Keeney and Von Winterfeldt (1989) the interaction process is followed by simple mathematical combining, such as equal weighting, of the individual experts' assessments in order to obtain a single (aggregated) probability density function per variable. Fixed interaction procedures can be applied, or alternatively, the study team could design a dedicated procedure to suit a particular application. Both mathematical approaches with some modelling and behavioural approaches seem to provide results that are inferior to simple mathematical combination rules (Clemen and Winkler 1999). Furthermore, a group of experts tends to perform better than the average solitary expert, but the best individual in the group often outperforms the group as a whole (Clemen and Winkler 1999). This motivates the elicitation of the assessments of individual experts without any interaction, followed by simple mathematical aggregation in order to obtain a single assessment per variable, thereby weighting the individual experts' assessments based on their quality.

Over the last twenty years the Delft University of Technology has developed methods and tools to support the formal application of expert judgement. Several applications were made for both chemical

substances and nuclear accident consequence assessments, among other fields of interest. Techniques can be applied to give quantitative assessments or qualitative and comparative assessments. The former give rise to assessments of uncertainty in the form of probability distributions, from which nominal values of parameters can be derived for practical applications. The latter lead to rankings of alternatives. Over 25 cases of expert judgement have been executed with the Delft method, about 30,000 elicitation questions were answered. Examples are flange leaks, crane risks, space shuttle propulsion, and composite materials, space devris, groundwater transport, several nuclear consequences (see section 8, example 3), toxicity of chemical substances (see section 8, example 2), water pollution (see section 8, example 1), corrosion in gas pipelines, moveable barriers flood risks, river channel risks, volcano predictions, dike ring failures, bovine diseases, *Campylobacter* in chicken processing industries, falls from ladders, option trading, and prime rent predictions.

The resources required for an expert judgement study vary greatly depending on size and complexity of the study. A trained uncertainty analyst is required for defining the issues and processing the results. Past studies have used between four and twenty experts. The amount of expert time required for making the assessments depends on the subject and may vary between a few hours and a week, per expert. Total time required for studies in the past varies between one man-month to one man-year. Other variables determining the resource commitment are travel, training given to experts in subjective probability assessments and level of documentation. Processing and write up of the results are greatly facilitated by software support.

3 WHAT IS UNCERTAINTY?

The“Uncertainty is that which is removed by becoming certain”. In practical scientific and engineering contexts, certainty is achieved through observation, and uncertainty is that which is removed by observation. Hence uncertainty is concerned with the results of possible observations. Uncertainty must therefore be distinguished from ambiguity. Ambiguity is removed by linguistic conventions regarding the meaning of words. To be studied quantitatively, uncertainty must be provided with a mathematical representation, for instance, as probability.

Within the *subjective* interpretation of probability, uncertainty is a degree of belief of one person, and can be measured by observing choice behaviour. Viewed from the theory of rational decision an assessor’s probabilities are as good as another assessor’s probabilities. There is no rational mechanism for persuading individuals to adopt the same degrees of belief.

A structured uncertainty analysis is indicated for a decision problem when the following features are present:

- Decision making is supported by quantitative models.
- The modelling is associated with potentially large uncertainties.
- The consequences predicted by the models are associated with utilities and disutilities in a non-linear way (threshold effects are the most common instance of this).
- The choice between alternative courses of action might change as different plausible scenarios are fed into the quantitative models.

Expert judgement has always played a large role in science and engineering. Increasingly, expert judgement is recognised as just another type of scientific data, and methods are developed for treating it as such. For applications in uncertainty analysis, we are mostly concerned with random variables taking values in some continuous range. Strictly speaking, the notion of a random variable is defined with respect to a probability space in which a probability measure is specified, hence the term "random variable" entails a distribution. We therefore prefer the term "uncertain quantity", which assumes a unique real value, but we are uncertain as to what this value is. Our uncertainty is described by a subjective probability distribution for uncertain quantities with values in a continuous range.

In the absence of sufficient field or experimental data it is important that the expert assessments are subjected to some kind of performance measure. The measures of performance used apply to discrete events and uncertain quantities. They are designed to be objective and (largely) scale invariant, so that performance on different sets of variables measured on different scales can be compared. Moreover, performance measures should be conservative in the sense that they tie in closely with familiar notions for measuring performance in other areas. They require that experts assess variables whose values become known to the experts *post hoc*. These variables are termed “performance variables”, “calibration variables” or “seed variables”. Performance is measured in two dimensions, namely calibration and informativeness (see section 8).

When expert judgements are cast in the form of distributions of uncertain quantities, the issues of conditionalisation and dependence are important. When

uncertainty is quantified in an uncertainty analysis, it is always conditional on *something*. It is essential to make clear the background information conditional on which the uncertainty is to be assessed.

The Procedures Guide document (Cooke and Goossens 2000) provides details of the protocol for a full expert judgement exercise. The protocol refers in particular to expert judgement exercises with the aim of achieving uncertainty distributions for uncertainty analyses. In that field of application the methods developed at Delft University of Technology have benefited from experiences gained with expert judgement in the US with the NUREG-1150 protocol. For sake of clarity, the Procedures Guide represents a mix of these developments and is not applicable for NUREG-1150 type applications only. The protocol consists of 15 steps (Table 1).

Table 1. Steps in the protocol of structured expert judgement as outlined in the Procedures Guide (Cooke and Goossens 2000)

Preparation for Elicitation:
(1) Definition of case structure
(2) Identification of target variables
(3) Identification of query variables
(4) Identification of performance variables
(5) Identification of experts
(6) Selection of experts
(7) Definition of elicitation format document
(8) Dry run exercise
(9) Expert training session
Elicitation
(10) Expert elicitation session
Post-Elicitation
(11) Combination of expert assessments
(12) Discrepancy and robustness analysis
(13) Feedback
(14) Probabilistic inversion analyses
(15) Documentation

4 NECESSARY CONDITIONS FOR RATIONAL CONSENSUS USING EXPERT JUDGEMENT

The goal of applying structured expert judgment techniques is to enhance rational consensus. Necessary conditions for achieving this goal are laid down as methodological principles (see Cooke 1991) in Table 2. We claim that these are *necessary* conditions for rational consensus, we do not claim that they are sufficient as well. Hence, a rational subject could accept these and yet reject a method which implements them. In such a case, however, (s)he incurs a burden of proof to formulate additional conditions for rational consensus which the method putatively violates.

The requirement of empirical control will strike some as peculiar in this context. How can there be empirical control with regard to expert subjective probabilities? To answer this question we must reflect on the question 'when is a problem an expert judgment problem?' We would not have recourse to expert judgment to determine the speed of light in a vacuum. This is physically measurable and has been measured to everyone's satisfaction. Any experts we queried would give the same answer. Neither do we consult expert judgment to determine the existence of god. There are no experts in the operative sense of the word for this issue. A problem is susceptible for expert judgment, if there is relevant scientific expertise. This entails that there are theories *and* measurements relevant to the issues at hand, but the quantities of interest themselves cannot be measured in practice. For example, toxicity of a substance for humans is measur*able* in principle, but is not measured for obvious reasons. However, there are toxicity measurements for other species which might be relevant to the question of toxicity in humans. Or again, we may be interested in the dispersion of a toxic airborne release at 50 km from the source. Although it is practically impossible to measure the plume spread at 50 km, it is possible to measure this spread at 1 km. If a problem is an expert judgment problem, then necessarily there will be relevant experiments which can in principle be used to enable empirical control.

Table 2. Methodological principles of rational consensus as defined in the Procedures Guide (Cooke and Goossens 2000)

Scrutability/ Accountability	All data, including experts' names and assessments, and all processing tools are open to peer review and results must be 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 are not pre-judged, prior to processing the results of their assessments

5 STRUCTURED EXPERT JUDGEMENT

This section gives a brief overview of methods for utilising expert judgement in a structured manner.

For more complete summaries see Hogarth (1987), Granger Morgan and Henrion (1990), and Cooke (1991).The subject is broken down according to the form in which expert judgement is cast. A final subsection addresses conditionalisation and dependence. In all cases, the judgements of more than one expert are elicited. The questions of measuring performance of experts and combining their judgements are addressed more fully in succeeding sections.

In the world of engineering technical expertise is generally separated from value judgements. Engineering judgement is often applied to bridge the gap between hard technical evidence and mathematical rules on the one hand and unknown characteristics of a technical system. Numerical data have to be derived suitable for the practical problem at hand. Engineers are quite able to provide these required engineering data which are essentially subjective data driven by engineering models and experience. The same is true for expert judgements. Engineering models and experience largely drive the subjective experts' assessments. That is why certain professionals become experts in certain fields of interest.

5.1 *Point values*

In earlier methods, most notably the Delphi method (Helmer 1966), experts are asked to guess the values of unknown quantities. Their answers are single point estimates. When these unknown values become known through observation, the observed values can be compared with the estimates. There are several reasons why this type of assessment is no longer widespread.

First, any comparison of observed values and estimates must make use of some scale on which the values are measured, and the method of comparison must inherit the properties of the scale. For example, percentages are measured on an absolute scale between 0 and 100; mass is measured on a ratio scale (values are invariant up to multiplication by a positive constant), wealth is often referred to an interval scale (values are invariant up to a positive constant and a choice of zero). In other cases values are fixed only as regards rank order (an ordinal scale); a series of values may contain the same information as the series of logarithms of values, etc. To be meaningful, the measurement of discrepancy between observed and estimated values must have the same invariance properties as the relevant scales on which the values are measured. The meaning of "close" and "far away" is scale dependent. This makes it very difficult to combine scores for variables measured on different scales.

A second disadvantage with point estimates is that they give no indication of uncertainty. Expert judgement is typically applied when there is substantial uncertainty regarding the true values. In such cases it is essential to have some picture of the uncertainty in the assessments.

A third disadvantage is that methods for processing and combining judgements are typically derived from methods for processing and combining actual physical measurements. This has the effect of treating expert assessments as if they were physical measurements in the normal sense, which they are not. On the positive side, point estimates are easy to obtain and can be gathered quickly. These types of assessments will therefore always have a place in the realm of the quick and dirty. For psychometric evaluations of Delphi methods see Brockhoff (1966) and Gustafson *et al* (1973), and see Cooke (1991) for a review.

5.2 *Paired comparisons*

In the paired comparison method, experts are asked to rank alternatives pair wise according to some criterion like preference, beauty, feasibility, etc. If 20 items are involved in total, 190 comparisons must be made; each item is compared with the 19 others. Since each item is compared with all the other items, there is a great deal of redundancy in the judgement data. Various processing methods are proposed for distilling a rank order from the pair wise comparison data. According to the method chosen and the availability of some measured values, the data can be further reduced to an interval or even a ratio scale. Paired comparisons were originally introduced for studying psychological responses (Thurstone 1927), and have been applied to consumer research (Bradley 1953), to the assessment of human error probabilities (Comer *et al* 1984), and to the assessment of failure probabilities (Goossens *et al* 1989) and accompanying safety management options (Goossens and Cooke 1997, Hale *et al* 1999). For a mathematical review see David (1963). As with point value assessments, the method of paired comparisons yields no assessment of uncertainty. Methods for evaluating the degree of expert agreement and consistency are available.

5.3 *Discrete event probabilities*

An uncertain event is one that either occurs or does not occur, though we don't know which. The archetypal example is "rain tomorrow". Experts are asked to assess the probability of occurrence of uncertain events. The assessment takes the form of a single

point value in the [0,1] interval, for each uncertain event. The assessment of discrete event probabilities must be distinguished from the assessment of limit relative frequencies of occurrence in a potentially infinite class of experiments (the so-called reference class). The variable "limit relative frequency of rain in days for which the average temperature is 20 degrees Celsius" is not a discrete event. This is not something that either occurs or does not occur; rather this variable can take any value in [0,1], and under suitable assumptions the value of this variable can be measured approximately by observing large finite populations. If we replace "limit relative frequency of occurrence" by "probability", then careless formulations can easily introduce confusion. Confusion is avoided by carefully specifying the reference class whenever discrete event probabilities are not intended.

Methods for processing expert assessments of discrete event probabilities are similar in concept to methods for processing assessments of distributions of random variables. For an early review of methods and experiments see Kahneman *et al* (1982); for a discussion of performance evaluation see Cooke (1991).

5.4 *Distributions of continuous uncertain quantities*

For applications in uncertainty analysis, we are mostly concerned with random variables taking values in some continuous range. Strictly speaking the notion of a random variable is defined with respect to a probability space in which a probability measure is specified, hence the term "random variable" entails a distribution. We therefore prefer the term "uncertain quantity". An uncertain quantity assumes a unique real value, but we are uncertain as to what this value is. Our uncertainty is described by a subjective probability distribution.

We are concerned with cases in which the uncertain quantity can assume values in a continuous range. An expert is confronted with an uncertain quantity, says X, and is asked to specify information about his subjective distribution over the possible values of X. The assessment may take a number of different forms. The expert may specify his cumulative distribution function, or his density or mass function (whichever is appropriate). Alternatively, the analyst may require only partial information about the distribution. This partial information might be the mean and standard deviation, or it might be several quantiles of his distribution. For r in [0,1], the r-th quantile is the smallest number x_r such that the expert's probability for the event $\{X \leq x_r\}$ is equal to r. The 50% quantile is the median of the distribution. Typically, only the 5%, 50% and 95% quantiles are requested, and distributions are fitted to the elicited quantiles.

5.5 *Conditionalisation and dependence*

When expert judgement is cast in the form of distributions of uncertain quantities, the issues of conditionalisation and dependence are important. When uncertainty is quantified in an uncertainty analysis, it is always uncertainty conditional on something. It is essential to make clear the background information conditional on which the uncertainty is to be assessed. This is the role of the "case structure" (Step (1) of the Procedures Guide protocol, see Table 1). Failure to specify background information can lead experts to conditionalise their uncertainties in different ways and can introduce unnecessary "noise" into the assessment process. The background information will not specify values of all relevant variables. Obviously relevant but unspecified variables should be identified, though an exhaustive list of relevant variables is seldom possible. Uncertainty caused by unknown values of unspecified variables must be "folded into" the uncertainty of the target variables. This is an essential task of the experts in developing their assessments. Variables whose values are not specified in the background information can cause dependencies in the uncertainties of target variables. Dependence in uncertainty analysis is an active issue and methods for dealing with dependence are still very much under development. Suffice to say here, that the analyst must pre-identify groups of variables between which significant dependence may be expected, and must query experts about dependencies in their subjective distributions for these variables. Methods for doing this are discussed in Cooke and Goossens (2000) and Kraan and Cooke (2000).

6 PERFORMANCE MEASURES

For deriving uncertainty distributions over model parameters from expert judgements the so-called Classical Model has been developed in Delft (Bedford and Cooke 2001). Other methods to elicit expert judgements are available, for instance for seismic applications (Budnitz *et al* 1998) and nuclear applications (USNRC 1990). The European Union recently finalised a benchmark study among various expert judgement methods (Cojazzi *et al* 2000). As mentioned earlier, in a joint study by the European Communities and the Nuclear Regulatory

Commission the benefits of the latter method (the so-called NUREG-1150 method (Hora and Iman 1989)) have been used incorporating many elements of the Classical model (Goossens and Harper 1998).

The Classical model is a performance based linear pooling or weighted averaging model. The weights are derived from experts' calibration and information performance, as measured on calibration or seed variables. These are variables from the experts' field whose values become known to the experts *post hoc*. Seed variables serve a threefold purpose: (i) to quantify experts' performance as subjective probability assessors, (ii) to enable performance-optimised combinations of expert distributions, and (iii) 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. It contrasts with various Bayesian models.

The Classical model contains three different weighting schemes for aggregating the distributions elicited from the experts. These weighting schemes are equal weighting, global weighting, and item weighting. The different weighting schemes are distinguished by the means by which the weights are assigned to the uncertainty assessments of each expert. The equal weighting aggregation scheme assigns equal weight to each expert. If N experts have assessed a given set of variables, the weights for each density are $1/N$; hence for variable i in this set the decision maker's CDF is given by:

$$F_{ewdm,i} = (1/N) \sum_{j=1}^{N} f_{j,i} \qquad (1)$$

where $f_{j,i}$ is the cumulative probability associated with expert j's assessment for variable i.

Global and item based weighting techniques are termed performance based weighting techniques because weights are developed based on an expert's performance on seed variables. Global weights are determined, per expert, by the expert's calibration score and overall information score. The calibration score is determined per expert by his assessments of seed variables. The information score is related to the width of the uncertainty band and the placement of the median provided by the expert. As with global weights, item weights are determined by the expert's calibration score. Whereas global weights are determined per expert, item weights are determined per expert and per variable in a way that is sensitive to the expert's informativeness for each variable.

The performance based weights use two quantitative measures of performance, calibration and information. Calibration measures the statistical likelihood that a set of experimental results corresponds, in a statistical sense, with the experts' assessments. In particular, the calibration score is the p-value of a standard Chi-square goodness of fit test. Loosely, the calibration score is the probability that the divergence between the expert's probabilities and the observed values of the seed variables might have arisen by chance. A low score (near zero) means that it is likely, in a statistical sense, that the expert's probabilities are 'wrong'. Similarly a high score (near one, but bigger than, say, 0.05) means that the expert's probabilities are statistically supported by the set of seed variables. Information represents the degree to which an expert's distribution is concentrated, relative to some user-selected background measure. The overall information score is the mean of the information scores for each variable. This is proportional to the information in the expert's joint distribution relative to the joint background measure, under the assumption of independence. Independence in the experts' distributions means that the experts would not revise their distributions for some variables after seeing realizations for other variables. Scoring calibration and information under the assumption of independence reflects the fact that expert learning is not a primary goal of the study.

"Good expertise" corresponds to good calibration (high statistical likelihood) and high information. The weights in the classical model are proportional to the product of statistical likelihood and information. When a combined expert has been formed, we can also measure the calibration and information of this combined expert. For more detail see Cooke (1991), Bedford and Cooke (2001) and Cooke *et al* (1988). Calculations are performed with the EXCALIBUR software available through the M.Sc. Risk and Environmental Modeling website: http://ssor.twi.tudelft.nl/~risk/.

In the classical model calibration and information are combined to yield an overall or combined score with the following properties:

1. Calibration dominates over information, information serves to modulate between more or less equally well calibrated experts,
2. The score is a long run proper scoring rule, that is, an expert achieves his/her maximal expected score, in the long run, by and only by stating his/her true beliefs. Hence, the weighting scheme, regarded as a reward structure, does not bias the experts to give assessments at variance with their real beliefs, in compliance with the principle of neutrality.

3. Calibration is scored as 'statistical likelihood with a cut-off'. An expert is associated with a statistical hypothesis, and the seed variables enable us to measure the degree to which that hypothesis is supported by observed data. If this likelihood score is below a certain cut-off point, the expert is unweighted. The use of a cut-off is driven by property (2) above. Whereas the theory of proper scoring rules says that there must be such a cut off, it does not say what value the cut-off should be.
4. The cut-off value for (un)weighting experts is determined by optimising the calibration and information performance of the combination.

A fundamental assumption of the Classical model (as well as Bayesian models) is that the future performance of experts can be judged on the basis of past performance, as reflected in the seed variables. Seed variables enable empirical control of any combination schemes, not just those that optimise performance on seed variables. Examples of expert judgement studies using seed variables are available and references are provided in this paper. Therefore, choosing good seed variables is of general interest, see Goossens *et al* (1996, 1998) for backgrounds and details. The Classical model follows the steps of the Procedures Guide summarised in Table 1.

7 SEED QUESTIONS

A fundamental assumption of the classical (as well as the Bayesian) model is that the future performance of experts can be judged on the basis of past performance, reflected in the so-called seed variables. **Seed variables** are variables of which the true values are known by the analyst or can be found within the time span of the study. The performance of the experts on the seed variables is taken as indicative for the performance on the variables of interest. Therefore the seed variables must resemble as much as possible the variables of interest. The more seed variables the better, but ten is certainly sufficient. The success of any implementation depends to a large measure on defining relevant variables whose true values become known in a reasonable time frame. This requires resourcefulness on the part of the analyst as well as the sympathetic cooperation of the experts themselves. It is essential that the experts understand the model and generally appreciate its potential usefulness.

Letting p_{DM} denote the decision maker's distribution for an uncertain item, and letting p_e [e = 1,2,...,E] denote the distributions of experts 1,2,...,E for the same item, then p_{DM} is a weighted combination of p_1, p_2,..., p_E if:

$$p_{DM} = \Sigma_e w_e p_e \quad e = 1,2,...,E \qquad (2)$$

where w_e is expert's e weight, and $\Sigma_e w_e = 1$ and $w_e \geq 0$.

The weights are determined by the "theory of proper scoring rules", and by measures of **calibration** and **informativeness**. The mathematical details can be found in Cooke (1991). We give here a rough idea of these concepts and how they are used. Informativeness measures the degree to which an information is "concentrated". Calibration is a measure derived from the classical theory of hypothesis testing, and reflects the degree to which the experts' performance on seed variables "supports" the hypothesis that the expert's probability statements "correspond with reality". To give a rough idea how this works, suppose we elicit the 5%, 50% and 95% quantiles, or percentiles for each (seed) variable, or item. For each item, the expert's probability is 5% that the true value falls beneath his 5% quantile, etc. If we distinguish the four possible "interquantile ranges" into which the true values can fall, then for each item the expert's distribution $p = p_1,...,p_4$ over these four ranges is

$$p_1 = 5\%, p_2 = 45\%, p_3 = 45\%, p_4 = 5\%. \quad (3)$$

According to the expert's distribution there is, for example, a 10% probability that any true value falls outside his/her 90% central confidence band, i.e. falls below his/her 5% quantile or above his/her 95% quantile. If this actually occurred for 50% of the seed variables, then we should say that this seed data gives little support to the hypothesis that the expert's probabilities correspond with reality. Let $s = s_1,...,s_4$ denote the sample distribution, reflecting the relative frequencies with which the true values fall in the interquantile ranges. Using standard statistical techniques we can measure calibration as "statistical likelihood", that is as

the probability that we should see at least as much disagreement between s and p as found in the expert's performance on the seed variables, supposing that the distribution p were really correct.

High values of this probability correspond to good calibration, low values to poor calibration. "Good expertise" corresponds to good calibration (high statistical likelihood) and high information. Moreover, calibration should "dominate" over information: we do not want very highly informative distributions unless they are well calibrated, rather we want information to discriminate between more or less equally well

calibrated experts. The weights in the classical model are proportional to the product of statistical likelihood and information, and it turns out that this product is indeed dominated by calibration.

One additional ingredient in the weights is derived from the requirement that the (unnormalized) weights be strictly proper scoring rules. A scoring rule is a method of assigning a number (a score) to a set of probability assessments, on the basis of observed realizations. The weights used in the classical model (see formula (2)) are eo ipso scores in this sense. A score is called *strictly proper* if an expert can achieve his highest expected score by and only by stating his true opinion. The measures for information and calibration described above must be combined in such a way that the result is (in the long run) a strictly proper scoring rule. This requires that the measurement of calibration be combined with "classical significance tests". Briefly, there must be some value $\alpha > 0$, such that if the expert's statistical likelihood drops below α, his/her weight becomes zero. The theory of strictly proper scoring requires that such an α be used, but does not say what α should be. In the classical model, α is chosen such that the resulting calibration and informativeness of the decision maker (DM) is optimal.

The seed variables (or calibration variables) are not only important in determining the weights for combining experts' assessments, but also they provide empirical evidence of the performance of the combined assessment (the "optimised decision maker") and thus form an important feedback to the experts. Two types of seed variables are available:

1. Domain variables: these variables fall in particular in the field of the experts.
2. Adjacent variables: these variables fall into fields which are adjacent to the field of expertise of the experts in question.

Crucial issues to be investigated are:

- the number of seed variables
- the dependence beween seed variables; i.e., seed variables are experimental results; if several seed variables are taken from the same experiment they may be dependent on the true values
- the selection of adjacent seed variables in cases where domain seed variables are not available, e.g. in cases of fatal responses to high toxic doses
- preknowledge of an expert on the true value of some of the seed variables; the true values of the seed variables are supposed not to be known directly by the experts; however, if an expert has access to the results of a particular experiment from which seed variables are taken, (s)he may know the true value and may provide subjective assessments with high informativeness contents; such situations need be avoided.

8 EXAMPLES OF EXPERT JUDGEMENT STUDIES

Three examples will be presented to illustrate the practical use of expert judgements. The examples come from fields adjacent to the field of assessment and management of environmental risks. In the first example risk management is modelled to derive the water pollution risks of establishments under the European Commission's Seveso-Directive for Major Hazards Control. The Classical model is used to derive the relative failure rates of various types of chemical activities in industry. This example shows in particular the application of adjacent seed variables and the treating of small differences in DM-scores for the three weighting schemes as outlined in Section 6.

In the second example the Classical model is used to derive the coefficients of the dose-response-relations describing the fatal consequences of exposure to large amounts of toxic chemicals. This example shows in particular the structured choice of seed variables and the application of probabilistic inversion techniques to provide uncertainty distributions of the coefficients of the probit relations.

In the third example the Classical model was used to derive uncertainty distributions of the important model parameters of nuclear accident consequence modelling. Although the decision was made to output only according to the equal weighting scheme, a diversity of the item weights has been made possible which shows how different performance based weighting results can be.

8.1 *Example 1: Water pollution risk management*

The EC-directive on Major Hazards Installations (the so-called Seveso-Directive) is implemented in the Netherlands in 1988. The extended EC-Directive with regard to environmental risks was established in 1992. The Dutch Ministry of Housing, Physical Planning and Environment (responsible for the implementation of the EC-Directive in the Netherlands) had decided to provide industries with a dedicated methodology in order to fulfil these EC-Directive requirements, and had developed a software package (VERIS) for ready use by industries.

The methodology determines the environmental risks of an installation from accident scenarios and their consequences to surface water pollution. Plant

operators can assess the consequences directly by applying on-site data of quantities of hazardous materials, distances of installations to surface waters, and by taking into account the impact of mitigation measures provided by the plant (for example, second containments, dikes and effluent treatment).

The assessments of the frequencies of the accident scenarios cannot be derived directly from company data and is therefore guided by a *generic framework*. Instead of requiring a full probabilistic approach, industry management is asked to rate several generic features of their installation's management on a four point scale ranging from very good performance (lowest point on scale) to poor performance (upper most point on scale). The data points on the four point scale were derived based on standard questions of which the answers were controllable to regulatory bodies afterwards. For instance, one such question asks for the frequency of a specific inspection: on the four point scale this may be once a week, once a month, once every year, and almost never, expressing essentially a range from very good performance to poor performance. The numerical values associated with the range of the four point scale in the generic framework was derived by expert judgement.

In the methodology eight basic chemical activities are defined covering all activities which take place in the industries (see left columns of Table 3). For each basic activity eight groups of *influential factors* are defined (see right columns of Table 3), which cover all possible failure causes. The total capability of failure of each chemical activity is reflected by the effectiveness of each individual influential factor compared to its specific contribution. Suppose, for instance, that for tank storage, the "lay-out in general" is assessed to contribute 7 percent of the total contribution of all influential factors. Poor performance (the lower limit) of the "lay-out in general" of a specified storage tank in company X will then already contribute 7 percent. A second influential factor may contribute 17 percent, but in case the performance is very good, this factor does not add up to the total failure of the storage tank.

Table 3. Basic activities and influential factors of risk management in the chemical industries

Basic activity	Influential factor
Tank Storage	Lay-out in general
Storage in warehouse	Organisation in general
Continuous process	Procedures in general
Batch process	Emergency precautions
(Un)loading trains/cars	Supervision/operators
(Un)loading ships	Design/condition install.
Transfer to small containers	Specific procedures
Transfer in units	Maintenance

For each basic activity paired comparisons were used to derive the relative contributions of the influential factors in water pollution risk management. In order to achieve a numerical output for the representation of water pollution frequencies, generic failure data were required for the range of very good performance to poor performance of the eight basic chemical activities. These generic data were derived by using the Classical model.

This expert judgement exercise aimed at getting a subjective assessment of all individual plants in the Netherlands which fall under the post-Seveso-directive guidelines. The study can be characterised as what it is not intending to do: it is not a generic picture of the Dutch chemical installations, and it is not an average picture of the Dutch installations. Indeed, it is a current picture of the Dutch plants under the current state-of-the-art of chemical installations designs. In words, the experts were expected to sit back and consider the whole present Dutch chemical industry and were asked to subjectively assess the relative failure rates of the defined chemical activities. In this case, relative means relative to each other (comparison of activities).

Suppose, the Dutch chemical industries would have X_N installations of each specified activity, the experts were asked to consider the failure rate of each activity compared to one specified activity, namely storage tanks. The median failure rate of all X_N storage tanks was set at a value equal to one. The experts were asked to assess the failure rates of all X_N continuous reactors in the Netherlands, and they had to note down their median assessment of the continuous reactors in terms of a factor, with which the failure rate of the storage tanks has to be multiplied to get the median continuous reactor failure rate. And so forth.

They then provided a subjective assessment of the 90 percent central confidence band for the continuous reactors indicating the range of (relative) failure rates within which $0.9 * X_N$ continuous reactors would fall.

Summarising, the target variables were *individual cases* for which *relative assessments* were to be made. Furthermore, there were *no data available* on the target variables. No domain variables were available either. The choice of seed variables was driven by the following considerations:

- adjacent variables were necessary which covered an identical type of subjective assessments; data from equipment failures in the chemical industries and incident registration data seemed appropriate

- the seed variables were mostly phrased in terms of relative assessments, for instance, comparing failure rates of two comparable pieces of equipment
- generic equipment failure rates from data books were used; the data generally come from individual behaviour of equipment, and not from averages over a large population of equipments.

The results for the 'virtual weights' of the DM scores are summarised in Table 4. 'Virtual weight' is the weight that the combination would receive if added to the expert panel as an additional virtual expert. A virtual weight of one half or more indicates that the combination would receive more weight than the real experts cumulatively.The difference in DM-scores is rather low, albeit that the item weights show a bit better performance.

Table 4. Summary of DM-scores for all three weighting schemes

weighting scheme	calibration score	relative informativeness score	DM-scores (virtual weights)
item wts	0.35	1.872	0.650
global wts	0.25	1.802	0.560
equal wts	0.35	1.381	0.576

8.2 *Example 2: Dose-response relations for toxic substances*

Under the same Seveso-Directive the Dutch Ministry of Housing, Physical Planning and Environment has developed procedures to establish quantitative risk analyses, in order to meet quantitative limits for the accepted (individual and group) risks of major hazards installations. For example, the individual risk is defined as the probability of a fatality as a result of an incident while being constantly present at specified distances from a major hazards installation. By adjusting the limits of the accepted risk level to this definition, the calculated risks of a particular installation can be judged to be acceptable or not. Iso-risk contours are used to indicate risk profiles at distances around the installation. In the Netherlands, for instance, the iso-risk contour describing 10^{-6} deaths per year is used as a land-use planning instrument to mark a zone between the (new) installation and housing.

A major part of the risks associated with chemical installations arises from the exact dose-response relationship, relating exposure concentrations and exposure times of inhaled toxic chemicals to the (lethal) response of the exposed individuals. Although many dose-response representations are possible, the calculations must use probit relations for the risk assessments as required under the Dutch law. In practical applications the probit relation is expressed by

$$Pr = a + b \,.\, \ln(C^n t) \qquad (5)$$

in which a is a dimensionless constant indicative for the dose at which lethal effects begin, b represents the slope of the probit relation, C is the concentration of hazardous materials (in ppm or mg/m^3), t is the exposure time (in minutes), and n is the exponent indicating the relative influence of C to the probit value with respect to values of t.

The experts were asked to provide assessments on observable quantities only and not on the coefficients in the mathematical formula of the probit relation. These coefficients were derived from the assessments by probabilistic inversion techniques in step (14) of the Procedures Guide (see Table 1). The experts were asked to assess three quantile points of the concentrations C of toxic substances if exposed during t = 30 minutes at three lethality levels (10%, 50% and 90% lethality).

The Classical model was used to derive values for the probit coefficients of five chemical substances: acrylonitrile, ammonia, hydrogen fluoride, sulphur trioxide and azinphos-methyl (a pesticide). For defining the performance variables in step (4) of the procedure a *classification model of inhalation* was developed. The main purposes of the model were:

- to characterize the toxic material in all phases of the toxic process
- to find animal models sharing properties in one or more steps of the model with the general human model
- to identify other toxics with similar properties within one or more steps of the model.

The Classification Model can be represented by a number of 'dimensions':

- kinetics: (quantitative) properties concerning the rates of absorption, distribution, metabolism and elimination of the substance
- mechanisms (or dynamics): (qualitative) properties concerning the types of reaction, and the formation of metabolites, during absorption down to excretion
- target organs: the organs where the toxic impact will occur
- functional disturbances: (pathophysiogical) changes in organ-functioning as a result of the toxic impact
- health effects: clinical expression of the organ-function disturbances.

Input into the model is a range of concentrations, or dose rates, of a certain toxic substance. The output of the model contains the values derived within the

various parts of the model. In this application, the model is quantified with values leading to acute lethal responses for a human population. These properties define one or more paths throughout the model, and these paths provide a basis for the classification of the chemical substance. Table 5 presents an overview of the performance variables matched on the classification model of inhalation.

Table 5. Distribution of 'dimensions' of the Classification Model of Inhalation over the seed variables for five chemical substances (K=kinetics, M=mechanics, TO=target organs, FD=functional disturbances, HE=health effects). NB. Some seed variables covered more than one dimension, for which reason the sum of dimensions may be larger than the number of seed variables applied

chemical substance	# of seed variables	'dimensions' K	M	TO	FD	HE
acrylonitrile	10	8	4	2	-	2
ammonia	10	3	1	3	3	3
hydrogen fluoride	9	6	-	-	-	3
sulphur trioxide	10	2	1	3	1	6
azinphos-methyl	10	6	2	1	1	5

Table 6. Summary of DM-scores ('virtual weights') of three chemical substances: PW = performance weights, EW = equal weights

chemical substance		calibration score	information score	DM-score (virtual weights)
acrylonitrile	PW	0.2400	3.186	0.500
	EW	0.2800	1.511	0.233
ammonia	PW	0.1100	1.672	0.341
	EW	0.2800	1.075	0.457
sulphur trioxide	PW	0.1400	3.904	0.745
	EW	0.1400	2.098	0.611

chemical substance	probit relation
acrylonitrile PW	$Pr = -8.17 + 1.12 \ln(Ct)$
ammonia PW	$Pr = -36.4 + 2.01 \ln(C^2t)$
sulphur trioxide PW	$Pr = -2.85 + 0.68 \ln(C^2t)$

Extensive literature surveys of quantitative aspects of each chemical's acute toxicity regrouped a preliminary list of world-wide experts into 'major contributors' and 'less often contributing' experts. Experts were selected on either of the following criteria: 1) major contribution to a criterion document, monograph or review article, 2) at least leading author of two scientific research papers, of which one within the last five years, or 3) known to the 'chemical's scientific community'. The selected experts come from institutions having academic interests, and from industrial or regulatory environments. Twenty-seven experts were selected. Results of the expert judgement exercise are published elsewhere (Goossens *et al* 1998). The 'virtual weights' (for explanation of the virtual weight, see the end of Section 8.1) of the DM-scores and values of the probit coefficients are summarised in Table 6. Only three out of the five chemicals were successful in deriving a probit relation.

The distributions on the probit coefficients are the result of probabilisic inversion. For details on this application, see Cooke (1994) and for the current state of the art, see Kurowicka and Cooke (2005a and 2005b).

8.3 *Example 3: Nuclear accident consequence risk modelling*

The U.S. Nuclear Regulatory Commission (USNRC) and the European Commission (EC) have both developed probabilistic accident consequence codes: MACCS (Chanin *et al* 1990) in the United States and COSYMA (Kelly 1991) in Europe. Uncertainty analyses have been performed with predecessors of both codes, whereby the probability distributions utilised were assigned primarily by the consequence code developers rather than by phenomenological experts in the many different scientific disciplines that provide input to a complete consequence code. For that reason, the decision was made to execute a full uncertainty analysis on each code separately, whereby most of the uncertainty distributions of the code input parameters were derived using formal expert judgement (Goossens and Harper 1998). See also Goossens (2005) in the Proceedings of this Workshop. An overview of the joint expert judgement studies are shown in Table 7.

Table 7. Phenomenological areas with expert panels and number of questions in the EC/USNRC joint project (NOTE: the countermeasures panel was performed as an EC project only)

Expert panel	Year of panel	NUREG / CR - *EUR* *report*	# of experts in panel	# of elicitation questions	# of seed questions
Atmospheric dispersion	1993	6244 *15855/15856*	8	77	23
Deposition (dry and wet)	1993	6244 *15855/15856*	8	87	14 dry 19wet
Behaviour of deposited materials and its related doses	1995	6526 *16772*	10	505	none
Food chain on animal transfer and behaviour	1995	6523 *16771*	9	80	8
Food chain plant/soil transfer and processes	1995	6523 *16771*	6	244	31
Internal dosimetry	1996	6571 *16773*	9	332	55
Early health effects	1996	6545 *16775*	10	489	15

Late somatic health effects	1996	6555 *16774*	10	106	8
Counter-measures	1999	n/a *18821*	10	111	none

The experts, who do not necessarily have to be familiar with the codes, were neither forced to provide uncertainty distributions on code input parameters, nor to believe in the models used in the codes. Instead, they were asked to provide assessments on variables, which, in principle, are observable and measurable. The results are published in EUR-/NUREG-reports (Table 7) and summarised in a special issue of *Radiation Protection Dosimetry* (Goossens and Kelly 2000).

For programmatic reasons of assignment, the aggregation process was done using equal weights for all panels of experts. As the individual expert's assessments differed from each other, equal based aggregation resulted in relatively wide uncertainty distributions of the decision maker's distributions. For most of the panels, seed questions were also available (see Table 7) to test the differences with the general equal weighting outcomes (see Table 8). For the late health effects panel the seed questions referred to future outcomes of the Japanese atomic bomb survivors' data. For the deposited materials and countermeasures panels no seed questions were available. Table 8 shows the performance based combination and the equal weight combination for the other seven panels. For each panel, Table 8 shows the calibration score (1 is maximal, 0 is minimal), the mean information score (0 is minimal), and the 'virtual weight' (for explanation of the virtual weight, see the end of Section 8.1).

Table 8. Performance based and equal weight combinations

Panel	Weighting scheme	Calibration score	Information score	DM-score (virtual wts)
DISP	Item wts	0.9000	1.024	0.80545
	Equal wts	0.1500	0.811	0.33166
DEPOS				
dry dp	Item wts	0.5200	1.435	0.50000
	Equal wts	0.0010	1.103	0.00168
wet dp	Item wts	0.2500	1.117	0.93348
	Equal wts	0.0010	0.793	0.07627
ANIML	Item wts	0.7500	2.697	0.50000
	Equal wts	0.5500	1.778	0.19204
SOILPL	Item wts	0.0010	1.024	0.13369
	Equal wts	0.0010	0.973	0.12779
DOSIM	Item wts	0.8500	0.796	0.52825
	Equal wts	0.1100	0.560	0.09217
EARLY	Item wts	0.2300	0.216	0.98749
	Equal wts	0.0700	0.165	0.94834

Apart from the SOIL/PLANT case, the performance based combination performs well; the calibration scores are not alarmingly low, and the virtual weight is high. The equal weight combination sometimes returns good calibration and high virtual weight, but these scores are lower than those of the performance based combination. In the case of SOIL/PLANT, we must conclude that the evidence gathered from the seed variables does not establish the desired confidence in the results. Although it might be argued that 31 seed variables constitutes a rather sever test of calibration, reducing the effective number of seed variables to 10 still yields poor performance (calibration scores 0.04 and 0.01 for the performance based and equal weight combinations respectively). In general, the number of effective seed variables is equal to the minimum number assessed by some expert. Hence the effective number in INTERNAL DOSIMETRY is 28 and in ANIMAL is 6. Experts are scored on the basis of the effective number of seed variables; lowering this number is comparable to lowering the power of a statistical test. Thus we cannot directly compare calibration scores of different panels without first setting the effective number of seed variables equal.

In DISPERSION, ANIMAL and INTERNAL DOSIMETRY, the results of equal weighting are not dramatically inferior to the performance based combination. In such cases, a decision maker giving priority to *political* rather than *rational* consensus might apply equal weight combination without raising questions of performance. In the other cases the evidence for degraded performance in the equal weight combination, in our opinion, is strong.

9 CONCLUSIONS

The Delft method has by now generated a substantial amount of experience with structured expert judgment. Over 30,000 individual elicitations have been performed, and there is extensive data on expert assessments for uncertain quantities for which the true values are known post hoc. This data, in suitably scrubbed form, is available upon request from the second author.

The overall conclusions from this experience may be summarized as follows:

1. Valid measures of performance for subjective probability assessors exist and can be applied.
2. Experts' performance as subjective probability assessors is highly variable. There is strong variation across panels, and within panels.

3. Equal weight combinations generally lead to statistically acceptable performance, often with a very significant loss of informativeness.
4. Performance-based combinations of expert judgments outperform the equal weight combination, and the best expert in the overwhelming majority of cases.
5. Experts have no problem in quantifying their uncertainty, if the questions are carefully formulated and directed to observable quantities with which the experts are familiar.
6. Experts are generally quite supportive of performance measurement, and positively appreciate the introduction of objective criteria to validate performance, both their own and that of any resulting combination.

References

Bedford, T..J. & Cooke R.M. (2001). *'Probabilistic risk analysis, foundations and methods'*. Cambrigde University Press.

Bradley, R. (1953). 'Some statistical methods in taste testing and quality evaluation'. *Biometrica*, vol 9, 22-38.

Brockhoff, K. (1975). 'The performance of forecasting groups in computer dialogue and face to face discussions'. In: Linstone, H., & Turoff, M. (eds). '*The Delphi Method, Techniques and Applications.'* Addison Wesley, pp 291-321.

Budnitz, R.J., Apostolakis, G., Boore, D.M., Cluff, L.S., Coppersmith, K.J., Cornell, C.A. & Morris, P.A. (1998). 'Use of technical expert panels: Applications to probabilistic seismic haxard analysis'. *Risk Analysis* vol 18, 463-469.

Canvey Island, (1998). Report on Canvey Island risk assessments.

Chanin, D.I., Sprung, J.L., Ritchie, L.T. & Jow, H.-N. (1990). '*MELCOR Accident Consequence Code System (MACCS): User's Guide*'. Report NUREG/CR-4691, Albuquerque/NM/USA.

Clemen, R. T., & Winkler, R. L. (1999). 'Combining probability distributions from experts in risk analysis'. *Risk Analysis* vol 19, 187-203.

Cojazzi, G., Fogli, D. & Grassini, G. (2000). '*Benchmark exercise on expert judgement techniques in PSA Level 2, Phase I ; Summary and evaluation of unstructured and structured expert judgment results'*. Report to the EC. EUR 19738.

Comer, K., Seaver, D., Stillwell, W. & Gaddy, C. (1984). '*Generating Human Reliability Estimates Using Expert Judgement'*. NUREG/CR-3688.

Cooke, R.M. (1991). '*Experts in uncertainty*'. Oxford University Press.

Cooke, R.M. (1994). 'Parameter fitting for uncertain models: modelling uncertainty in small models;'. *Reliability Engineering and System Safety*, vol 44, 89-102.

Cooke, R.M., & Goossens, L.H.J. (2000). '*Procedures guide for structured expert judgement'*. European Commission. Report EUR 18820.

Cooke, R.M., Kraan, B.C.P. & Goossens, L.H.J. (1999). 'Rational consensus under uncertainty: Expert judgement in the EC-USNRC uncertainty study'. *VALDOR Values in Decisions*, Conference Proceedings, Stockholm, Sweden, June 13 – 17, 1999, 9 pages.

Cooke, R.M., Mendel, M. & Thys, W. (1988). 'Calibration and information in expert resolution: a classical approach'. *Automatica*, vol 24, 87-94.

David, H. (1963). '*The Method of Paired Comparisons'*. Charles Griffin.

Goossens, L.H.J. (2005). *'Expert judgement elicitation on probabilistic accident consequence codes.'* Appearing in this Proceedings.

Goossens, L.H.J. & Cooke, R.M. (1997). 'Applications of some risk assessment techniques: Formal expert judgement and accident sequence precursors'. *Safety Science*, vol 26, 35-48.

Goossens, L.H.J. & Cooke, R.M. (2001). 'Expert judgement elicitation in risk assessment'. In: Linkov, I. & Palma-Oliveira, J. (eds.). '*Assessment and management of environmental risks'*. Kluwer Academic Publishers, 411-426.

Goossens, L.H.J., Cooke, R.M. & Kraan, B.C.P. (1996). '*Evaluation of weighting schemes for expert judgement studies'*, Report prepared for the European Commission, Delft University of Technology.

Goossens, L.H.J., Cooke, R.M. & Kraan, B.C.P. (1998). 'Evaluation of weighting schemes for expert judgement studies'. In: Mosleh, A. & Bari, A. (Eds.). '*Probabilistic Safety Assessment and Management'*. Springer, vol 3, 1937-1942.

Goossens, L.H.J., Cooke, R.M., & van Steen, J. (1989). '*Expert Opinions in Safety Studies'*, vols. 1 – 5. Philosophy and Technical Social Sciences, Delft University of Technology.

Goossens, L.H.J., Cooke, R.M., Woudenberg, F. & van der Torn, P. (1998). 'Expert judgement and lethal toxicity of inhaled chemicals'. *J. Risk Res*. vol 1, 117-133.

Goossens, L.H.J. & Harper, F.T. (1998). 'Joint EC/USNRC expert judgement driven radiological protection uncertainty analysis'. *J. Radiol. Prot*. vol 18, 249-264.

Goossens, L.H.J. & Kelly, G.N. (2000). 'Expert judgement and accident consequence uncertainty analysis'. Special Issue. *Radiat. Prot. Dosim*. vol 90, 293-381.

Granger Morgan, M. & Henrion, M. (1990). *'Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis'*. Cambridge University Press.

Gustafson, D., Shulka, R., Delbecq, A., & Walster, A., (1973). 'A comparative study of differences in subjective likelihood estimates made by individuals, interacting groups, Delphi groups and nominal groups'. *Organ. Behaviour and Human Performance* vol 9, 280-291.

Hale, A.R., Costa, M.A.F., Goossens, L.H.J. & Smit, K. (1999). 'Relative importance of maintenance management influences on equipment failure and availability in relation to major hazards'. In: Schuëller, G.I. & Kafka, P. '*Safety and Reliability*', ESREL '99, vol. 2, 1327-1332.

Helmer, O. (1966). '*Social Technology'*. Basic Books.

Hogarth, R. (1987). *'Judgement and choice'*. Wiley.

Hora, S. & Iman, R. (1989). 'Expert opinion in risk analysis: the NUREG-1150 methodology'. Nucl Sci Engineering, vol 102, 323-331.

IEEE STD 500 (1977). *'IEEE Guide to the collection and presentation of electrical, electronic and sensing component reliability data of nuclear power generation stations.'*

Kahneman, D., Slovic, P. & Tversky, A. (eds) (1982). '*Judgement under Uncertainty, Heuristics and Biases*,'. Cambridge University Press.

Kaplan, S. (1992). ''Expert information' versus 'expert opinions'. Another approach to the problem of eliciting/combining/using expert knowledge in PRA'. *Reliability Engineering and System Safety,* vol 35, 61-72.

Keeney, R. L., & Von Winterfeldt, D. (1989). On the Uses of Expert Judgment on Complex Technical problems. *IEEE Transactions on Engineering Management,* vol 36, 83-86.

Kelly, G.N. (1991). '*COSYMA: A new programme package for accident consequence assessment*'. Report EUR 13028.

Kraan, B. & Cooke, R.M. (2000). 'Processing expert judgements in accident consequence modelling'. *Radiat. Prot. Dosim,*vol 90, 311-316.

Kurowicka, D. & Cooke, R.M. (2005a). 'Techniques for generic probabilistic inversion'. *Comput. Stat. and Data Anal.* (appearing).

Kurowicka, D. & Cooke, R.M. (2005b). '*Uncertainty analysis and high dimensional dependence modelling*'. Wiley (appearing).

T-Book (1992). '*Reliability Data of Components in Nordic Nuclear Power Plants'*. 3rd edition, prepared by the ATV Office and Studsvik AB,Vattenfall AB.

Thurstone, L.L.(1927). 'A law of comparative judgment'. *Psycho.Review.* vol 34, 273-286.

USNRC (1975). '*Reactor safety study WASH-1400'*. NUREG 751014.

USNRC. (1990). '*Severe accident risks: An assessment for five US nuclear power plants'*. Report NUREG-1150

USNRC (1997). '*Guidance on uncertainty and use of experts.*' Report to NRC.