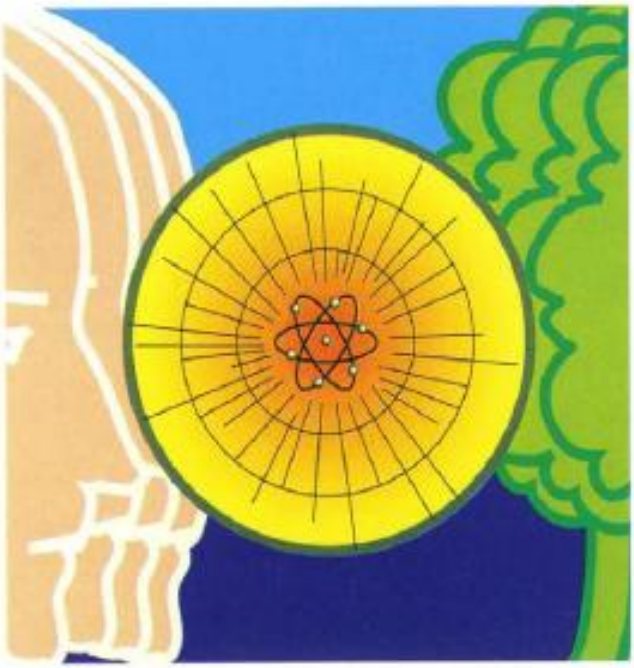
European Commission

Community research

Project report

Nuclear science and technology

Procedures guide for structured expert judgment



EURATOM

EUR 18820 EN

European Commission

nuclear science and technology

Procedures guide for structured expert judgment

R. M. Cooke, L. J. H. Goossens

Delft University of Technology
Delft, the Netherlands

Contract No FI4P-CT95-0006

Work performed as part of the European Atomic Energy Community's R & T
specific programme 'Nuclear fission safety 1994–98'
Area D: 'Radiological impact on man and the environment'

Directorate-General for Research

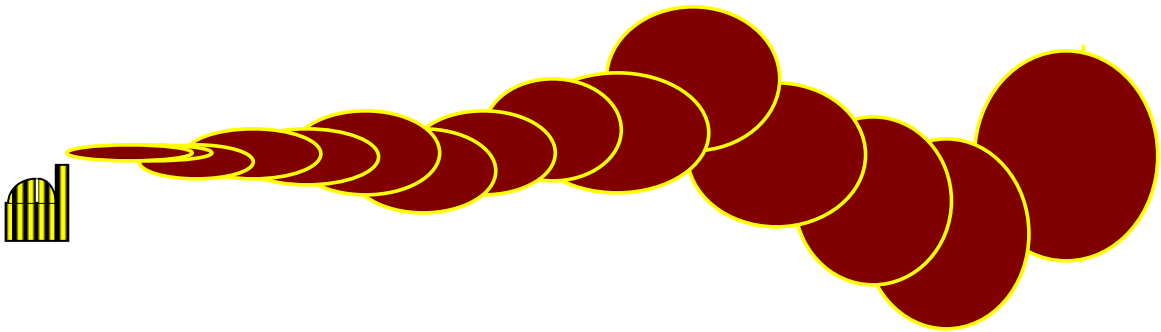
2000

EUR 18820 EN

PROCEDURES GUIDE

for

STRUCTURED EXPERT JUDGMENT



R.M. Cooke, L.H.J. Goossens
Delft University of Technology Delft,
The Netherlands
Report prepared under contract No. ETNU-CT93-0104-NL for the Commission of European Communities
Directorate-general XI (Environment and Nuclear Safety) Directorate D (DG 11 CCJH)
June 1999

EUR 18820, Luxembourg/Brussels

FORWARD

This document is a guide for using structured expert judgment to quantify uncertainty in quantitative models. The methods applied here have been developed by a host or researchers over the last 30 years. During the years 1990 - 1999, the European Commission and the United States Nuclear Regulatory Commission undertook a joint uncertainty study of accident consequence codes for nuclear power plants using structured expert judgment. The purpose was not only to perform an uncertainty analysis of the US accident consequence code MACCS and the European accident consequence code COSYMA. The wider purpose was to form a baseline for the state of the art in using structured expert judgment for quantifying uncertainty. The reports emerging from this work are intended to be useful outside the community of nuclear accident consequence modeling. Indeed the quantification of uncertainty in the modeling of dispersion, deposition, foodchain transport, and cancer induction, may be used in many fields of environmental modeling and health protection. In the same spirit the methods for using structured expert judgment to quantify uncertainty are applicable far beyond the accident consequence modeling community.

Results of the EU-USNRC joint research with regard to expert judgment are summarized in APPENDIX I. Reports from this joint research are listed in APPENDIX II. Training material developed in the joint research is included as APPENDIX V.

We gratefully acknowledge the enthusiastic and wholehearted participation of all the experts in this study. All experts provided written rationales explaining how they arrived at their assessments. These rationales, reproduced in the published reports, are highly recommended to those readers who wish deeper insight into the sources of uncertainty in these issues. We also gratefully acknowledge the help of many institutes, in particular the National Radiological Protection Board, the Forschungszentrum Karlsruhe and the Energy Centrum Netherlands, in developing seed variables.

TABLE OF CONTENTS

PART I Generic Issues

1. What is Uncertainty
2. When and how Should Uncertainty Analysis be Performed?
3. Structured Expert Judgment
4. Performance Measures
5. Combinations of Expert Judgments
6. Dependence

PART II Procedures

1. Introduction
2. Preparation for Elicitation
3. Elicitation
4. Post-Elicitation

APPENDIX I

Summary Results of the EC-USNRC Uncertainty Study

APPENDIX II

Reports published as a result of the joint ec/usnrc project on uncertainty analysis of probabilistic accident consequence codes (under the third ec-framework programme).

APPENDIX III

Glossary of terms for Uncertainty Analysis

APPENDIX IV

Glossary of terms for performance based combinations of expert judgments

APPENDIX V

Training material

REFERENCES

PART I: GENERIC ISSUES

1. WHAT IS UNCERTAINTY?

Uncertainty is that which disappears when we become 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 be distinguished from ambiguity. Ambiguity is removed by linguistic conventions regarding the meaning of words. The two notions of uncertainty and ambiguity become contaminated when observations are described in an ambiguous language. Much of the work of an uncertainty analyst consists of developing a sufficiently clear language for expressing the possible observations. We assume that it is always possible to reduce any given ambiguity to any desired level. It is impossible to remove all ambiguity, and the work of disambiguation goes on until the residual ambiguities are not worth the effort required to remove them.

To be studied quantitatively, uncertainty must be provided with a mathematical representation. The 'representation of uncertainty' has received much attention of late from groups which were not previously in contact and which do not share a common vocabulary or background. Without wishing to adjudicate the many theoretical questions raised by this recent activity, the following proposes a perspective from which practical work can proceed without closing the door to interesting theoretical questions.

The problem of representing uncertainty is similar to the problem of representing reasoning in logic. A logic adequate for the representation of mathematical reasoning was developed at the end of the 19th century. The first half of the 20th century witnessed a proliferation of "alternative logics". Some were specifically meant to capture reasoning in other domains, e.g. moral reasoning, temporal reasoning, and reasoning with possibility and necessity. Paradoxes in physics were used to motivate such innovations as additional truth values and non-Boolean logical operations. Some innovations have earned a permanent place in logic. Many more are gratefully forgotten. Through this development, a core theory emerges and forms the theme on which variations are given. This core theory is roughly the most convenient representation which is adequate for the class of problems most people deal with. The present core theory is not sacred; the marginal variation of today may develop into the core of tomorrow as the class of 'core problems' evolves. This picture of a core theory with variations remains valid as long as an active research community is focused on a common set of problems.

In the same spirit, we suggest that there is a core theory for the representation of uncertainty, namely as subjective probability. Every aspect of this representation has been challenged, relaxed or strengthened. These variations remain just that, variations on a theme. Within the subjective interpretation of probability, uncertainty is a degree of belief of one person, and can be measured by observing choice behavior. In well-defined circumstances, subjective probabilities will coincide with observed relative frequencies.

For more background and discussion on uncertainty, the reader is referred to (Granger Morgan and Henrion 1991, Cooke 1991, Hogarth 1987).

Perhaps one final remark is useful. Viewed from the theory of rational decision (Savage 1954) one subjective probability is as good as another. There is no rational mechanism for persuading individuals to adopt the same degrees of belief, just as there is no mechanism for persuading them to adopt the same preferences. When observations are made, however, degrees of belief will change and under appropriate circumstances will converge. Uncertainty analysis is not about getting people to agree about uncertainty, rather it explores the consequences of uncertainty with respect to quantitative models. That having been said, in practical uncertainty analysis, we are typically interested in the uncertainty of experts in relation to quantitative models. Hence structured expert judgment provides a key input to uncertainty analysis.

2. WHAT IS UNCERTAINTY ANALYSIS?

This chapter presents a first pass at uncertainty analysis, with the intention of orienting the reader in the variety of ideas, terminology and techniques that are applied in quantitative uncertainty analysis. A glossary of terms for uncertainty analysis is provided in Appendix III.

2.1 When is uncertainty analysis indicated?

Uncertainty analysis is indicated for a decision problem when the following features are present:

- Decision making is supported by quantitative models
- The modeling 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

A simple example illustrates these features. Suppose a firm is considering alternative investment programs. Quantitative models predict the profit resulting from each program as a function of factors like market share, price and production costs. If these factors were known with certainty and if the models were known to be correct, then there would be no uncertainty in the models' output, and no need for uncertainty analysis.

Even though the above factors are not known with certainty, it might be argued that the expected market share, expected price and expected production costs could be fed into the models to yield a "best estimate" of the profit. If the models were linear in these variables, then the "best estimate" would be the expected profit; in other cases the meaning of "best estimate" is unclear. If the models were indeed linear and if the firm's (dis)utilities were linear in profit, then this would provide a suitable basis for reaching a decision. The firm wouldn't care whether the expected profit were realized with certainty, or whether the expected profit was computed from a "favorable scenario" yielding high profits and an "unfavorable scenario" involving negative profits (losses). This may well be the case when the potential losses are not too heavy and the profits not too large. However, if the losses exceed a certain threshold the firm goes bankrupt. The disutilities of negative profit exhibit threshold behavior: losing twice as much is more than twice as bad if the solvency threshold is crossed. Even this is not enough to motivate an uncertainty analysis. It might be the case that the excessive losses arise from "acts of God" which would produce excessive losses no matter which of the proposed alternatives were adopted. The firm is then confronted with a risk which it cannot avoid; it needn't quantify this risk in order to decide between the alternatives at hand.

Suppose however, that different alternative entail different risks of excessive loss. Further rational deliberation of alternatives now requires identifying the possible scenarios, quantifying the probability the each scenario occurs, and running each scenario through the models, yielding a distribution of possible profits associated with each alterative investment program. This is uncertainty analysis.

Notice that uncertainty analysis does not solve the decision problem. The firm must still decide how to trade heavy loss off against high profits. It is clear, however, that uncertainty analysis is an essential ingredient in responsible decision making when the features listed above apply. In many contexts we are not dealing with a full decision problem. For example, a regulatory authority is typically charged with regulating the risks from one type of activity. The choice between alternatives is made at a different level, where the trade-off of utilities against disutilities of different stake holders is factored in. It is nonetheless incumbent upon the regulatory authority to provide such information as is deemed necessary for responsible decision making.

For background on uncertainty analysis, consult (Baverstam et al 1993, Cooke 1995, Glaser et al 1994, Kalos and Whitlock 1986, Ripley, 1987. Ross, 1990, Rubenstein 1981, Dagpunar 1988, de Ruyter van Steveninck 1994).

2.2 How is Uncertainty Analysis Performed?

An uncertainty analysis is performed with respect to a given quantitative model. It may be broken into three steps:

- Assessing uncertainty over the model input
- Propagating uncertainty through the model

- Communicating results to decision makers.

Assessing uncertainty over model input

Sophisticated physical models, such as those met with in accident consequence modeling, contain a large number of parameters. Few of these are known with certainty. It is seldom feasible to devote full resources to quantify the uncertainty on all parameters. Typically the parameters are divided into two groups. The first group contains those parameters whose uncertainty is thought to have a large impact on the uncertainty of the model output. The second group contains those variables whose uncertainty does not have significant impact. Methods for defining the significant variables are discussed in (Iman et al 1981, Iman and Helton 1988, McKay 1988), and will not be pursued here.

Once a set of potentially important parameters has been identified, uncertainty over these parameters must be quantified. This is done by assigning (joint) probability distributions to the parameters. When sufficient statistical data is available, these are used to generated distributions. When statistical data is not available, uncertainty is quantified via structured expert judgment. The following sections provide an overview of expert judgment methods.

Propagating uncertainty through the model

When probability distributions have been obtained on the model parameters, these distributions must be "pushed through" the model. In very simple cases this can be done analytically. In practice the models to be analyzed are much too complex to allow analytic propagation. For complex models, the distribution of model output is determined by simulation. To perform one simulation run, a value for each parameter is drawn from the appropriate uncertainty distribution, the model is run with these parameter values and the result is stored. This is repeated many times until a distribution over model output is built up.

Communicating results to decision makers

Complex models compute many quantities of interest, or "endpoints". In accident consequence modeling endpoints might include, acute and chronic fatalities, area of land denial, number of persons evacuated, economic damages, etc. In fact, the accident consequence codes analyzed in (Harper et al 1995) compute, on each run, distributions arising from uncertainty in meteorological conditions at the time of a hypothetical accident. In performing an uncertainty analysis, distributions over model input parameters generate distributions over the meteorologically induced distributions for each endpoint. The question how best to compress all this information in a form which can be used by decision makers deserves careful attention. It is often important to distinguish different 'kinds' of uncertainty. Usually, this is done to distinguish those uncertainties which we might reduce with future research, and those uncertainties which we cannot reduce.

Many authors distinguish "objective" or "statistical" uncertainty from "subjective" uncertainty. In this context, statistical uncertainty refers to natural variation incorporated into the model endpoint calculations. In other words, the models compute statistical distributions of endpoints based on measured natural variation. The most familiar example is natural variation due to weather. The consequence codes contain a meteorological file reflecting different possible weather scenarios. In running the code, the endpoints are computed for each weather scenario and the distribution of endpoint values is output, which is taken to reflect natural or statistical variation in endpoint values due to weather. The computations of these distributions call parameters whose values are in fact uncertain, even though they are represented by a single number in the code. In such cases we speak of (subjective) uncertainty over statistical distributions. We perform an uncertainty analysis; we repeatedly sample these parameters from the subjective uncertainty distributions and compute a large number of statistical distributions. In general the distinction of objective and subjective aspects is often less straightforward than the above example would suggest.

Consider for example deposition velocity. The accident consequence codes have one deposition velocity for a given contaminant to 'pasture'. On the other hand it is known that deposition velocities may vary over an order of magnitude according to the species of grass. We may distinguish three cases:

(i) We believe that the pasture in the vicinity of our nuclear power plant consisted of one species of grass, but we don't know which. A distribution over this deposition velocity should then reflect our subjective uncertainty with regard to the species of grass.

(ii) We believe that the pastures have different kinds of grass. In this case, if we religiously distinguish objective and subjective uncertainties, we should incorporate this 'natural variation' into the model output. If we are uncertain how exactly the grass species are distributed over the pastureland, we should quantify our uncertainty over which natural variation is really the case, perform an uncertainty analysis and output a set of possible 'natural variations' (and we should do the same for all other contaminants and all other surfaces),

(iii) We are not sure which of the above situations is true. Then if we really must distinguish subjective from objective uncertainty, we should quantify our uncertainty in each possibility, and perform both analyses.

Case (i) conforms to current practice, but this is more a matter of convenience than conviction.

Mathematically, of course, there is no need to distinguish "objective" and "subjective" uncertainties. Indeed, there is ultimately one aggregated uncertainty measure. We can represent this measure as an integral over sets of conditional measures, where we conditionalize on possible values of some uncertain quantity. This is what we do in separating "objective" and "subjective" uncertainties: conditional on the values of all subjectively uncertain quantities, the remaining uncertainty is (by definition) objective. The question is, why do we separate some 'objective' uncertainties and not others? In fact, this decision is taken by the code developers and not by the uncertainty analyst, since the code developers decide which natural variations to incorporate in the code output. The uncertainty analyst is not empowered to alter the code on which the uncertainty analysis performed. The role of the uncertainty analyst is define clearly the various uncertainties and indicate which might be reduced by various means. The decision makers may also wish to see which of the input variables contribute most to the uncertainty of various endpoints, and which variables involve correlated uncertainties. Techniques for defining "importance" in the context of an uncertainty analysis are developing rapidly. The reader is referred to a forthcoming book (Saltelli, appearing) and a special issue of Computer Physics Communications (1999) devoted to these subjects.

3 STRUCTURED EXPERT JUDGMENT

Expert judgment has always played a large role in science and engineering. Increasingly, expert judgment is recognized as just another type of scientific data, and methods are developed for treating it as such. This section gives a brief overview of methods for utilizing expert judgment in a structured manner. For more complete summaries see (Hogarth 1987, Granger Morgan and Henrion 1990, Cooke, 1991).

In this section, the subject is broken down according to the form in which expert judgment is cast. A final sub-section addresses conditionalization and dependence. In all cases, the judgments of more than one expert are elicited. The questions of measuring performance of experts and combining their judgments are addressed more fully in succeeding sections. Mathematical terms used in this section are defined in Appendix IV.

3.1 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 in 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 judgment 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 judgments 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.

3.2 Paired comparisons

In the paired comparison method, experts are asked to rank alternatives pairwise 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 judgment data. Various processing methods are proposed for distilling a rank order from the pairwise 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). 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.

3.3 Discrete event probabilities

An uncertain event is one which either occurs or does not occur, though we don't know which. The archetypical 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 which 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).

3.4 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 (see Appendix III) 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, say 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-r-th quantile is the smallest number x_r such that the expert's probability for the event $\{X < x\}$ 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. This is treated further in section 4.2 and Appendix III.

3.5 Conditionalization and dependence

When expert judgment is cast in the form of distributions of uncertain quantities, the issues of conditionalization 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" (see Part II). Failure to specify background information can lead experts to conditionalize 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 Part II.

4 PERFORMANCE MEASURES

The measures of performance discussed here 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. These methods are developed in (Cooke et al 1988, Cooke 1991) and applied in a wide variety of studies, including recent EU-USNRC uncertainty study of accident consequence codes (Harper et al 1994). They require that experts assess variables whose values become known to the experts post hoc. These variables are termed "performance variables" or "calibration variables" or "seed variables", and they form a distinguishing feature of the present methods.

The variables of interest for an uncertainty analysis are typically not of such a nature that their true values become known within the time frame of the study. In this case additional variables must be introduced. The identification of appropriate seed variables is a major task of the uncertainty analyst. Seed variables may sometimes have the same physical dimensions as the variables of interest. This arises when the variables of interest are not practically measurable for reasons of scale, e.g. great distances, long times, high temperatures; whereas measurements can be performed at other scales. In this case, unpublished measurements or experiments can be used as seed variables. When such seed variables are not available, variables can be chosen which "draw on the relevant expertise" yet do not have the same dimensions as the variables of interest. As a loose criterion, a seed variable should be a variable for which the expert may be expected to make an educated guess, even if it does not fall directly within the field of the study at hand. Seed variables should be chosen so as to avoid dependencies so far as possible. The number of seed variables for assessments of uncertain quantities with continuous ranges is typically between 10 and 30.

The purpose of performance is twofold. First, verifying good performance of the experts and of the combination of experts' judgments enhances the credibility of the study and helps build rational consensus. Second, performance measures can be used to construct performance based combinations of expert judgments. The latter subject is treated in the following section. To facilitate discussion, we assume that experts have assessed 5%, 50% and 95% quantiles for a number of seed variables.

Performance is measured in two dimensions, namely calibration and informativeness. Performance is discussed from the viewpoint of a decision maker who will use the experts' assessments in his decision problem. Details on these notions are found in Appendix IV, a qualitative discussion is given below.

4.1 Calibration

Calibration measures statistical likelihood, very loosely characterized as "correspondence with reality". In scoring calibration, each expert is regarded as a statistical hypothesis, namely:

The realizations of the seed variables may be regarded as independent samples from a distribution corresponding to the expert's quantile assessments.

The "distribution corresponding to the expert's quantile assessments" is the distribution which says, in effect, that the probability of a realization falling in an inter-quantile range is just the difference of the corresponding quantiles. Thus, the probability that a realization falls between the 50% quantile and the 5% quantile is 45%. The decision maker wants experts for whom the corresponding statistical hypothesis is well supported by the data gleaned from the seed variables. This is sometimes expressed as 'the decision maker wants probabilistic assessments which correspond to reality'.

We may sketch the matter very crudely as follows. If an expert gives 90% confidence bands for a large number of variables, then we might expect that only 10% of the variables will actually fall outside his bands. If the expert has assessed 20 variables for which the realizations are known post hoc, then 3 or 4 of the 20 variables falling outside these bands would be no cause for alarm, as this can be interpreted as sampling fluctuations. The above hypothesis would still be reasonably supported by the data. If 10 of the 20 variables fell outside the bands, we should be worried, as it is difficult to believe that so many outliers should result from fluctuations; we should rather suspect that the expert chooses his bands too narrowly. Statistical likelihood measures the degree to which data support the corresponding statistical hypothesis. More precisely:

A calibration score of 0.01 for an expert based on 20 realizations means that there is a 1/100 chance of seeing a discrepancy between observed and predicted frequencies as great or greater than that observed for this expert on these 20 realizations.

The calibration score is always between zero and one, and higher scores are better. As is evident from the above, the proper interpretation of a calibration score requires knowledge of the number of realizations on which it is based, and scores based on different numbers of realizations cannot be compared directly. The can be compared by rescaling the 'effective number' of seed variables.

Calibration is a fairly "fast" function; that is, on the basis of say 10 realizations we can easily distinguish four or more orders of magnitude in calibration. Empirical scientists are frequently puzzled by this on first acquaintance, so it is worthwhile to offer a brief explanation.

Suppose we toss a coin ten times and observe six heads. Expert Nr. 1 has assessed the probability of heads on each toss as 1/10. The statistical likelihood of the corresponding statistical hypothesis is the probability of observing six or more heads on ten tosses, if the probability per toss is 1/10/ This likelihood is about 0.0001. Expert Nr. 2 has assessed the probability of heads per toss as 1/2. His statistical likelihood is about 0.38.

The speed of the calibration function entails that adding or deleting a realization can have a noticeable impact on the calibration score of the experts. The performance based decision makers discussed in the next section depend on the ratio of expert calibration scores, and this is considerably less sensitive. Experience with many data sets supports the following rule of thumb: If there are 20 realizations, then some of these when removed individually may change the calibration score of an expert (or the decision maker) by a factor 2. This typically has a small impact on the actual distributions of the performance based decision maker. However, an analysis of robustness against removal of seed variables should always be carried out.

4.2 Informativeness

To measure informativeness, a background measure is assigned to each query variable. Probability densities are associated with the assessments of each expert for each query variable in such a way that (i) the densities agree with the expert's quantile assessments, and (ii) the densities are minimally informative with respect to the background measure, given the quantile constraints (Kullback 1959). The background measures are typically either uniform or loguniform. For these background measures, it is necessary to choose an "intrinsic range", i.e. to truncate the range of possible values. The resulting information scores can be shown to be very robust against the choice of the intrinsic range (Cooke 1991). In practice, the intrinsic range is chosen by the "10% overshoot rule": the smallest interval containing all quantile assessments is extended by 10% above and below. Information scores on different sets of variables can be compared only if the background measures and intrinsic ranges agree setwise.

Informativeness is scored per variable per expert by computing the relative information of the expert's density for that variable with respect to the background measure. Overall informativeness per expert is the average of the information scores over all variables. This average is proportional to the relative information in the expert's joint distribution over all variables under the assumption that the variables are independent. Information scores are always positive, and other things being equal, experts with high information scores are preferred. The information score is a positive number with increasing values indicating greater information relative to the background measure. Since the intrinsic range depends on the expert assessments, this range can change as experts are added or deleted, and this can exert a small influence on the information scores of the remaining experts.

Information is a "slow" function; that is, large changes in the quantile assessments produce only modest changes in the information score. On a data set with 20 realizations and five experts, calibration scores typically vary over four orders of magnitude, but information scores seldom vary by more than a factor 3. In the performance based combinations discussed below, this feature prevents a poor calibration performance being compensated by a very high information score. Information serves to modulate between experts who are more or less equally well calibrated.

4.3 Expert Performance

It is natural to ask how experts perform, and whether the uncertainty of an expert in matters relating to his/her expertise is better, in the sense of calibration and information, than the uncertainty of non-experts.

There is not an abundance of literature on this issue. However one psychometric experiment compared “experienced experts” (teachers at a technical training institution) with “inexperienced experts” (students at said institution) on two types of items, namely technical items and general knowledge items. On technical items the experienced experts performed significantly better than the inexperienced experts, with regard to both calibration and information. On the general knowledge items there was no significant difference in either regard.

An extended study of weather forecasting in The Netherlands (Murphy and Daan, 1982, 1984) was reanalyzed in (Roeleven et al 1991). Experts predicted precipitation, visibility and wind speed for 5 six hour periods into the future, over a period of six years. Not surprisingly, it emerged that experts’ distributions were less informative as the variables predicted were further into the future; although the loss of information was not great. Somewhat less obvious was the result that experts’ distributions were much better calibrated for variables up to 12 hours in the future, than for variables further into the future. The experts’ scientific modeling tools are of course better for near future variables. This suggests that that the experts’ scientific expertise contributes to producing well calibrated assessments.

A study of Dutch project managers assessing the probability that project proposals (Bhola et al 1991) showed that younger experts were better probability assessors than their older colleagues, though other explanatory variables than experience may play a decisive role. Projects in The Netherlands are more predictable than projects abroad, and younger project leaders have proportionally more projects in The Netherlands.

5. COMBINATIONS OF EXPERT JUDGMENT

Decision makers want to take, and want to be perceived to take, 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 involved, 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.

Rational consensus imposes certain constraints which a method of combining expert judgments must satisfy. These constraints are expressed as principles and constitute necessary conditions for rational consensus. If a method of combination violated one of these principles, then a rational stakeholder would have sufficient grounds for withholding prior commitment.

5.1 Principles for Rational Consensus

The principles forming the basis of the performance based methods are (Goossens et al 1989, Cooke 1991):

- 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.
- Fairness: experts are not pre-judged.
- Neutrality: methods of elicitation and processing must not bias results.
- Empirical control: quantitative assessments are subjected to empirical quality controls.

When expert judgments are cast in the form of probability distributions, these distributions must be combined, for each assessed variable, to yield distributions for a "decision maker". Many mathematical functions for this purpose have been proposed and studied. This subject is not reviewed here. Suffice to say that strong arguments can be given for restricting attention to "linear pooling", that is taking weighted averages of the experts' distributions, where the weights are non-negative and sum to one (French, 1985). The problem of combining experts' distributions is then reduced to the problem of determining the weights.

A variety of methods have been used in practice. The simplest method, which is always preferred in the absence of something better, is to take all weights to be equal. While equal weight combinations have an obvious appeal, they also have drawbacks. One expert whose distributions differ strongly from the rest can have a large impact on the resulting decision maker. This is a drawback if this expert's assessments cannot be defended on the basis of performance. As more and more experts are brought into the study, the equal weight decision maker can tend to become quite diffuse.

In cases where performance has been measured by the methods sketched above, a general conclusion is that performance based decision makers outperform the equal weight decision maker. The difference is sometimes marginal, often mild, but sometimes severe (Goossens et al. 1998). We discuss briefly the principles of performance based combinations of expert judgments. Details can be found in Appendix IV.

5.2 Proper scoring rules

In developing combinations of expert judgments, the principle of neutrality is of particular interest. This principle says that the method of combination must not bias the assessments. Put differently, the method of combination should not reward experts from giving an assessment at variance with their true opinion. Implementing the principle of neutrality leads naturally to the theory of strictly proper scoring rules. In the simplest case, a scoring rule is a function assigning a real number to an assessed distribution plus a realization. A scoring rule is strictly proper if, whatever an assessors true distribution, his maximal expected score (computed before the realization is known) is obtained when his assessment coincides with his true distribution.

The idea is best explained by considering a popular improper scoring rule. Suppose an uncertain quantity X can assume one of three possible values {1,2,3}. Let $P = (p_1, p_2, p_3)$ be an assessor's true probability distribution for outcomes 1,2,3. A scoring rule is a function R(P,i) where i is the realized value.

Consider the so-called direct rule:

$R(P,i) = p_i$;

When the expert believes P, yet reports distribution Q as his assessment, his expected score is

$E_P(R(Q,X) = q_1p_1 + q_2p_2 + q_3p_3$.

If he decides to choose Q so as to maximize his expected score, what should he choose? Suppose $p_1 > p_2$; $p_1 > p_3$; then it is easy to see that expected score is maximized by choosing $q_1 = 1$, $q_2 = q_3 = 0$.

R is strictly proper if for all P, the expected score $E_P(R(Q,X)$ is maximized if and only if Q = P. This type of scoring rule is a function of a single assessment plus realization. In combining expert judgments we must assign scores to sets of assessments and sets of realizations, and we are interested in the long run behavior of scoring rules as the number of variables gets large. This complicates matters considerably. Suffice to say that the product of the calibration and information scores introduced above is a strictly proper scoring rule in an appropriate long run sense.

5.3 Combinations

Experts give their uncertainty assessments on query variables in the form of, say, 5%, 50% and 95% quantiles. An important step is the combination of all experts assessments into one combined uncertainty assessment on each query variable. All combination schemes considered here are examples of "linear pooling"; that is the combined distributions are weighted sums of the individual experts' distributions, with non-negative weights adding to one. Different combination schemes are distinguished by the method according to which the weights are assigned to densities. These schemes are designated "decision makers". Three decision makers are described briefly below.

Equal weight decision maker

The equal weight decision maker results by assigning equal weight to each density. 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 density is given by:

$f_{eqdm,i} = (1/N) \Sigma_{j=1\ldots N} f_{j,i}$

where fj,i is the density associated with expert j's assessment for variable i.

Global weight decision maker

The global weight decision maker uses performance based weights which are determined, per expert, by the normalized product of expert's calibration score and his(her) overall information score, and by an optimization routine described below. The calibration score is determined per expert by his(her) assessments of the seed variables. The overall information score is the (relative) information in the expert's joint distribution. For expert j, the same weight is used for all variables assessed. Hence, for variable i the global weight decision maker's density is:

$f_{gwdm,i} = \Sigma_{j=1\ldots N} w_j \, fj_{,i}$; $\Sigma_{j=1\ldots N} w_j = 1$.

These weights satisfy a "proper scoring rule" constraint. That is, under suitable assumptions, an expert achieves his(her) maximal expected weight, in the long run, by and only by stating quantiles which correspond to his(her) true beliefs.

Item weight decision maker

As with global weights, item weights are performance based weights which satisfy a proper scoring rule constraint, and are based on calibration and informativeness, with an optimization routine described below. Whereas global weights use an overall measure of informativeness, item weights are determined per expert and per variable as the product of calibration and information for the given item.This enables an expert to up- or down-weight him(her)self for each variable by choosing a more or less informative distribution for that variable. Roughly speaking, more informative distributions are gotten by choosing quantiles which are closer together whereas less informative distributions result when the quantiles are farther apart. For the item weight decision maker, the weights depend on the expert and on the item. Hence, the item weight decision maker's density for variable i is:

$f_{iwdm,i} = \Sigma_{j=1\ldots N} w_{j,i} \, f_{j,i}$; $\Sigma_{j=1\ldots N} w_{j,i} = 1$.

Optimization

The proper scoring rule constraint entails that an expert should be unweighted if his/her calibration score falls below a certain minimum, $\alpha > 0$. The value of α is determined by optimization. That is, for each possible value of α a certain number of experts will be unweighted, and the weights of the remaining experts will be normalized to sum to unity. For each value of α a decision maker dm_α is computed. Dm_α is scored with respect to calibration and information, and the "virtual weight" of dm_α is computed, this is the weight which this dm would receive if he were added as a 'virtual expert'. The value of α for which the virtual weight of dm_α is greatest is chosen as the cut-off value for determining the unweighted expert. For more details see Appendix IV.

The appeal to first principles, even to secondary and tertiary principles, cannot lead to a unique mathematical model for combining expert judgment. Ad hoc choices must still be made. One example of such a choice is the selection of the intrinsic range discussed above. Another such choice relates to the parameters "size" and "power" in selecting statistical tests, see Appendix IV.

6. Dependence

It has long been known that ignoring dependencies between uncertainties (Apostolakis and Kaplan 1981) can cause significant errors in uncertainty analysis. New techniques for estimating and analyzing dependencies in uncertainty analysis have been developed in the course of the Joint EU-NRCaccident consequence uncertainty analysis. These are described here. For the mathematics of dependence modeling see Cooke (1995) and the references therein. Post-processing (Kraan and Cooke 1996) also represents a way of assessing dependencies, but this not discussed here. We discuss how to elicit dependencies from experts, how to combine them, and how to analyse dependencies in the output of an uncertainty analysis.

6.1 Lumpy and smooth elicitation strategies

The best source of information about dependencies is often the experts themselves. The most thorough approach would be to elicit directly the experts' joint distributions. The practical drawbacks to this approach have forced analysts to look for other dependence elicitation strategies. One obvious strategy is to ask experts directly to assess a (rank) correlation coefficient. Even trained statisticians have difficulty with this type of assessment task (Gokhale and Press 1982).

Two approaches have been found to work satisfactorily in practice. The choice between these depends on whether the dependence is lumpy or smooth. Consider uncertain quantities X and Y. If Y has the effect of switching various processes on or off which influence X, then the dependence of X on Y is called lumpy. In this case the best strategy is to elicit conditional distributions for X given the switching values of Y, and to elicit the probabilities for Y (Krzykacz and Hofer 1988). This might arise, for example, if corrosion rates for under ground pipes are known to depend on soil type, where the soil type itself is uncertain. In other cases the dependence may be smooth. For example, uncertainties in the biological half lives of cesium in dairy cows and beef cattle are likely to be smoothly dependent.

Within the joint EU-NRC study a strategy has been employed for eliciting smooth dependencies from experts. When the analyst suspects a potential smooth dependence between (continuous) variables X and Y, experts first assess their marginal distributions for X and Y. They are then asked:

Suppose Y were observed in a given case and its value were found to lie above y_{50}, the median value for Y; what is your probability that, in this same case, X would also lie above x_{50}, the median value for X?

Experts quickly became comfortable with this assessment technique and provided answers which were meaningful to them and to the project staff. If F_x and F_y are the (continuous inevitable) cumulative distribution functions (cdf's) of X and Y respectively, the experts thus assess

$P_{50}(X,Y) = P(F_x(X) > 0.50 \mid F_y(Y) > 0.50)$.

Consider all joint distributions for having marginals, F_x, F_y, having minimum information relative to the distribution with independent marginals F_x , F_y, and having rank correlation *rnk*(X,Y), *rnk* $\in$ [-1, 1]. For each such minimum information distribution there is a unique value for $P_{50}(X,Y)$, and conversely, each value of $P_{50}(X,Y) \in [0,1]$ is associated with a unique minimum information distribution with a rank correlation $\in$ [-1,1].

More generally, for each *rnk* $\in$ [-1, 1], and each r $\in$ (0,1) we may associate the number

$H(rnk, r) = P(F_x(X) > r \mid Fy(Y) > r)$;

where the probability P is taken from the minimum information distribution with marginals F_x, F_y, and rank correlation *rnk*.

Holding *rnk* fixed, we get a functions of r, r $\in$ [0,1]. These functions have been computed with the uncertainty analysis package UNICORN (Cooke 1995). The results for *rnk* = -0.90, -0.80, …0.90 are shown in Figure 1.

When there is a single expert having assessed $P_{50}(X,Y)$, then we simply use the minimum information joint distribution with rank correlation *rnk* solving:

$H(rnk, 0.50) = P(F_x(X) > 0.50 \mid F_y(Y) > 0.50)$

from Figure 1.

When several experts are combined via linear pooling, a complication arises. Since the medians for X and Y will not be the same for all experts, the conditional probabilities $P_{50}(X,Y)$ cannot be combined via the linear pooling. However, the marginal distributions can be pooled, resulting in cdfs $F_{x,DM}$ and $F_{y,DM}$ for X and Y for the Decision Maker (DM). Let x* and y* denote the medians for DM's distribution for X and Y. With each expert we associate a minimum information joint distribution; for each such distribution we can compute the conditional probabilities $P(X>x^*|Y>y^*)$. Since these conditional probabilities are defined over the same events for all experts, they can be combined via the linear pool. This yields a value for p_{50} for DM, for which we can find the corresponding rank correlation.

A simple graphical method for doing this can be given if for each experts the numbers $F_x(x^*)$ and $F_y(y^*)$ are not too different. In this case we associate with each expert a number $r_e = (F_x(x^*)+F_y(y^*))/2$. We then use this value of r_e together with the value of rnk_e determined from the experts response p_{50} to read the value of $P(X>x^*|Y>y^*)$ from Figure 1. The steps are summarized as follows.

1) For each expert e query $P_{50,e}(X,Y) = P(F_x(X) > 0.50 \mid F_y(Y) > 0.50)$
2) For each expert e, find rnk_e which solves $H(rnk_e, 0.50) = P_{50,e}(X,Y)$ (using Fig. 1)
3) Take linear pooling of experts' marginals, find DM's medians x*, y*
4) For each expert find $r_e = (F_y(y^*)+F_x(x^*))/2$, and $P_{r,e}$, using rnk_e from Fig. 1
5) Using linear pooling, define $P_{DM} = \Sigma_{j=1\ldots N} w_j P_{r,e} = P_{DM}(F_{x,DM}(X) > x^* \mid F_{y,DM}(Y) > y^*)$
6) Find *rnk* from Figure 1 which solves $H(rnk, 0.50) = P_{DM}(X,Y)$ (using Fig. 1)

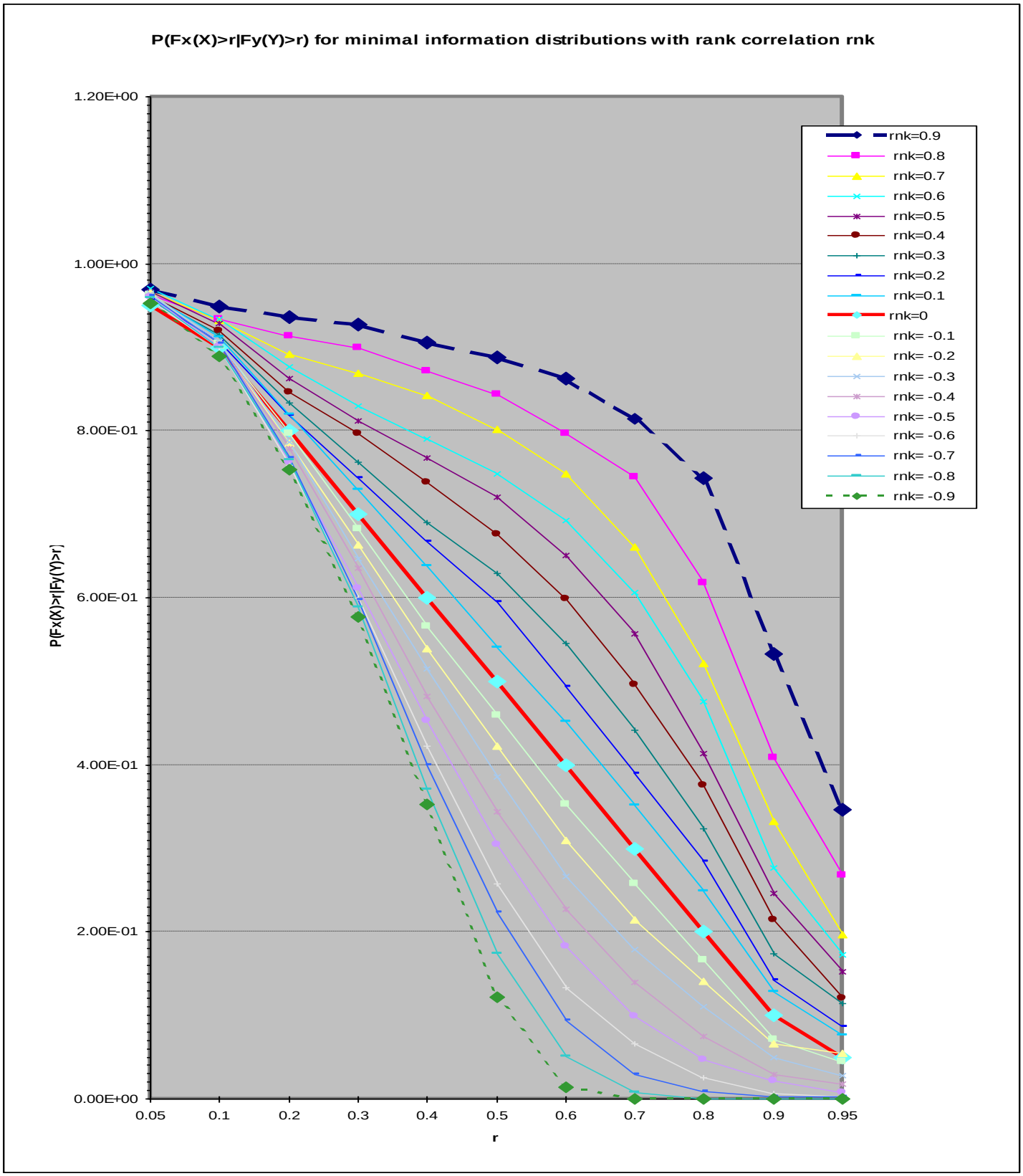


Figure 1. p_r for minimum information distributions with rank correlation constraints.

6.2 Analyzing dependence in output with cobweb plots

The problem of understanding and communicating information about dependencies in the results of an uncertainty analysis is inverse, as it were, to the problem of eliciting dependencies. If the dependencies are smooth then familiar regression, rank regression and partial rank correlation measures can convey useful information about how the uncertainty in output variables depends on uncertainty in input variables. Unfortunately, such measures may easily miss important information. Using only such measures, one never knows whether something is being missed.

For this reason a graphical tool, called cobweb plotting has been developed to study dependence. To form a cobweb plot, selected variables are represented as parallel vertical lines on which the variables' percentiles have been marked. Each sample from an uncertainty analysis is mapped as a jagged line intersecting the vertical lines at the percentile points realized in that sample. The result of plotting a few hundred samples suggest a cobweb, and contains all the information in the empirical joint distribution of the variables quantile functions. To study and extract information from a cobweb plot, the user can filter all lines passing through selected intervals on the vertical lines. This is equivalent to conditionalizing the joint distribution on percentile intervals of the variables.

Figure 2 shows an unconditional cobweb for variables dispersion and deposition variables for stability class E. The left most variable is dry deposition velocity. The next four variables are dispersion coefficients from the power laws:

$\sigma_y(x) = a_y x^{by}$;

$\sigma_z(x) = a_z x^{bz}$.

The remaining variables are ground concentrations at various down wind distances. “|x,y” indicates the ground concentration at down wind distance x and cross wind distance y from the centerline.

From Figure 2 we can see that the concentrations tend overall to be rather strongly correlated with each other. However, this overall pattern need not hold if we look at very high or low concentrations at specific distances. Figure 3 shows the lowest 5% of the concentrations at half a kilometer from the source at the centerline.

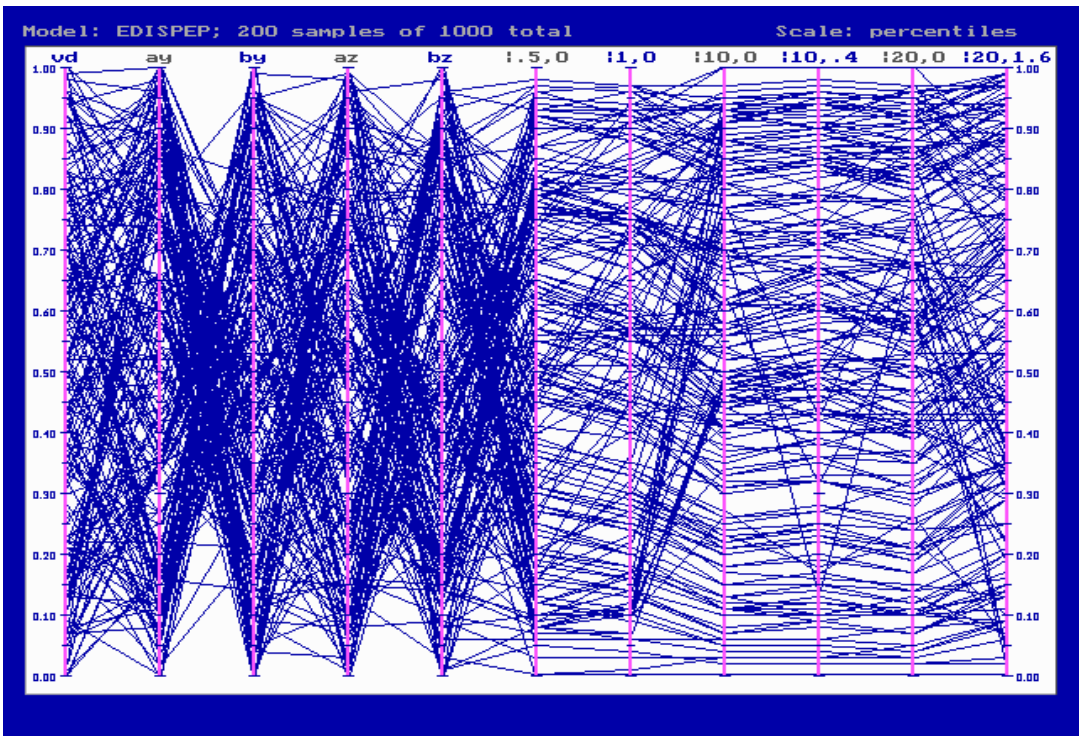


Figure 2, Unconditional cobweb plot, ground concentrations, stability class E

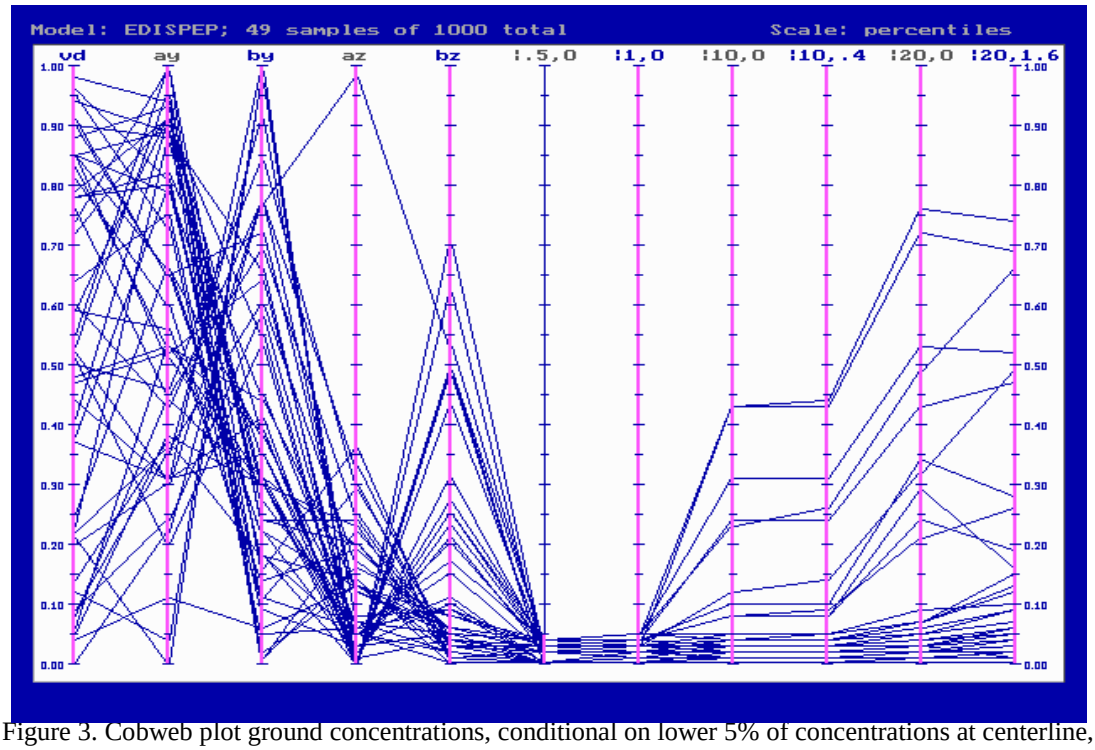


Figure 3. Cobweb plot ground concentrations, conditional on lower 5% of concentrations at centerline, ½ km downwind.

PART II: PROCEDURES

1. INTRODUCTION

Part II of this Procedures Guide provides the details of the protocol for a full expert judgment exercise. This protocol is based in large measure on experience gained within the joint EU-USNRC research effort, and reflects contributions from both the European and US experience (Hora and Iman 1988). Of course, the method itself is applicable outside the nuclear sector. In the general field of risk analysis the following studies performed by or in collaboration with the T.U. Delft have also contributed to the experience base for structured expert judgment.

- Crane Risk (DSM in collaboration with TU Delft, Akkermans 1989)
- Space Debris (TU Delft for the European Space Agency, Meima 1990)
- Safety Analysis Composite Materials (European Space Agency in collaboration with TU Delft, Offerman 1990)
- Groundwater Transport (DSM chemical plant in collaboration with TU Delft, Claessens, 1990)
- Dose Response Relations for Hazardous Substances (TU Delft for Dutch Ministry of Environment, Goossens et al 1992)
- Water Pollution (TU Delft for Dutch Min. of Environment, VROM 1994)
- Failure of Moveable Water Barriers (Dutch Ministery of Water Management in collaboration with TU Delft, van Elst 1997)
- Safety Factors for Airline Pilots (Aspinall and Associates for British Air, Aspinall 1996)
- Expert Judgment at Montserrat (Aspinall and Associates for governor Montserrat, Aspinall 1996)
- Expert Judgment for Serviceability Limit States (Ter Haar et al 1998).
- Expert Judgment for Uncertainty Analysis of Inundation Probability (in Dutch) (Frijters et al 1999).

For a review of recent applications and references to literature, see (Goossens et al 1998).

This protocol is broken into three sections, Preparation for elicitation, Elicitation and Post Elicitation. Material for training is collected in Appendix V. In the EU-NRC joint project, the uncertainty analysis of probabilistic accident consequence codes was broken down into the following sub models. For each individual sub model the following steps were followed.

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) Feed back
(14) Post-processing analyses
(15) Documentation

Steps (1) to (7) may contain several iterations prior to proceeding with step (8). Each step will be discussed separately.

2. PREPARATION FOR ELICITATION

Expert judgment is used to obtain results from experiments and/or measurements, which are physically possible, but not performable in practice. Such experiments are 'out of scale' financially, morally, or physically in terms of time, energy, distance, etc. they may be compared to thought experiments in physics. Since these experiments cannot in fact be performed, experts are uncertain about the outcomes, and this uncertainty is quantified in a formal expert judgment exercise. Preparation for elicitation is really nothing more than carefully designing these hypothetical experiments, so as to obtain the information that we require.

2.1. Definition of case structure

Defining the case structure for an expert judgment study involves studying the physical/biological/environmental models for whose quantification expert judgment is to be employed. The parameters in the models have of course been assigned values, but these values may be uncertain. The activity of the analyst in this phase may be described as answering the following questions:

Which values are uncertain?

The first job of the analyst is to retrieve the basis for the assignment of values to parameters, and to determine which values are uncertain. If the model documentation is not sufficient for this purpose, the model builders and source material must be consulted.

Can the uncertainty be quantified by historical and/or measurement data?

If historical and/or measurement is available to quantify uncertainty, then of course this should be preferred to expert judgment. Such data must be compiled under conditions which agree with the assumptions which the model builders have made. For example, accident consequence codes use parameters to describe plume spread under given atmospheric stability conditions. If measurement data is to be employed to quantify these parameters, the measurements must use the same stability classification scheme.

When the analyst concludes that sufficient historical and/or measurement data is not available for quantifying uncertainty, he is justified in applying expert judgment to this end.

Which (hypothetical) measurements would be used to quantify the parameters?

If a parameter is uncertain, and if the uncertainty cannot be quantified by data, then the analyst must ask how the values *would* be determined if suitable measurements could be preformed. These experiments will be hypothetical, i.e. they cannot be preformed in practice, but they must be physically possible. Again, care must be taken to insure that the hypothetical measurement conditions agree with the assumptions of the models in question.

The results of this activity should be written up in a case structure document. This document is distributed to the experts in the expert training (see paragraph 2.9)

2.2. Identification of target variables

Once the case structure of a specific model is defined, the target variables over which uncertainty distributions are required must be identified. Target variables are those parameters of the model in question which satisfy three criteria:

1. The values of the parameters are uncertain.
2. The uncertainty cannot be quantified with historical and/or measurement data.
3. The uncertainty is expected to have a significant impact on the uncertainty of one or more endpoints of the model.

The third criterion is applied when the number of variables satisfying the first two criteria is very large and exceeds the resources of the study. A formal method for implementing the third criterion is not available. In the joint EU-USNRC study, variables were selected on the basis of past experience and computing resources. In deciding to apply

expert judgment to assess the uncertainty concerning target variables, it is incumbent upon the analyst to document the available historical data and indicate his reasons for not using this data. Even when existing historical data is not used, it is important to document this data as a type of check for the results of expert assessments.

2.3. Identifying query variables

Variables which will be presented to the experts in the elicitation and for which they will quantify their uncertainty are termed *query variables*. We assume that the target variables to be quantified via structured expert judgment have been identified. The question is how to obtain uncertainty distributions over these variables. In defining the variables about which experts will be asked, two golden rules apply:

1. Ask for values of observable or potentially observable quantities.
2. Formulate questions in a manner consistent with the way in which an expert represents the relevant information in his knowledge base.

If a target variable satisfies the above two rules, then experts can be asked directly to quantify their uncertainty with regard to this variable. In this case the target variable is a query variable. If a target variable does not satisfy these requirements, then experts cannot be asked directly about this variable, and other variable(s) must be found which do satisfy these requirements and from which the required information can be extracted. We can best illustrate with some examples.

Deposition velocities:

In some cases the model parameters correspond to physically measurable quantities with which the experts are familiar. For example, deposition velocities to various surfaces under various conditions are directly measurable. The measured values are known to depend on a large number of physical parameters which cannot all be measured or controlled on any given experiment. Moreover, the functional form of the dependence is not known. Hence, if a controlled experiment is repeated many times, different values will be found reflecting different values of uncontrolled and unknown physical parameters. If a measurement set-up is described to an expert, (s)he can express his/her uncertainty via a subjective distribution over possible outcomes of the measurement. In such cases the experts are questioned directly about uncertainty with respect to model parameters.

In other cases experts cannot be questioned about model parameters as illustrated in the following two examples.

Dispersion coefficients:

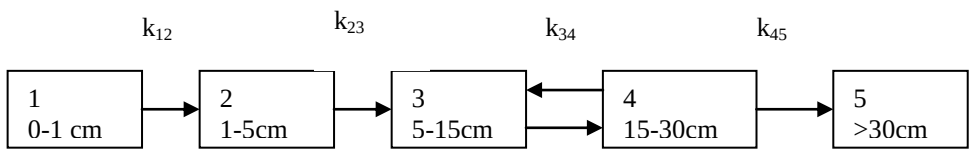
Lateral plume spread σ_y is modeled as a power law function of downwind distance x from the source of a release:

$$\sigma_y(x) = a_y x^{by}$$

where the dispersion coefficients a_y and b_y depend on the stability of the atmosphere at the time of the release. The above equation is not derived from underlying physical laws; rather, the coefficients are fit to data from tracer experiments. For the uncertainty analysis, we require distributions on a_y and b_y which, when pushed through the above equation, will yield the uncertainty on σ_y for each down wind distance x. a_y and b_y are target variables. Although the experts have experience with measured values of σ_y under various conditions, it is unrealistic to expect them to be able to quantify their uncertainty in terms of the target variables a_y and b_y. Indeed, the dimension of a_y must be [meters^{1-by}]. In this case we define *query variables* $\sigma_y(x_i)$; for down wind distances $x_1,...x_n$, and elicit uncertainty distributions on these. The problem then arises how to translate these elicited distributions into distributions on the target variables a_y and b_y. This type of problem is called probabilistic inversion, and is discussed in the paragraph on post-processing.

Transfer coefficients.

The migration of radioactive material through various depths of soil is modeled using a so-called box model, shown below.



k_{43}

Figure 4. Box model for soil transfer

The target variables in this case are the transfer coefficients k_{ij}, representing the proportion of material moved from box i to box j in a small time interval. The model in determines a set of first order differential equations which, under appropriate initial conditions, fully specifies the movement of material between the boxes.

The transfer coefficients in this model cannot be measured directly and therefore cannot be query variables. Query variables were chosen to be the times T_i at which half of an initial mass starting in box 1 at time 0 has passed beyond box i. From this information, a distribution on the transfer coefficients must be determined.

We see that in the second two examples above, the target variables cannot be query variables. The uncertainty analyst must choose query variables and must design methods for translating distributions over query variables into distributions on the target variables. It is essential to develop and test an appropriate post-processing technique before finalizing the choice of query variables (see paragraph 4.4).

Elicitation format for query variables

We assume that the query variables concern results of hypothetical experiments which take values in an effectively continuous range. For example, our query variable might be

The root mean square difference in time integrated concentration in the crosswind direction, at 10 km downwind from a unit airborne release.

The experts' uncertainty distributions may be represented by a number of quantiles; we assume that the 5%, 50% and 95% quantiles are elicited (see section 3.4 of Part I).

The elicitation will generally be cast as a hypothetical experiment. In describing this experiment it is important to identify the physical factors which may influence the outcome of the experiment. Each relevant physical factor will fall into one of two classes:

1. The case structure assumptions.
2. The uncertainty set.

Some relevant factors will have their values stipulated by the assumptions of the study, as reflected in the case structure. Thus models using simple straight line Gaussian dispersion models assume simple terrain. Further the models will read certain physical factors from the environment. In the above example, these might be the time of year, time of day, release height, degree of insolation, average hourly wind speed and average hourly wind direction. Hence these factors are assumed known and form part of the case structure assumptions.

Other factors may influence the outcome of the hypothetical experiment, but their values are *not* stipulated by the case structure. These factors belong to the uncertainty set. The experts should be made aware that these factors are uncertain, and should fold this uncertainty into their distributions on the outcome of the hypothetical experiment. Thus, the model may assume a simple classification of atmospheric stability involving the percentage of the sky covered by clouds. The case structure may then contain the assumption "insolation is 5/8" meaning that 5/8 of the sky is covered by cloulds. An expert may remark:

"yes, but are they high or low cloulds? that makes a big difference in the solar energy reaching the ground."

To such a question the analyst must answer

"This factor belongs to the uncertainty set for this experiment, and your uncertainty should take account of different effects of high versus low clouds, together with the probability of high versus low clouds".

A general format for elicitation may then be given as follows:

Conditional on

< values of factors in the case structure assumptions >

Please give the 5%, 50% and 95% quantiles of your uncertainty in

< Hypothetical experiment >

taking into account that values of

< uncertainty set >

are unknown.

Figure 5. Format for eliciting continuous varibles

Eliciting uncertainty on probabilities

Many models contain parameters which are probabilities, and which are also uncertain. In reasoning about 'uncertainty of probabilities' or 'uncertainty of uncertainty' conceptual muddles easily arise. It is therefore useful to treat this case separately.

A probability may be interpreted either as a limiting relative frequency, or as a degree of belief. In either case, a probability is not directly observable. Hence, the analyst should not ask ' What is your uncertainty on the probability of event X'; for a number of reasons:

1. It is ambiguous whether "probability" is to be interpreted objectively or subjectively.
2. If "probability" is interpreted subjectively, i.e. as degree of belief, then it is not meaningful to quantify uncertainty, as uncertainty concerns potential observations.
3. If "probability" is interpreted objectively, then it must be interpreted as limiting relative frequency, and the physical dimensions of frequency must be specified.

Probability is dimensionless. However probability may sometimes be interpreted as limiting relative frequency, and frequency has the dimensions: [#/units] (i.e. # per hour, per cycle, per mission, etc).

We can quantify uncertainty in relative frequency in large populations. To do this we basically follow the rules set forth above (see also Cooke and Jager, 1998, Frijters, 1999). First we specify a large virtual population; this might be the population of all pumps. Then we specify an experimental sub population selected randomly but so as to satisfy the case structure assumptions. For example, we select 1,000,000 pumps which are motor driven and with a given power rating. We now ask, how many of these pumps fail before 1,000 hours of continuous service. Operating environment, maintenance regime, ambient air temperature, for example, may belong to the uncertainty set.

The general format for "uncertainty over the probability of characteristic C" may be given as follows:

Conditional on

< N items satisfying in the case structure assumptions >

Please give the 5%, 50% and 95% quantiles of your uncertainty on

< number K of items exhibiting characteristic C >

taking into account that values of

< uncertainty set >

are unknown.

Figure 6. format for eliciting uncertainty on probabilities

Choosing N large, the distribution over K/N obtained in this way may be interpreted as the expert's uncertainty in the probability of C under the conditions specified in the case structure.

The output of the query variable identification activity is a list of query variables together which a schematically filled in elicitation format. The above formats are filled in telegraph style for each variable.

Dependencies

At this stage, the query variable elicitation formats are checked for overlapping uncertainty sets. If two query variables depend on the values of the same physical factor, and if the value of this physical factor is not specified in the case structure, then this creates a probabilistic dependence between these two query variables.

The analyst should note all pairs of query variables where the effect of common factors in their respective uncertainty sets may give rise to significant dependencies. These should be checked with substantive experts to arrive at a final list of dependencies.

A format for eliciting dependencies may be taken from Part I, paragraph 6.1.

2.4. Identification of seed variables

Empirical control is built in the elicitation procedure by asking experts to assess calibration or seed variables. Seed variables are variables whose values are or will be known to the analyst within the frame of the exercise but not to the expert. Seed variables are important for assessing the performance of the combined experts' assessments. Seed variables also form an important part of the feedback to experts, helping them to gauge their subjective sense of uncertainty against quantitative measures of performance.

It is impossible to give an effective procedure for generating meaningful seed variables. If the analyst undertakes to generate his own seed variables, he must exercise a certain amount of creativity, perhaps supported the experts themselves. General guidelines and tips will be provided here.

Classification

Seed variables falling squarely within the experts' field of expertise are called *domain variables*. In addition to domain variables, it is permissible to use variables from fields which are adjacent to the experts' proper field. These are called *adjacent variables*. Adjacent variables are those about which the expert should be able to give an educated guess. It will often arise that a given seed variable is a domain variable for one expert and an adjacent variable for another expert.

In the literature one sometimes encounters the use of general knowledge variables in an expert judgment context ("what is the population of Wroclaw, Poland"). For such variables an educated guess from the expert could not be expected, and the use of general knowledge variables is not recommended. Psychometric experiments indicate that experts are not better than general public on general knowledge items (Cooke et al 1988).

Seed variables may also be distinguished according to whether they concern predictions or retrodictions. For predictions the true value does not exist at the time the question is answered, whereas for retrodictions, the true value exists at the time the question is answered, but is not known to the expert.

In general, domain predictions are the most meaningful, in terms of proximity to the items of interest, and are also the hardest to generate. Adjacent retrodictions are easier to generate, but are less closely related to the items of interest. Combining these notions we arrive at the following table with a crude evaluation:

	PREDICTIONS	RETRODICTIONS
Domain	+++	++
Adjacent	++	+

Table 1. Classification of seed variables, crude evaluation

The use of adjacent retrodictions is sanctioned by the supposition that performance on such variables correlates with performance on the items of interest. There is no direct proof of this supposition at present, but it has been found that "experience" positively correlates with performance on adjacent retrodictions, and does not correlate with performance on general knowledge variables (see Cooke et al. 1988). Generally, the field of interest has a relatively small community of experts who tend to be familiar with current experiments. It is therefore not opportune to identify the source of the data during the elicitation; however, the name and institute of the experimenter are important for post hoc review.

Practical issues

Practical issues regarding the seed variables are:

1. The seed variables should sufficiently cover the case structures for elicitation. Particularly, when one expert panel should tackle different sub fields, seed variables must be provided for all sub fields.

2. For each sub field at least ten seed variables are needed, preferably more. Distinct seed variables may be drawn from the same experiment, but there should be sufficient alternative experiments to provide independence among the seed variables data.

3. Seed variables may be, but need not be identified as such in the elicitation.

4. If possible, the analyst should be unaware of the values of the seed variables during the elicitation.

2.5. Identification of experts

The term "expert" is not defined by any quantitative measure of resident knowledge. Rather "expert for a given subject" is used here to designate a person whose present or past field contains the subject in question, and who is

regarded by others as being one of the more knowledgeable about the subject. Such persons are sometimes designated in the literature as "domain" or "substantive" experts, to distinguish them from "normative experts", i.e. experts in statistics and subjective probability. Identifying experts for a given subject therefore means identifying persons whose work terrain contains the subject and whom others regard as knowledgeable. The procedure for identifying the set of experts for a given subject is known as the Round Robin:

1. Some names of potential experts are generated within the organization responsible for the study (if the organization could not do this, they could not perform the study in the first place). These persons are approached and asked:
   - what is your background and knowledge base with regard to the subject?
   - which other persons are knowledgeable with regard to the subject?
2. The persons named in the first round are approached with the same two questions.
3. Step 2 is iterated until (a) no new names appear, of (b) it is judged that a sufficiently diverse set of experts is obtained.

The knowledge base describes the type of information on which the experts' assessments would be based. This may be either

- articles in journals or technical reports
- experimental or observational data
- computer modeling.

After the set of experts is identified, a choice is made which experts to use in the study. In general, the largest number of experts consistent with the level of available resources should be used. In any event at least four experts for a given subject should be chosen. The choice should made so as to diversify the knowledge bases and institutions of employment. The experts' time commitment to the study must be assessed and budgeted. This may include time spent researching the questions and composing a written rationale for his assessments.

2.6 Selection of experts

After the set of experts has been identified, a choice is made which experts to use in the study. In general, the largest number of experts consistent with the level of resources should be used. In any event, at least four experts for a given subject should be chosen. A panel of eight experts is to be recommended as a rule of thumb. The choice should be made so as to diversify the knowledge bases and institutions of employment. At least two experts should be from outside the institution performing the study, in cases internal expertise is at stake only. The following general selection criteria are used:

- reputation in the field of interest
- experimental experience in the field of interest
- number and quality of publications in the field of interest
- diversity in background
- awards received
- balance of views
- interest in and availability for the project

The nature and the broadness of a panel may require experts with very broad experience for which only a few are available (generalists). Panels may need a diversity of in-depth expertise; a mix of generalists and specialists. The requirement for the specialists is that they cover the whole panel's field sufficiently. For instance, in a panel on health effects, specialists may be required on the various organs of a human body, whereby only a few generalists have experience in the whole field of health effects. The following selection procedure for experts is recommended:

1. All potential experts named during the expert identification phase will be contacted (by mail and later by personal contact) to find out whether they are interested and whether they consider themselves a potential expert for that particular panel. During personal contacts potential experts are also asked to name other potential experts.
2. Potential experts send in a CV (curriculum vitae) indicating their expertise and availability for that specified panel.

3. All CV's will be reviewed by a nomination committee consisting of one to three persons from the project staff and one or two additional persons with thorough expertise in the field of interest not being involved as a potential expert themselves.

When the list of candidate experts is determined, these are contacted and invited to participate in the study. It is essential to clarify the conditions of participation, including:

1. Type of assessment task
2. Remuneration
3. Distribution of study results
4. Use of the experts name
5. Feedback of expert judgment data

Use of Experts' Names

Every expert is very jealous of his/her name and professional reputation. It is essential to clarify how the names will be used. The following procedure, developed over a number of years, seeks to satisfy the demands of openness and objectivity in science, as well as demands of freedom from conflict of interest, harassment and legal liability which may legitimately be raised by the experts themselves. If indeed expert judgment is scientific data, then it must be open to peer review. On the other hand, the expert's affiliation and professional activities may create a conflict of interest if his/her name is associated with the actual assessments. If told that his name would be published with his assessments, an expert in toxicology working at a pharmaceutical company might well say, "if you want the company viewpoint, ask the president of the company".

The proposed procedure is the following:

1. Expert names and affiliations are published in the study.
2. All information, including expert names and assessments, is available for competent peer review, but is not available for unrestricted distribution.
3. Individual assessments are available for unrestricted distribution, assessments are not associated with names but identified as "expert 1, 2,3,…" etc.
4. Expert rationales are available for unrestricted distribution.
5. Each individual expert receives feedback on his/her own performance assessment.
6. Any further published use of the expert's name requires the expert's approval.

2.7 Preparation of elicitation document

When the case structure is finalized, the target and query variables identified, the potential dependencies identified and the seed variables defined, the elicitation document can be prepared.

For the query variables, the schematic elicitation formats from paragraph 2.3 must be fleshed out. For seed variables, a similar type of question format must be prepared. In the case of seed variables, however, the case structure assumptions need not necessarily agree with the case structure as defined in paragraph 2.1. Needless to say, the factors assumed to be known in the elicitation of seed variables must be consistent with the conditions under which the actual measurement/experiment is performed.

The elicitation document should also contain any generally known data, tables, graphs etc. which the expert might like to consult. This is partly for convenience, but it often helps in clarifying the questions.

The elicitation document will be taken into the elicitation, and perhaps sent to the experts in advance. Therefore, it should be intelligible without the analyst's explanations. The following format for the elicitation document is recommended:

1. Brief statement of the purpose of the study
2. Statement of the conditions for participation
3. Brief description of subjective probability assessment, including illustrations of quantiles etc.

4. Brief explanation of the performance measures
5. Elicitation questions for query variables (including seed variables)
6. Elicitation questions for dependencies.
7. Graphs, tables, and other common reference material.

The conditions for participation include the issues raised in paragraph 2,6; time, remuneration, feedback and use of experts' name.

2.8. Dry run exercise

The dry run exercise aims at finding out whether the case structure document and the elicitation format document are unambiguously outlined and whether they capture all relevant information and questions. One or two persons experienced in the field of interest should be asked to provide comments on both documents. The dry run experts are asked to study the documents and comment on the following during the dry run session:

- is the case structure document clear
- are the questions clearly formulated
- is the additional information provided with each question appreciated
- is the time required to complete the elicitation too long or too short.

The dry run experts should preferably come from outside the selected panel members. If that is difficult to achieve expert panel members may be asked to do the dry run. After the dry run exercise the case structure document and the elicitation format document will be finalized and sent to the experts of the panel.

2.9. Expert training session

The experts are requested to provide subjective assessments on the query variables. They will represent their subjective assessments in terms of quantile points (e.g., 5%, 50% and 95%). Most experts are unfamiliar with quantifying their degree of belief in terms of quantiles. For that reason a training session is recommended with all experts present. Such a meeting is also useful to discuss the case structures. The training session organized in the Joint EU USNRC study typically lasted two days. Examples of training material are included in Appendix V. A general outline of the topics of the session are listed below:

1. introduction to the project and the expert panel performed by the project leader (typical duration: 30 to 45 minutes)
2. probabilistic training presentation performed by a project staff member with a scientific background in subjective probability theory (typical duration: two hours)
3. overview of the code(s) or models for which the exercise is done performed by a project staff member who is or was involved in the code or model development (typical duration: a half to one hour)
4. introduction to the panel's field of interest referring to the case structure document and elicitation format document performed by a project staff member who has a scientific background and experience in the field of interest, and who is or was preferably involved in the code or model development (typical duration: one to one and a half hour)
5. probabilistic training exercise by the experts supervised the project staff member who performed the probabilistic training presentation. For this purpose questions from available experiments may be used. As the experts will be asked to fill in the training exercise format on-the-spot, the experiment may be in principle known to the experts; the questions should not require extensive computing.
6. issues related to uncertainties in the field of interest performed and supervised by a project staff member preferably the person who presented the introduction to the panel's field of interest; in this part ample time for discussion with the experts must be planned (typical duration: one to two hours) .
7. introduction to the processing of the experts' assessments and the performance measures by a project staff member with expertise in the field of probability theory and the computer programs used (typical duration: one hour)
8. introduction to the issue of dependencies among uncertainties of the query variables by a project staff member with a scientific background in probability theory (typical duration: one hour)
9. assessment of training exercise by the project staff member who performed the probabilistic training presentation and exercise (typical duration: half an hour)

10. explanation of the elicitation procedure and what is expected from the experts; in particular, the experts need to writing is required for the rationales (typical duration: half an hour to one hour)
11. during the whole training session sufficient time must be taken for discussions with the experts on any relevant matter.

3 ELICITATION

3.1. Expert elicitation session

With each expert an individual elicitation session must be held in which all results will be reviewed and discussed. In the session a normative and a substantive analysts will be present. The normative analyst is a project staff member who is experienced with subjective probabilities and has experience in expert elicitation; this person leads the session. The substantive analyst is a project staff member who has experience in the field of interest and who has preferably contributed to the case structure and elicitation format documents. The duration of the elicitation session should not exceed four hours. Nominal duration is two to three hours.

3.2 Expert rationales

The expert rationales form a very valuable element in an expert judgment study, and may also constitute a considerable burden for the experts. If the elicitation requires expert to consult sources, do some modeling, do some calculations, run some codes, then it may not be much *additional* work to write this up in readable form.

The experts should be encouraged to bring their written rationales, at least in draft form, to the elicitation. Not only will this facilitate the clarification of substantive issues, but it helps to meet deadlines.

4. POST-ELICITATION

4.1 Combination of experts assessments

The combination of expert judgments is discussed amply in Part I. We assume that a convenient software tool (for example EXCALIBR, Cooke and Solomatine 1992) (for explanations of technical terms, see Appendix IV). A typical output is shown below.

We see the scores of 8 experts, and of the performance based (in this case the item weight) and equal weight decision maker (dm). Note that the performance based dm is better calibrated and more informative than the equal weight dm, and the individual experts. The optimal significance level (the parameter α in paragraph 5.3 of Part I) is 0.20. Experts with calibration score less than 0.20 are unweighted. For other experts their "unnormalized weight" is the product of their calibration score and their information score for seed variables. The normalized weights are blank, as the acutal weights for the item weight dm vary from item to item. Note that the unnormalized weight of the performance based dm is higher than that of any expert, whereas the unnormalized weight of the equal weight dm is lower than that of expert number 4.

The last column shows "weight with dm"; this is the weight which each expert *would* receive if the dm had been added as an extra expert. Note that both dm's would attract more weight than the original experts.

Output from other modules is provided in Appendix 1.

```
Case name  : DISPERSION       31. 3.99                    CLASS version 3.3
_____________________________________________________________________________

  Results of scoring experts.
    Bayesian updates : no.      Weights : item.   DM optimisation : yes.
  Significance level : 0.020  Calibration power :     1.0
```

Expert name	Calibr.	Mean rel.infor. total	Mean rel.infor. realiz.	Number realiz.	UnNorm. weight	Normalized weight no DM	Normalized weight with DM

1	Expert1	0.00010	2.078	1.281	23	0		0
2	Expert2	0.00010	1.594	1.431	23	0		0
3	Expert3	0.00100	1.504	1.285	23	0		0
4	Expert4	0.13000	1.286	1.242	23	0.16142		0.13288
5	Expert5	0.03000	1.092	1.622	23	0.04867		0.04006
6	Expert6	0.00500	1.590	1.540	23	0		0
7	Expert7	0.01000	1.508	1.506	23	0		0
8	Expert8	0.02000	1.840	1.312	23	0.02625		0.02160
	Perform	0.90000	1.024	1.087	23	0.97849		0.80545
	Equal	0.15000	0.811	0.862	23	0.12935		0.33166

Table 2. Expert scores for dispersion

A few practical remarks may be worth while.

- The decision whether to use the equal weight decision maker, the global weight decision maker or the item weight decision maker may be motivated by the performance of these decision makers, but may depend on other factors as well.
- It is a truism in statistical hypothesis testing that all hypotheses can be rejected with sufficient data. In the same vein, all experts will eventually receive low calibration scores if the number of seed variables is large enough. Calibration power adjusts the effective number of seed variables, as a percentage of the total number. Values less than one may be used to counteract very low group scores resulting from a very large number of seed variables, or to compare different expert panels by equalizing the effective number of seed variables.
- The choice of parameters in the calculation, such as intrinsic range, background measure, and calibration power, should be made on the basis best judgment, and must not be done to influence the scores of the decision makers. Indeed, lowering the calibration power is equivalent to lowering the effective number of seed variables, and this will always produce higher calibration scores.
- If the equal weight decision maker is used, it is nonetheless important to verify the performance.

4.2. Discrepancy and Robustness analysis

After the expert data has been collected and analyzed, a number of questions must be addressed. These may be grouped under the headings discrepancy analysis and robustness analysis.

Discrepancy analysis

One would like to know how much the experts agree among themselves. The best way to examine this is through the use of “range graphs”. The following figure shows the range graphs for a few items from the soil panel:

```
                          Range graph of input data
Item no. :    2     Item name:  S1_SM_CS_7Y        Scale : LOG
Experts
  1 [────────────────#─────────────────────]
  2 [────────────────#─────────]
  3                             [────────────────────────────────────#────]
  4                                                     [────────────#───]
DM  [══════════════════════════════#═════════════════════════════════════════]
Real░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░▓░░░░░░░░░░░░░░░░
                                                                0.42
    0.02                                                              0.98
                          Range graph of input data
Item no. :    3     Item name:  S1B_SL1_3Y_1C        Scale : LOG
Experts
  1                                                   [───────#──────────]
  2                   [───────────────#───────────────]
  3                 [───#─────]
  4                                                [──────────────#──────────────]
DM                  [══════════════════════════════════#══════════════════════]
Real▓░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░
        0.41
    0.41                                                              0.929
                          Range graph of input data
Item no. :    4     Item name:  S1B_SL1_3Y_5C        Scale : LOG
Experts
  1       [─────#────────────]
  2 [────────────────────#───────────────]
  3     [────────────────────────────────────────────────────#──────────────────]
  4           [──────────────────────#────────────────────────────────]
DM    [════════════════════#══════════════════════════════════════════════]
Real▓░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░░
        0.15
    0.15                                                              0.44
```

```
                           Range graph of input data
Item no. :    5     Item name:  S1B_SL1_3Y_10        Scale : LOG
Experts
  1                         [#──]
  2                      [───────────────────────#──────────────]
  3 [─────────────────────────────────────────────────────#────────────────────]
  4                           [─────#──────────────────]
DM       [═════════════════════#══════════════════════════════════════════]
Real
                           0.05
    0.03                                                                  0.25
```

Figure 7: Range graphs for Soil panel.

The “[“’s denote the 5% quantiles, “]”’s denote the 95% quantiles, and “#” denotes the medians. The four experts assessments are shown, and the dm is shown as a double line. When a realization is available, it is shown beneath the dm’s assessment, and numerical information is provided.

We may note that the experts disagree in their median assessments. On Items 2,3 there is little overlap in their uncertainty ranges. This means that the experts consider the assessments of other experts as improbable. On item 5 there is a great difference in the uncertainty ranges.

It is of the nature of expert judgment that there should be differences. An expert judgment study cannot have the goal to alter or reduce expert uncertainties; rather the goal is to form a clear picture of these uncertainties, and to communicate that picture accurately.

An overall indication of mutual agreement is obtained by computing the average relative information for eqch expert, with respect to the equal weight decision maker. These overall indications are useful in robustness analysis.

Robustness analysis

Robustness analysis is concerned with the questions, to what degree do the decision maker’s distributions depend on the particular choice of experts and seed variables?

The procedure for addressing these questions is straightforward. For robustness on experts, experts are excluded from the analysis one at a time, and the resulting decision maker is re-calculated. The relative information of these ‘perturbed decision makers’ may be compared to the overall agreement among the experts themselves. If the perturbed decision makers resemble the orginal decision maker more than the experts agree with each other, then we may conclude that the results are robust against choice of experts.

For robustness on items the same procedure is followed. Seed variables are excluded one at a time and the decision maker is recalculated. If the perturbed decision makers resemble the orginal decision maker more than the experts agree with each other, then we may conclude that the results are robust against choice of seed variables.

Examples of robustness analysis are included in Appendix I.

4.3 Feedback

The experts must have access to
- their assessments
- their calibration and information scores
- their weighing factors
- passages in which their name is used.
- whether the expert shows a tendency toward over- or underconfidence
- whether the expert shows a tendency to over- or underestimate.

Barring exceptional circumstances it is strongly recommended to make the final report available to the experts. This may strengthen their insight into the nature and purpose of expert judgment data.

4.4. Post-processing analysis

Post-processing is required in case target variables are not suitable as query variables. In this case, other query variables are chosen, which can be expressed as functions of the target variables (see paragraph 2.3). The decision maker's uncertainty distributions over the query variables must be 'pulled back' so as to yield a distribution over the target variables. The pull-back distribution on the target variables must have the property that, when pushed through onto the query variables, we retrieve as nearly as possible the original distribution from the combined expert assessments. This is a problem in probabilistic inversion, and such problems are mathematically hard.

A simple solution suitable for models was proposed in (Cooke, 1994 A) which, however does not always return good results in more complex situations. More advanced techniques are described in (Kraan, and Cooke 1996) and (Jones et al, appearing as EUR 18827). This is an area which is currentldy under active development.

4.5 Documentation

Finally, the results must be documented in a report. No specific guidelines are given for this, as the level of reporting will depend on requirements of the problem owners.

APPENDIX I

Summary Results of the EC-USNRC Uncertainty Study

The performance measures and performance based weighting of Part I were applied in the eight expert panels shown in Table 1. The experts for each panel are internationally recognized in their fields, and were selected according to the method described in Part II. The seed variables for the Late Health Effects panel are defined in terms of the follow-up of the Nagasaki and Hiroshima survivors, to be published in 2001. Hence the values of these variables are not available at present. For the other panels seed variables were queried. Table 1 shows the performance based combination and the equal weight combination for the other seven panels. For each panel, Table 1 shows the calibration score (1 is maximal, 0 is minimal), the mean information score (0 is minimal), and the 'virtual weight'. 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.

CASE	WEIGHTING	Calibr.	Mean inform	Number seed	virtual weight
DISPERSION	Perform	0.90000	1.024	23	0.80545
	Equal	0.15000	0.811	23	0.33166
DRY DEPOSITION	Perform	0.52000	1.435	14	0.50000
	Equal	0.00100	1.103	14	0.00168
WET DEPOSITION	Perform	0.25000	1.117	19	0.93348
	Equal	0.00100	0.793	19	0.07627
ANIMAL	Perform	0.75000	2.697	8	0.50000
	Equal	0.55000	1.778	8	0.19204
SOIL/PLANT	Perform	0.00010	1.024	31	0.13369
	Equal	0.00010	0.973	31	0.12779
INTERNAL DOSE	Perform	0.85000	0.796	55	0.52825
	Equal	0.11000	0.560	55	0.09217
EARLY HEALTH	Perform	0.23000	0.216	15	0.98749
	Equal	0.07000	0.165	15	0.94834
LATE HEALTH	Equal	*******	0.280	0	0

Table 1 Performance based and equal weight combinations

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.[1] In DISPERSION, ANIMAL and INTERNAL DOSE, 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. Table 2 shows the individual expert scores for the results in Table 1.

[1] 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.

DISPERSION

Expt	Cal	Mean Inf	# seed
1	0.0001	2.078	23
2	0.0001	1.594	23
3	0.0010	1.504	23
4	0.1300	1.286	23
5	0.0300	1.092	23
6	0.0050	1.590	23
7	0.0100	1.508	23
8	0.0200	1.840	23
Prf	0.9000	1.024	23
Eq	0.1500	0.811	23

DRY DEPOSITION

Expt	Cal	Mean Inf	# seed
1	0.0001	1.953	14
2	0.5200	1.435	14
3	0.0010	1.702	14
4	0.0010	1.732	14
5	0.0001	1.792	14
6	0.0010	2.234	14
7	0.0010	1.695	14
8	0.0005	1.985	14
Prf	0.5200	1.435	14
Eq	0.0010	1.103	14

WET DEPOSITION

Expt	Cal	Mean Inf	# seed
1	0.0001	2.638	19
2	0.0100	1.979	19
3	0.0010	1.009	19
4	0.0001	1.028	19
5	0.0010	1.565	19
6	0.0001	1.946	19
7	0.0001	1.252	19
Prf	0.2500	1.117	19
Eq	0.0010	0.793	19

ANIMAL

Exprt	Calibr.	Mean Inf	# seed
1	0.00100	2.658	8
2	0.00100	2.730	8
3	0.09000	1.689	8
4	0.75000	2.697	8
5	0.01000	2.835	6
6	0.64000	2.888	8
7	0.02000	2.821	7
Prf	0.75000	2.697	8
Eq	0.55000	1.778	8

SOIL/PLANT

Expt	Cal	Mean Inf	# seed
1	0.0001	2.376	31
2	0.0001	1.309	31
3	0.0001	1.346	31
4	0.0001	1.607	31
Prf	0.0001	1.024	31
Eq	0.0001	0.973	31

INT. DOSIMETRY

Expt	Cal	Mean Inf	# seed
1	0.0010	1.671	39
2	0.7300	0.822	55
3	0.0001	2.003	50
4	0.0001	2.366	39
5	0.0001	1.205	39
6	0.0050	0.838	28
Prf	0.8500	0.796	55
Eq	0.1100	0.560	55

EARLY HEALTH

Expt	Cal	Mean Inf.	# seed
1	0.0001	0.834	15
2	0.0001	1.375	15
3	0.0001	1.008	15
4	0.0001	0.966	15
5	0.0001	1.115	15
6	0.0001	0.573	15
7	0.0001	0.410	15
Prf	0.2300	0.216	15
Eq	0.0700	0.165	15

LATE HEATH

Expt	Cal	Mean Inf	# seed
1	*****	0.440	0
2	*****	1.379	0
3	*****	1.024	0
4	*****	0.507	0
5	*****	0.836	0
6	*****	0.599	0
7	*****	0.616	0
8	*****	0.988	0
Eq	*****	0.280	0

Table 2. Expert scores

The mean information of the performance based combination is usually slightly lower than that of the least informative experts, and the calibration score is typically substantially higher. This reflects the dominance of calibration over information in this weighting scheme. The equal weight combination has wider confidence bands still, and the calibration is typically lower than the best calibrated experts. Inspecting the data in Table 2, we see that the performance based combination for DRY DEPOSITION and ANIMAL, actually coincides with one of the experts. In other words, performance is optimized by assigning weight one to a single expert. This naturally raises the question of robustness with regard to expert choice. How much would the results differ if this one expert happened not to be available? One way to address this question is to repeat the analyses, leaving this expert out. If the differences between the original and the 'perturbed' combination are smaller than the differences among the experts themselves and if the performance is still acceptable, then there is no strong indication that the results are unrobust against choice of experts. Table 3 shows the results of these comparisons. Experts are excluded one at a time and the performance based combination is recalculated. Columns 2 and 3 show the mean information and calibration of the 'perturbed' combination. The differences of the experts among themselves are reflected in the last column, which shows the relative information of each expert with respect to the equal weight combination.

ROBUSTNESS ON EXPERTS: ANIMAL

Expert Excluded	Mean Inf.	Calibration	Rel.Inf Original	Rel.Inf Eq.Wgt
None	2.697	0.750	0	
1	2.697	0.75000	0	1.084
2	2.697	0.75000	0	0.987
3	2.045	0.75000	0	0.374
4	2.695	0.64000	0.569	0.719
5	2.697	0.70000	0	0.835
6	2.690	0.75000	0	0.818
7	2.697	0.75000	0	0.988

ROBUSTNESS ON EXPERTS: DRY DEPOSITION

Expert Excluded	Mean Inf	Calibration	Rel.Inf Original	Rel.Inf Eq.Wgt
None	1.435	0.520	0	
1	1.435	0.52000	0	0.852
2	1.245	0.05000	0.858	0.420
3	1.435	0.52000	0	0.555
4	1.435	0.52000	0	0.608
5	1.435	0.52000	0	0.651
6	1.446	0.52000	0	1.137
7	1.431	0.52000	0	0.618
8	1.435	0.52000	0	0.860

Table 3. Robustness on experts

We see from Table 3, that the robustness on experts for ANIMAL is satisfactory, whereas for DRY DEPOSITION it is marginal. Lack of robustness is always a danger when performance is optimized. The equal weight combination is almost always more robust, but the price of course is lower performance.

Finally, Table 4 compares cancer risks at various cites of the EU-USNRC study with those of other studies, for high dose, high dose-rate. These results are obtained from the LATE HEALTH panel and hence reflect the equal weight combination.

	EC-USNRC (+90% confidence)[2]	BIER V[3]	ICRP 60[4]	UNSCEAR[5]	COSYMA[6]
BONE	0.035 (<0.001, 0.88)	-	-	-	0.01
COLON	0.98 (0.011, 3.35)	-	3.24	0.6	2.24
BREAST	0.78 (0.11, 3..78)	0.35	0.97	1.0	0.80
LEUKEMIA	0.91 (0.026, 2.33)	0.95	0.95	1.1	0.52
LIVER	0.086 <0.001, 2.02)	-	-	1.2	-
LUNG	2.76 (0.59, 8.77)	1.70	2.92	2.50	0.90
PANCREAS	0.17 (<0.001, 1.26)	-	-	-	-
SKIN	0.039 (<0.001, 0.37)	-	0.03	-	0.01
STOMACH	0.30 (<0.001, 4.01)	-	0.51	1.4	-
THYROID	0.059 (,0.001, 0.71)	-	-	-	0.17
ALL OTHER	2.60 (<0.001, 10.8)	-	-	-	-
ALL CANCERS	10.2 (3.47, 28.5)	7.90	12.05	12.0	5.02

Table 4. Comparison of elicited high dose and high dose-rate lifetime low LET cancer risks for a general EU/US population with those derived from other sources ($10^{-2}Gy^{-1}$)

Although the median values of the EC-USNRC study generally agree with the values from the other studies in Table 4, the 90% central confidence intervals are sometimes significantly wider than the spread of values from these studies. Indeed, the spread of assessments in the last four columns of table 4 is *not* an assessment of uncertainty.

Conclusions

We collect a number of conclusions regarding the use of structured expert judgment.

1. Experts' subjective uncertainties may be used to advance rational consensus in the face of large uncertainties, in so far as the necessary conditions for rational consensus are satisfied.
2. Empirical control of experts' subjective uncertainties is possible.
3. Experts' performance as subjective probability assessors is not uniform, there are significant differences in performance.
4. Experts as a group may show poor performance.
5. A structured combination of expert judgment may show satisfactory performance, even though the experts individually perform poorly.
6. The performance based combination generally outperforms the equal weight combination.
7. The combination of experts' subjective probabilities, according to the schemes discussed here, generally has wider 90% central confidence intervals than the experts individually; particularly in the case of the equal weight combination.

We note that poor performance as a subjective probability assessor does *not* indicate a lack of substantive expert knowledge. Rather, it indicates unfamiliarity with quantifying subjective uncertainty in terms of subjective probability distributions. Experts were provided with training in subjective probability assessment, but of course their formal training does not (yet) prepare them for such tasks.

Finally we include the actual output for the expert panels.

```
Case name  : DISPERSION       31. 3.99                    CLASS version 3.3
```

[2] Radiation exposure-induced deaths (REID) for joint current EU-US population.
[3] BIER V calculates excess cancer deaths for current US population
[4] ICRP calculates REID average of risks for current UK and US populations.
[5] UNSCEAR calculates REID for current Japanese population.
[6] REID

Results of scoring experts.
Bayesian updates : no. Weights : item. DM optimisation : yes.
Significance level : 0.020 Calibration power : 1.0

Expert name	Calibr.	Mean rel.infor. total	Mean rel.infor. realiz.	Number realiz.	UnNorm. weight	Normalized weight no DM	Normalized weight with DM
1 Expert1	0.00010	2.078	1.281	23	0		0
2 Expert2	0.00010	1.594	1.431	23	0		0
3 Expert3	0.00100	1.504	1.285	23	0		0
4 Expert4	0.13000	1.286	1.242	23	0.16142		0.13288
5 Expert5	0.03000	1.092	1.622	23	0.04867		0.04006
6 Expert6	0.00500	1.590	1.540	23	0		0
7 Expert7	0.01000	1.508	1.506	23	0		0
8 Expert8	0.02000	1.840	1.312	23	0.02625		0.02160
Perform	0.90000	1.024	1.087	23	0.97849		0.80545
Equal	0.15000	0.811	0.862	23	0.12935		0.33166

Case name : DRY DEPOSITION 31. 3.99 CLASS version 3.3

Results of scoring experts.
Bayesian updates : no. Weights : global. DM optimisation : yes.
Significance level : 0.520 Calibration power : 1.0

Expert name	Calibr.	Mean rel.infor. total	Mean rel.infor. realiz.	Number realiz.	UnNorm. weight	Normalized weight no DM	Normalized weight with DM
1 Expert1	0.00010	1.953	1.779	14	0	0	0
2 Expert2	0.52000	1.435	1.339	14	0.69651	1.00000	0.50000
3 Expert3	0.00100	1.702	1.503	14	0	0	0
4 Expert4	0.00100	1.732	1.820	14	0	0	0
5 Expert5	0.00010	1.792	1.792	14	0	0	0
6 Expert6	0.00100	2.234	2.457	14	0	0	0
7 Expert7	0.00100	1.695	1.869	14	0	0	0
8 Expert8	0.00050	1.985	1.581	14	0	0	0
Perform	0.52000	1.435	1.339	14	0.69651		0.50000
Equal	0.00100	1.103	1.184	14	0.00118		0.00168

Case name : WET DEPOSITION 31. 3.99 CLASS version 3.3

Results of scoring experts.
Bayesian updates : no. Weights : global. DM optimisation : yes.
Significance level : 0.001 Calibration power : 1.0

Expert name	Calibr.	Mean rel.infor. total	Mean rel.infor. realiz.	Number realiz.	UnNorm. weight	Normalized weight no DM	Normalized weight with DM
1 Expert1	0.00010	2.638	2.338	19	0	0	0
2 Expert3	0.01000	1.979	0.593	19	0.00593	0.73846	0.04912
3 Expert4	0.00100	1.009	1.154	19	0.00115	0.14366	0.00956
4 Expert5	0.00010	1.028	1.718	19	0	0	0
5 Expert6	0.00100	1.565	0.947	19	0.00095	0.11788	0.00784
6 Expert7	0.00010	1.946	1.719	19	0	0	0
7 Expert8	0.00010	1.252	1.815	19	0	0	0
Perform	0.25000	1.117	0.451	19	0.11276		0.93348
Equal	0.00100	0.793	0.726	19	0.00073		0.07627

(c) 1995 TU Delft, SoLogic.

Case name : ANIMAL 31. 3.99 CLASS version 3.3

Results of scoring experts.
Bayesian updates : no. Weights : global. DM optimisation : yes.
Significance level : 0.750 Calibration power : 1.0

Expert name	Calibr.	Mean rel.infor. total	Mean rel.infor. realiz.	Number realiz.	UnNorm. weight	Normalized weight no DM	Normalized weight with DM
1 EXPERT1	0.00100	2.658	2.658	8	0	0	0
2 EXPERT2	0.00100	2.730	2.730	8	0	0	0
3 EXPERT3	0.09000	1.689	1.689	8	0	0	0
4 EXPERT4	0.75000	2.697	2.697	8	2.02257	1.00000	0.50000
5 EXPERT5	0.01000	2.835	2.835	6	0	0	0
6 EXPERT6	0.64000	2.888	2.888	8	0	0	0
7 EXPERT7	0.02000	2.821	2.821	7	0	0	0
Perform	0.75000	2.697	2.697	8	2.02257		0.50000
Equal	0.55000	1.778	1.778	8	0.97768		0.19204

(c) 1995 TU Delft, SoLogic.

Case name : SOIL 31. 3.99 CLASS version 3.3

Results of scoring experts.
Bayesian updates : no. Weights : global. DM optimisation : yes.
Significance level : 0 Calibration power : 1.0

Expert name	Calibr.	Mean rel.infor. total	Mean rel.infor. realiz.	Number realiz.	UnNorm. weight	Normalized weight no DM	Normalized weight with DM
1 EXPERT1	0.00010	2.376	2.376	31	0.00024	0.35801	0.31015
2 EXPERT2	0.00010	1.309	1.309	31	0.00013	0.19714	0.17079
3 EXPERT3	0.00010	1.346	1.346	31	0.00013	0.20273	0.17562
4 EXPERT5	0.00010	1.607	1.607	31	0.00016	0.24212	0.20975
Perform	0.00010	1.024	1.024	31	0.00010		0.13369
Equal	0.00010	0.973	0.973	31	0.00010		0.12779

(c) 1995 TU Delft, SoLogic.

Case name : INTERNAL DOSIMETRY 31. 3.99 CLASS version 3.3

Results of scoring experts.
Bayesian updates : no. Weights : global. DM optimisation : yes.
Significance level : 0.005 Calibration power : 1.0

Expert name	Calibr.	Mean rel.infor. total	Mean rel.infor. realiz.	Number realiz.	UnNorm. weight	Normalized weight no DM	Normalized weight with DM
2 exp 2	0.00100	1.671	1.671	39	0	0	0
3 exp 3	0.73000	0.822	0.822	55	0.60009	0.99307	0.46848
4 exp 4	0.00010	2.003	2.003	50	0	0	0
5 exp 5	0.00010	2.366	2.366	39	0	0	0
6 exp 6	0.00010	1.205	1.205	39	0	0	0
8 exp 8	0.00500	0.838	0.838	28	0.00419	0.00693	0.00327
Perform	0.85000	0.796	0.796	55	0.67665		0.52825
Equal	0.11000	0.560	0.560	55	0.06158		0.09217

(c) 1995 TU Delft, SoLogic.

Case name : EARLY HEALTH 31. 3.99 CLASS version 3.3

Results of scoring experts.

Bayesian updates : no. Weights : global. DM optimisation : yes.
Significance level : 0 Calibration power : 1.0

Expert name	Calibr.	Mean rel.infor. total	Mean rel.infor. realiz.	Number realiz.	UnNorm. weight	Normalized weight no DM	Normalized weight with DM
1 exp 1 +	0.00010	0.834	0.834	15	0.00008	0.13272	0.00166
2 exp 2 +	0.00010	1.375	1.375	15	0.00014	0.21887	0.00274
3 exp 3 +	0.00010	1.008	1.008	15	0.00010	0.16053	0.00201
5 exp 5 +	0.00010	0.966	0.966	15	0.00010	0.15372	0.00192
6 exp 6 +	0.00010	1.115	1.115	15	0.00011	0.17756	0.00222
8 exp 8 +	0.00010	0.573	0.573	15	0.00006	0.09128	0.00114
9 exp 9 +	0.00010	0.410	0.410	15	0.00004	0.06532	0.00082
Perform	0.23000	0.216	0.216	15	0.04958		0.98749
Equal	0.07000	0.165	0.165	15	0.01153		0.94834

(c) 1995 TU Delft, SoLogic.

Case name : LATE HEATH 31. 3.99 CLASS version 3.3

Results of scoring experts.
Bayesian updates : no. Weights : equal. DM optimisation : no.
Calibration power : 1.0

Expert name	Calibr.	Mean rel.infor. total	Mean rel.infor. realiz.	Number realiz.	UnNorm. weight	Normalized weight no DM	Normalized weight with DM
1 exp 1	0.99990	0.440	0	0	0	0.12500	0
2 exp 2	0.99990	1.379	0	0	0	0.12500	0
3 exp 3	0.99990	1.024	0	0	0	0.12500	0
4 exp 4	0.99990	0.507	0	0	0	0.12500	0
5 exp 5	0.99990	0.836	0	0	0	0.12500	0
6 exp 6	0.99990	0.599	0	0	0	0.12500	0
7 exp 7	0.99990	0.616	0	0	0	0.12500	0
8 exp 8	0.99990	0.988	0	0	0	0.12500	0
Equal	0.99990	0.280	0	0	0		0

(c) 1995 TU Delft, SoLogic.

Table 5. Expert scores for all panels.

APPENDIX II:

REPORTS PUBLISHED AS A RESULT OF THE JOINT ec/usnrc PROJECT ON UNCERTAINTY ANALYSIS OF PROBABILISTIC ACCIDENT CONSEQUENCE CODES (under the Third EC-Framework Programme)

1 F.T. Harper, L.H.J. Goossens, R.M. Cooke, S.C. Hora, M.L. Young, J. Päsler-Sauer, L.A. Miller, B. Kraan, C. Lui, M.D. McKay, J.C. Helton and J.A. Jones
Probabilistic accident consequence uncertainty study: Dispersion and deposition uncertainty assessment
Prepared for U.S. Nuclear Regulatory Commission and Commission of European Communities
NUREG/CR-6244, EUR 15855 EN, SAND94-1453
Washington/USA, and Brussels-Luxembourg, November 1994, published January 1995
Volume I: Main report
Volume II: Appendices A and B
Volume III: Appendices C, D, E, F, G, H

2 R.M. Cooke, L.H.J. Goossens and B.C.P. Kraan
Methods for CEC\USNRC accident consequence uncertainty analysis of dispersion and deposition - Performance based aggregating of expert judgements and PARFUM method for capturing modeling uncertainty
Prepared for the Commission of European Communities, EUR 15856, Brussels-Luxembourg, June 1994, published 1995

3 J. Brown, L.H.J. Goossens, F.T. Harper, B.C.P. Kraan, F.E. Haskin, M.L. Abbott, R.M. Cooke, M.L. Young, J.A. Jones S.C. Hora, A. Rood and J. Randall
Probabilistic accident consequence uncertainty study: Food chain uncertainty assessment
Prepared for U.S. Nuclear Regulatory Commission and Commission of European Communities
NUREG/CR-6523, EUR 16771, SAND97-0335
Washington/USA, and Brussels-Luxembourg, March 1997, published June 1997
Volume 1: Main report
Volume 2: Appendices

4 L.H.J. Goossens, J. Boardman, F.T. Harper, B.C.P. Kraan, R.M. Cooke, M.L. Young, J.A. Jones and S.C. Hora
Probabilistic accident consequence uncertainty study: Uncertainty assessment for deposited material and external doses
Prepared for U.S. Nuclear Regulatory Commission and Commission of European Communities
NUREG/CR-6526, EUR 16772, SAND97-2323
Washington/USA, and Brussels-Luxembourg, September 1997, published December 1997
Volume 1: Main report
Volume 2: Appendices

5 F.E. Haskin, F.T. Harper, L.H.J. Goossens, B.C.P. Kraan, J.B. Grupa and J. Randall
Probabilistic accident consequence uncertainty study: Early health effects uncertainty assessment
Prepared for U.S. Nuclear Regulatory Commission and Commission of European Communities
NUREG/CR-6545, EUR 16775, SAND97-2689
Washington/USA, and Brussels-Luxembourg, November 1997, published December 1997
Volume 1: Main report
Volume 2: Appendices

6 M. Little, C.M. Muirhead, L.H.J. Goossens, F.T. Harper, B.C.P. Kraan, R.M. Cooke and S.C. Hora
Probabilistic accident consequence uncertainty study: Late health effects uncertainty assessment

Prepared for U.S. Nuclear Regulatory Commission and Commission of European Communities
NUREG/CR-6555, EUR 16774, SAND97-2322
Washington/USA, and Brussels-Luxembourg, September 1997, published December 1997
Volume 1: Main report
Volume 2: Appendices

7 L.H.J. Goossens, J.D. Harrison, F.T. Harper, B.C.P. Kraan, R.M. Cooke and S.C. Hora
Probabilistic accident consequence uncertainty study: Uncertainty assessment for internal dosimetry
Prepared for U.S. Nuclear Regulatory Commission and Commission of European Communities
NUREG/CR-6571, EUR 16773, SAND98-0119
Washington/USA, and Brussels-Luxembourg, February 1998, published April 1998
Volume 1: Main report
Volume 2: Appendices

ADDITIONAL STUDY AS A RESULT OF THE JOINT STUDY

8 L.H.J. Goossens, R.M. Cooke and B.C.P. Kraan
Evaluation of weighting schemes for expert judgement studies
Final report prepared under contract Grant No. Sub 94-FIS-040 for the Commission of European Communities, Directorate-General for Science, Reserach and Development, XII-F-6
Delft University of Technology, Delft/NL, December 1996, 75 p.

REPORTS TO BE PUBLISHED ON THE PROJECT UNCERTAINTY ANALYSIS OF THE PROBABILISTIC ACCIDENT CONSEQUENCE CODE COSYMA USING EXPERT JUDGEMENT (under the Fourth EC-Framewok Programme)

1: R.M. Cooke, L.H.J. Goossens, B.C.P. Kraan
Probabilistic Accident Consequence Uncertainty Assessment
Procedures Guide Using Expert Judgement
To be published as EUR 18820 (This document).

2: L.H.J. Goossens, J.A. Jones, J. Ehrhardt, B.C.P. Kraan
Probabilistic Accident Consequence Uncertainty Assessment
Countermeasures Uncertainty Assessment
To be published as EUR 18821

3 J.A. Jones, J. Ehrhardt, F. Fischer, I. Hasemann, L.H.J. Goossens, B.C.P. Kraan, R.M. Cooke
Probabilistic Accident Consequence Uncertainty Assessment Using COSYMA
Uncertainty from the Atmospheric Dispersion and Deposition Module
To be published as EUR 18822

4 J.A. Jones, J. Brown, F. Fischer, I. Hasemann, L.H.J. Goossens, B.C.P. Kraan, R.M. Cooke
Probabilistic Accident Consequence Uncertainty Assessment Using COSYMA
Uncertainty from the Food Chain Module
To be published as EUR 18823

5 J.A. Jones, F. Fischer, I. Hasemann, L.H.J. Goossens, B.C.P. Kraan, R.M. Cooke
Probabilistic Accident Consequence Uncertainty Assessment Using COSYMA
Uncertainty from the Health Effects Module
To be published as EUR 18824

6 J.A. Jones, F. Fischer, I. Hasemann, L.H.J. Goossens, B.C.P. Kraan, R.M. Cooke
Probabilistic Accident Consequence Uncertainty Assessment Using COSYMA
Uncertainty from the Dose Module
To be published as EUR 18825

7 J.A. Jones, J. Ehrhardt, L.H.J. Goossens, F. Fischer, I. Hasemann, B.C.P. Kraan, R.M. Cooke
Probabilistic Accident Consequence Uncertainty Assessment Using COSYMA

Uncertainty from the Complete System
To be published as EUR 18826

8 J.A. Jones, B.C.P. Kraan, R.M. Cooke, L.H.J. Goossens, F. Fischer, I. Hasemann
Probabilistic Accident Consequence Uncertainty Assessment Using COSYMA
Methodology and Processing Techniques
To be published as EUR 18827

References

Apostolakis G. and Kaplan S. "Pitfalls in risk calculations Reliability Engineering 1981:2, 135-145.

Aspinall W., (1996) Expert judgment case studies, Cambridge Program for Industry, Risk management and dependence modeling, Cambridge.

Baverstam, U. Davis, P. Garcia-Olivares, A. Henrich, E. and Koch, J. (1993) BIOMOVS II "Guidellines for Uncertainty Analysis" Technical Report No. 1 Stolkholm.

Beckman, R. and McKay, M. (1987) "Monte Carlo estimation under different distributions using the same simulation" Technometrics, vol 29 no 2, 153-160.

Best, D.J. and Roberts, D.E. (1974) "The percentage points of the chi square distribution, Appl. Stat. 22, 385-388.

Bier, V. (1983) "A measure of uncertainty importance for components in fault trees" PhD. thesis, Laboratory for Information and Decision systems, MIT Boston.

Bhola, B., Blauw, H., Cooke, R., and Kok, M. (1991) Expert opinion in project management, *European Journal of Operations Research,* 57, p 1-8.

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

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

Brown, J., Goossens, L.H.J., Harper, F.T., Haskin, E.H., Kraan, B.C.P., Abbott, M.L., Cooke, R.M., Young, M.L., Jones, J.A., Hora, S.C., and Rood, A., *Probabilistic accident consequence uncertainty analysis: Food chain uncertainty assessment*, Prepared for U.S. Nuclear Regulatory Commission and Commission of European Communities, NUREG/CR-6523, EUR 16771, Washington/USA, and Brussels-Luxembourg, (Volumes 1 and 2), 1997.

Claessens, M., (1990) An application of expert opinion in ground water transport (in Dutch) TU Delft, DSM Report R 90 8840.

Comer, K., Seaver, D., Stillwell, W. and Gaddy, C. (1984) Generating Human Reliability Estimates Using Expert Judgment, vols. I and II NUREG/CR-3688.

Cooke R.M., (1991) Expert judgment study on atmospheric dispersion and deposition Report Faculty of Technical Mathematics and Informatics No.01-81, Delft University of Technology.

Cooke, R. M., (1991) Experts in Uncertainty, Oxford University press.

Cooke R.M. (1994), Uncertainty in dispersion and deposition in accident consequence modeling assessed with performance-based expert judgment, Reliability Engineering and System Safety, , Vol. 45, pp.35-46.

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

Cooke, R.M., (1995) UNICORN Methods and Code for Uncertainty Analysis SRD Association, AEA.

Cooke R.M. and Solomatine, D. (1992) EXCALIBR integrated system for processing expert judgments version 3.0, User’s manual, prepared under contract for Directorate-General XII, Delft.

Cooke, R. M., and Kraan, B. (1996) "Dealing with dependencies in uncertainty analysis" Probabilistic Safety Assessment and Management, Proceedings ESREL-96-PSAM III (Cacciabue, P.C. and Papazouglou, I.A. eds.) vol. 2 pp 988-991, Springer, New York.

Cooke, R. M. and Waij, R. (1986) "Monte Carlo sampling for generalized knowledge dependence with application to human reliability" Risk Analysis, vol.6 no.3\pp 335-343.

Cooke, R. M., Goossens, L. and Kraan, B. (1994) "Methods for CEC\USNRC Accident Consequence Uncertainty Analysis of Dispersion and Deposition" EUR 15855 EN.

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

Cooke, R.M. and Jager, E., (1998) Failure frequency of underground gas pipelines: methods for assessment with structured expert judgment, Risk Analysis, vol 18 No.4 pp 511-527.

Crick, M. Hofer, E. Jones, H. and Haywood, S. (1988) "Uncertainty analysis of the foodchain and atmospheric dispersion modules of MARC. National Radiological Protection Board Report NRBP-R184, Chilton, Didcot, Oxon.

Dagpunar, J. (1988) Principles of Random Variate Generation, Carendon Press, Oxford.

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

De Ruyter van Steveninck, J. (1994) "Uncertainty analysis: an evaluation of methods, techniques and codes" ECN-TUD, NT-RA-94-09.

French, S. (1985) "Group consensus probability distributions: a critical survey" in Bernardo, J., De Groot, D., Lindley, D and smith, A (eds) *Bayesian Statistics* Elsevier, North Holland, pp 183-201.

Frijters, M., Cooke, R. Slijkuis, K. and van Noortwijk, J.(1999) Expert Judgment Uncertainty Analysis for Inundation Probability, (in Dutch) Ministry of Water Management, Bouwdienst, Rijkswaterstaat, Utrecht.

Glaser, H. Hofer, E. Kloos, M. and Skorek, T. (1994) "Uncertainty and sensitivity analysis of a post-experiment calculation in thermal hydraulics" Reliability Engineering and System Safety, vol. 45, nos 1,2; 3-19.

Gokhale, D. and Press, S. (1982) Assessment of a prior distribution for the correlation coefficient in a bivariate normal distribution, J.R. Stat. Soc. A vol. 145 P.2, 237-249.

Goossens LHJ, Cooke RM, and Kraan, BCP, (1996) Evaluation of weighting schemes for expert judgment studies, Final report prepared under contract Grant No. Sub 94-FIS-040 for the Commission of the European Communities., Directorate General for Science, Research and Development XII-F-6, Delft University of Technology, Delft, the Netherlands.

Goossens LHJ, Cooke RM, Woudenberg F and van der Torn P.(1992) Probit functions and expert judgment: Report prepared for the Ministry of Housing, Physical Planning and Environment, the Netherlands; Delft University of Technology, Safety Science Group and Department of Mathematics, and Municipal Health Service, Rotterdam, Section Environmental Health, October.

Goossens, L. Cooke, R. and Kraan, B. (1998) "Evaluation of weighting schemes for expert judgment studies" in Probabilistic Safety Assessment and Management,)Proceedings of the 4th International conference on Probabilistic Safety Assessment and Management (Mosleh, A. and Bari, R. eds), pp 1937-1942, Springer, New York.

Goossens, L., Cooke, R. and van Steen, J. (1989) Expert Opinions in Safety Studies vols. 1 - 5 Philosophy and Technical Social Sciences, Delft University of Technology, Delft, The Netherlands.

Goossens, L.H.J., and Harper, F.T., (1998) *Joint EC/USNRC expert judgement driven radiological protection uncertainty analysis,* Journal of Radiological Protection, Vol.18, No.4, 249-264.

Goossens, L.H.J. , Boardman, J., Harper, F.T., Kraan, B.C.P., Young, M.L., Cooke, R.M., Hora, S.C., and Jones, J.A., *Probabilistic accident consequence uncertainty analysis: Uncertainty assessment for deposited material and external doses*, Prepared for U.S. Nuclear Regulatory Commission and Commission of European Communities, NUREG/CR-6526, EUR 16772, Washington/USA, and Brussels-Luxembourg, (Volumes 1 and 2), 1997.

Goossens, L.H.J., Harrison, J.D., Harper, F.T., Kraan, B.C.P., Cooke, R.M., and Hora, S.C., *Probabilistic accident consequence uncertainty analysis: Internal dosimetry uncertainty assessment*, Prepared for U.S. Nuclear Regulatory Commission and Commission of European Communities, NUREG/CR-6571, EUR 16773, Washington/USA, and Brussels-Luxembourg, (Volumes 1 and 2), 1998.

Granger Morgan,M. and 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., and Walster, A., (1973) "A comparative study of differences in subjective likelihood estimates made by individuals, interacting groups, Delphi groups and nominal groups" Organizational Behaviour and Human Performance, vol. 9 pp 280-291.

Harper FT, Goossens LHJ, Cooke RM, Hora SC, Young ML, Päsler-Sauer J, Miller LA, Kraan BCP, Lui C, McKay MD, Helton JC, Jones JA. (1995) Joint USNRC/CEC consequence uncertainty study: Summary of objectives, approach, application, and results for the dispersion and deposition uncertainty assessment. Prepared for U.S. Nuclear Regulatory Commission and Commission of European Communities, NUREG/CR-6244, EUR 15855, Washington/USA, and Brussels-Luxembourg, (Volumes I, II, III).

Harper, F. Goossens, L. Cooke, R. Hora, S. Young, M. Päsler-Ssauer, J. Miller, L., Kraan, B. Lui, C. McKay, M. Helton, J. Jones, A. (1994) Joint USNRC\CEC consequence Uncertainty Study: Summary of Objectives, Approach, Application, and Results for the Dispersion and Deposition Uncertainty Assessment.Vol III, NUREG/CR-6244, EUR 15755 EN, SAND94-1453.

Haskin, F.E. , Goossens, L.H.J., Harper, F.T., Grupa, J., Kraan, B.C.P., Cooke, R.M., and Hora, S.C., *Probabilistic accident consequence uncertainty analysis: Early health uncertainty assessment*, Prepared for U.S. Nuclear Regulatory Commission and Commission of European Communities, NUREG/CR-6545, EUR 16775, Washington/USA, and Brussels-Luxembourg, (Volumes 1 and 2), 1997.

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

Hogarth, R. (1987) Judgement and Choice Wiley, New York.

Hora, S. and Iman, R. (1989) "Expert opinion in risk analysis: the NUREG-1150 methodology" Nuclear Science and Engineering, 102-323.

Iman R. and Helton J. (1985) "A comparison of uncertainty and sensitivity analysis techniques for computer models" NUREG/CR-3904 SAND84-1461 RG, Albuquerque.

Iman R. and Helton, J., (1988) "Investigation of uncertainty and sensitivity analysis techniques for computer models" Risk Analysis, 8, 71.

Iman, R. and Conover, W. (1982) "A distribution-free approach to inducing rank correlation among input varibles" Communications in Statistics-Simulation and Computation 11 (3) 311-334.

Iman, R. and Shortencarier, M. (1984) "A Fortran 77 program and user's guide for the generation of latin hypercube and random samples for use with computer models" NUREG/CR-3624, Sandia National Laboratories, Albuquerque.

Iman, R. Helton, J. and Cambell, J.(1981) "An approach to sensitivity analysis of computer models: Part I and II, J.of Quality Technology, 13 (4).

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

Kalos, M. and Whitlock, P. (1986) Monte Carlo Methods, Wiley and Sons, New York.

Kraan, B. and Cooke, R.M. ((1996) Post-processing techniques for the joint EU-NRCuncertainty analysis of accident consequence codes, *J. of Stat. Comp. and Simulation*, vol 55, no 2. Pp 243-261.

Krzykacz B. and Hofer, E. (1988) The generation of experimental designs for uncertainty and sensitivity analysis of model predictions with emphasis on dependences between uncertain parameters" in G. Desmet (ed) Reliability of Radioactive Transfer Models, Elsevier, London.

Kullback,, S. (1959) Information Theory and Statistics Wiley, New York.

Lehmann, E. (1963) "Some concepts of dependence" Ann. Math. Stat. 37, no 5, 1137-1153

Little, M. , Muirhead, C.M., Goossens, L.H.J., Harper, F.T., Kraan, B.C.P., Cooke, R.M. ,and Hora, S.C., *Probabilistic accident consequence uncertainty analysis: Late health uncertainty assessment*, Prepared for U.S. Nuclear Regulatory Commission and Commission of European Communities, NUREG/CR-6555, EUR 16774, Washington/USA, and Brussels-Luxembourg, (Volumes 1 and 2), 1997.

McKay, M. (1988) "Sensitivity and uncertainty analysis using a statistical sample of input values. Chapter 4 in Uncertainty Analysis (Y. Ronen ed) CRC Press, Boca Raton, Fla. 145-186.

McKay, M. Beckman, R. and Conover, W. (1979) "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code", Technometrics vol 21 no 2 239-245.

Meeuwissen, A. (1993) "Dependent random variables in uncertainty analysis" PhD dissertation, Dept. of Mathematics, Delft University of Technology, Delft.

Meeuwissen, A. and Cooke R. (1994) "Tree dependent random variables", Dept. of Mathematics, Delft University of Technology, Report 94-28, Delft.

Meima B., (1990) Expert opinion and space debris, Technological Designer's Thesis, Faculty ot Technical Mathematics and Informatics, Delft University of Technology, Delft.

Murphy, A.H. and Daan, H., (1982) Subjective probability forecasting in The Netherlands: some operational and experimental results, *Meteorol. Rsch.* 35, pp 99-112.

Murphy, A.H., and Daan, H. (1984) Impacts of feedback and experience of subjective probability forecasts: comparison of results from the first and second years of the Zierikzee experiment, *American Meteorological Society*, March, pp 413-423.

Nowak, E, and Hofer, E. (1988) "DIVIS a program package to support the probabilistic modeling of parameter uncertainties" in G. Desmet (ed) Reliability of Radioactive Transfer Models, Elsevier, London.

Offerman, J.(1990) Safety analysis of the carbon fibre reinforced composite material of the Hermes cold structure, TU-Delft/ESTEC, Noordwijk, the Netherlands.

Ripley, B. (1987) Stochastic Simulation, Wiley, New York.

Roeleven, D., Cooke, R.M., and Kok, M. (1991) Combining expert probabilistic weather foracsts, ISSN 0922-5641, report of the faculty of Technical Mathematics and Informatics, no 91-101,k Delft.

Ross, S. (1990) A Course in Simulation, Macmillan, New York.

Rubinstein, R. (1981) Simulation and the Monte Carlo Method John Wiley and Sons, Inc. New York.

Saltelli, A. (appearing) Mathematical and Statistical Methods for Sensitivity Analysis of Model Output, Wiley, New York.

Savage, L. (1954) The Foundations of Statistics, Dover, New York.

T-Book Reliability Data of Components in Nordic Nuclear Power Plants, 3rd edition, prepared by the ATV Office and Studsvik AB,Vattenfall AB, 1992.

Ter Haar, T.R., Retief, J.V. and Dunaiski, P.E. (1998) Towards a more rational approach of the serviceability limit states design of industrial steel structures paper no. 283, 2end World conference on steel in construction, San Sebastian, Spain.

Van Dorp, Rene (1991) Dependence Modeling for Uncertaintty Analysis, Technological Designers thesis, Department of Mathematics, Delft University of Technology, Delft.

Van Elst NP. (1997) Betrouwbaarheid beweegbare waterkeringen [Reliability of movable water barriers] Delft University Press, WBBM report Series 35.