Journal of Hazardous Materials, 29 (1992) 365–385
Elsevier Science Publishers B.V., Amsterdam

A perspective on structured expert judgment

Jacques F.J. van Steen

Department of Industrial Safety, Netherlands Organization for Applied Scientific Research TNO, P.O. Box 342, 7300 AH Apeldoorn (Netherlands)

(Received August 20, 1990; accepted in revised form July 31, 1991)

Abstract

Expert judgment has always been used in risk analysis, both in the qualitative and in the quantitative phases of such analyses. Focusing on the use of expert judgment for obtaining quantitative statements about uncertain quantities, it appears that expert judgment is a data source with a number of special characteristics. Consequently, the appropriate use of this source requires a structured approach and the development and introduction of specialized methodological tools. These formalized methods should deal with such aspects as the choice of experts, the encoding of uncertainty, the elicitation of judgments, and the evaluation and combination of judgments. Purpose of this paper is to review the state of the art of the use of expert judgment, in particular in risk analysis. The paper discusses the key characteristics which underline the need for a structured approach to expert judgment, and the procedures which may be used in applying structured expert judgment. Also, an impression is given of a number of practical experiences of the use of expert judgment. The paper concludes with a discussion of "lessons learned" and of unresolved issues which require further attention.

1. Introduction

Complex decision problems in both government and industry are often characterized by a lack of data, or by insufficient/inappropriate data. In such situations, one has to rely on the judgments of staff (and/or external) experts. Thus, expert judgment has always been used, and this also holds for risk analysis.

The *qualitative* analysis of the system, with which any risk analysis starts, makes up the bulk of expert judgment going into that analysis, and to date little attention has been given to formalizing the processes involved (e.g. in order to ensure some degree of completeness and reproducibility). Although this field of expert judgment application is beyond the scope of the present paper, it merits future attention.

This paper is concerned with the use of expert judgment as a source of data in the *quantitative* phases of risk analysis. Thus, the focus is on obtaining

Correspondence to: Dr. J.F.J. van Steen, Department of Industrial Safety, Netherlands Organization for Applied Scientific Research TNO, P.O. Box 342, 7300 AH Apeldoorn (Netherlands)

quantitative statements about uncertain quantities from experts. Important areas for such use of expert judgment include the assessment of human error probabilities and of failure rates of mechanical components. (It is noted by Suokas and Kakko [1] that the most common criticism on safety analysis has been focused on the uncertainties in component failure rate and human error data employed in the quantification of risks.) Typically, the role of experts is denoted by such frequently occurring phrases as "these data are based on engineering judgment".

A number of initial observations have to be made about expert judgment as data source. Firstly, expert judgment should never be substituted for objective data when the latter is available. Secondly, it should always be realized that the production and application of such "objective data" involves a great deal of expert judgment as well. (Mosleh [2] argues that "objective" data are nonexistent even when data is collected through detailed and careful review of plant operating records.) Thirdly, the availability of experts as source of data should not discourage analysts from collecting good data. Finally, and this issue is discussed in more detail in a later section, expert judgment appears to be a data source with a number of special characteristics.

The purpose of this paper is to review the state of the art of the use of expert judgment, in particular in risk analysis; see [3–11] for other surveys. The organization of the paper is as follows. Section 2 specifies the formalism to be used for representing uncertainty and the type of problem to be addressed. In Section 3, the key characteristics are introduced which underline the need for a structured approach to expert judgment. In Section 4, the procedures which may be used in applying structured expert judgment are discussed in more detail. Section 5 deals with a number of practical experiences of the use of expert judgment. The paper concludes with a discussion of "lessons learned" and of unresolved issues which require further attention.

2. Encoding of uncertainty and problem definition

Judgment is being applied in those situations where quantitative statements have to be made about uncertain quantities. Thus, an important aspect of incorporating expert judgment in science and engineering concerns the representation of uncertainty. Since a number of formalisms have been suggested for representing uncertainty, it is necessary to specify the formalism which is used in this paper. Moreover, different reasons for consulting experts can be distinguished, and thus it is also necessary to specify the type of problem which is being addressed.

2.1 Encoding of uncertainty

Uncertainty is encoded at two distinct points in the use of expert judgment: by the experts, in order to communicate it to the analyst, and by the analyst,

in order to use it in the analysis [11]. Moreover, a number of formalisms have been suggested for representing uncertainty: subjective probability and probability theory, membership functions and possibility theory, certainty factors, belief functions, and natural language (see [8,9,12–14] for further discussions). The question then is: which formalism to use for each task? It is noted in [11] that there is no a priori reason why the same formalism for representing uncertainty should be used for the two tasks mentioned above, nor that each expert should use the same formalism. Be that as it may, subjective probabilities have been used on a large scale in risk analysis since the *Reactor Safety Study* [15] in 1975 (see [9] for a short historical background). Thus, the discussion in this paper is limited to the use of (subjective) probability for encoding of uncertainty.

2.2 Problem definition

In a paper on group consensus probability distributions, French [16] distinguishes three types of problems in which experts are asked for advice:
(1) the expert problem;
(2) the group decision problem; and
(3) the textbook problem.

The first two types of problems are characterized by the existence of a real, pre-defined decision problem. In the expert problem, the group of experts is asked for their advice by a decision maker who is (or can be taken to be) outside the group, has the task of aggregating the judgments, has responsibility for the decision and is accountable for its consequences. The group itself may meet, or the experts may interact individually with the decision maker. In the group decision problem, the group of experts itself is responsible and accountable for the decision. In the textbook problem, there is no pre-defined decision problem: the group of experts are asked for their advice, which may be used in the future in as yet undefined circumstances.

The reason for distinguishing between these types of problems is that the responsibilities and accountabilities of the experts differ. This may have consequences for the appropriateness of different procedures for consulting experts. The usual type of problem in risk analysis can be considered to belong to the class of expert problems [11]. Consequently, the expert problem is the focus of this paper, where the roles of the person(s) responsible for the analysis and of those who are asked for advice are separated by referring to the analyst(s) and the experts.

3. The need for structured expert judgment

When uncertainty is represented in the form of subjective probability, then the use of expert judgment implies asking experts for probabilistic statements. This data source, however, has certain special characteristics, and these should

be addressed in procedures for the use of expert judgment. The most important characteristics are discussed in this section.

3.1 Spread

Expert assessments in risk analysis typically show a large spread, and thus an expert judgment methodology should address the issue of spread. A vivid example of this issue is found in the *Reactor Safety Study* (RSS) [15], which can be considered as the first "modern" probabilistic risk analysis and which made extensive use of expert subjective probability assessments: the estimates of the failure rate of high quality steel pipe of diameter ≥ 3 in. range from 5×10^{-6} to 1×10^{-10} (thirteen responses). Since the RSS, expert judgments have been used in various areas of application, and the existence of a substantial spread of opinion is a recurrent theme across all these applications: the analysis of nuclear risks [17], the analysis of seismic risks [18], and the analysis of health risks due to air pollution [19].

Ideally, an expert judgment methodology should be reproducible: it should be independent of the analyst performing the study. Reproducibility of results is related to the issue of spread, and has recently been investigated in various benchmark exercises in which independent teams of experts analyze the same system. Examples are a benchmark study on systems reliability organized by the Joint Research Centre of the Commission of the European Communities [20,21], and a benchmark study in the field of human reliability [22] on a draft version of the *Handbook of Human Reliability Analysis* [23]. A typical finding of these benchmark exercises is that both modelling and data uncertainties exist although modelling uncertainties may overwhelm uncertainties in the data.

3.2 Dependence

Expert judgments are likely to be dependent, and thus a methodology for using expert judgment should address the issue of dependence. Dependence may be encountered on different levels as indicated in [11]:

(1) Experts will usually share a common knowledge base. This type of dependence is called knowledge dependence [24,25].

(2) Experts may cluster into "optimists" and "pessimists": an expert who is optimistic (pessimistic) on one item may also be optimistic (pessimistic) on other items. See [9] for a further discussion.

(3) Experts' judgments may correlate with parameters reflecting common interests. This type of dependence might be called motivational dependence.

3.3 Calibration and information

One of the most important issues in using expert judgment is whether the experts' assessments are good. Winkler and Murphy [26] address this question extensively, defining two kinds of "goodness": normative goodness, relat-

ing to probabilistic considerations, and substantive goodness, relating to the assessor's knowledge. With reference to [26], Morgan et al. [27] distinguish four criteria for evaluating probability assessments: consistency, coherence, information and calibration. Probability assessments are consistent if they do not vary with the method used nor over time (unless the expert gets new information). They are coherent if they obey the laws of probability theory (which should be the case). Assessments should be informative concerning the true values of the events or variables in question. Finally, calibration is a measure of the degree of correspondence with reality: in the long run, the assessed probabilities should equal the actual frequencies of occurrence. Thus, a good probabilistic assessor gives consistent, coherent, informative and well calibrated probability judgments.

Since quantitative measures have been defined for the notions of calibration and information [28], the question comes up whether probability assessments in practice are good. Many experimental studies on the quality of expert judgment have been reported in the psychological literature. The general finding of these studies is that probability assessors are badly calibrated and show a significant degree of overconfidence (which is called the overconfidence bias: the uncertainty bands are too narrow) [29]. A second bias that often occurs is the location bias: the estimates are shifted to higher or lower values. Much research has been done into the underlying processes leading to biased assessments [30]. However, two problems have been identified concerning most of the experiments which have been performed to study the quality of probability assessments. The first problem is that many of these experiments took place in a laboratory situation, which has been criticized as being artificial [31]. The second problem is that in most cases the subjects were non-experts. The question then is whether experts are better probability assessors than non-experts; as Cooke et al. [28] note, the evidence is mixed on whether experts are better calibrated than non-experts. The next question then is: How about expert performance in risk analysis? Some examples are available, and these are briefly discussed below.

Some years after the publication of the *Reactor Safety Study* [15], sufficient operational experience with nuclear reactors became available to allow for a comparison with a number of RSS estimates for components and subsystems. This was done by Apostolakis et al. [32] by interpreting the RSS probability distributions as population variability or generic curves, reflecting plant-to-plant variations, and by using these distributions as prior distributions which were combined with statistical evidence in a Bayesian updating procedure. A similar approach was followed with failure rates obtained from another source of generic data: IEEE Standard-500 [33]. It appeared that in two of the three cases which were considered the posterior distributions were shifted to higher failure rates when compared with the initial prior distributions. It was concluded that the initial distributions might have been biased. This assumption

is supported by an analysis of data on operational experience which have been collected by the Oak Ridge National Laboratory [34,35]. For seven subsystems, Cooke [9] compared the failure frequencies as estimated on the basis of operational experience with the 90% confidence bounds as used in the RSS. It appeared that all the values from operating experience fall outside the RSS confidence bounds. Both analyses suggest that the RSS analysts are badly calibrated, with the occurrence of both a location bias (estimates too low) and an overconfidence bias (confidence bounds too narrow).

Snaith [36] and Mosleh et al. [7] compared observed and predicted values of reliability and maintenance parameters. Both studies show that the ratio of the observed to the predicted values is between $\frac{1}{4}$ and 4 for the majority of the predictions. In addition, Mosleh et al. also looked at the range factors, which indicate the degree of confidence in the predicted values (the range factor is defined as the square root of the ratio of the 95th and 5th percentiles of a log–normal distribution). They found that expert-estimated range factors are generally two to four times smaller than the observed range factors, which clearly indicates overconfidence.

4. Procedures in applying structured expert judgment

As discussed in the previous section, expert judgment is a data source with a number of special characteristics. Consequently, the appropriate use of this source requires a structured approach and the development and introduction of specialized methodological tools, in order to enhance the reproducibility and quality of the data obtained, leave an audit trail and build rational synthesis. Given this need for a structured approach to expert judgment, various methodological tools have been developed. These tools are introduced in this section, in connection with the key ingredients of an expert judgment process. Roughly speaking, the following ingredients can be distinguished:

(1) Problem analysis,
(2) Selection of experts,
(3) Elicitation of judgments,
(4) Processing and analysis, and
(5) Documentation and communication.

It should be noted that the above ingredients do not represent a rigid sequence of steps. In fact, an actual expert judgment process will usually be iterative in character and may also involve more specific steps, depending upon the complexity of the problem. For example, it may be more appropriate in a particular case to divide elicitation into two distinct ingredients by separating the training of experts from the actual data collection. Thus, Hora and Iman [37] describe a ten-step process. The above ingredients, however, are considered to be sufficient for introducing and discussing the various aspects of an expert judgment process.

4.1 Problem analysis

Elements which are associated with the particular problem under consideration appear in several stages of the process. Initially, problem analysis refers to the identification and selection of those issues, events or variables for which it is necessary to use expert judgment. At a later stage of the process, it refers to the formulation of questions; then it may also involve expert participation.

An important aspect in problem analysis is problem decomposition. Both Mosleh et al. [7] and Hora and Iman [37] underline the value of problem decomposition and argue that decompositions which are eventually expert-defined (initial decompositions may be proposed by the analyst) tend to improve the quality of assessments and the level of expert satisfaction. Ravinder et al. [38] provide a more general discussion of the use of decomposition from a psychometric measurement perspective.

4.2 Selection of experts

The selection of experts is an essential ingredient of any expert judgment process. Two steps can be distinguished: the identification of potential experts, and the eventual choice of experts. *Identifying* potential experts is a task for which little guidance exists in the literature. Important questions concerning the *choice* of experts include "How many experts must be consulted?" and "How does one choose between experts?" (provided that such a choice is necessary). Again, little guidance exists. It should be noted that, in practice, until now the selection of experts has depended more on practical considerations such as their geographical location and the availability of time and money than on matters directly related to their expertise [11].

4.3 Elicitation of judgments

A variety of procedures exists for the elicitation of judgments. These procedures differ in the following respects: the design for organizing the experts, the actual elicitation technique, and the "philosophy" for "dealing with" calibration.

Perhaps the most important aspect of the elicitation of judgments from a group of experts is the way in which the experts are organized. Various *designs* are possible, each with its own strenghts and weaknesses. In introducing these designs, it is also appropriate to indicate the method of aggregation, since these two aspects are not independent. Here it is sufficient to distinguish between behavioral, judgmental and mathematical aggregation; a more detailed discussion is given in the next subsection. The main designs are the following (see also [11]):

(1) The experts do not meet, but respond to questionnaires. The responses are analyzed and the results of this analysis are sent back to the experts, usually anonymously, upon which they may revise their original responses. The revised responses are analyzed and the whole process is iterated until some de-

gree of consensus is obtained. This approach is characteristic of the Delphi method [39]; see [9] for a critical discussion. The synthesis of judgments can be characterized as structured behavioral aggregation [7].
(2) The experts do not meet, but interact individually with the analyst. The synthesis of judgments is performed by the analyst and can be characterized as judgmental or mathematical aggregation.
(3) The group of experts meet and interact with each other, and produce consensus judgments. This approach is described by Kaplan [40]. The synthesis of judgments can be characterized as unstructured behavioral aggregation [7].
(4) Several independent teams of experts analyze the same problem. Each team produces consensus judgments, but the teams do not interact with each other. The synthesis of judgments within the various teams can be characterized as unstructured behavioral aggregation; the synthesis of the team judgments is performed by the analyst and can be characterized as judgmental or mathematical aggregation.

Whereas the above approaches are the main alternatives, variants or mixtures are also conceivable, such as a joint discussion followed by individual elicitation (synthesis by analyst), or individual elicitation followed by joint discussion (synthesis within group).

The *actual elicitation* of judgments takes the form of interaction between analyst and group of experts, or between analyst and individual experts. In the case of analyst–expert interaction, two alternatives exist: interaction through a questionnaire, or direct interaction. There have been many surveys of techniques for eliciting subjective probabilities from individual experts [8,9,41–43].

Finally, two alternative "approaches" exist for "*dealing with*" *calibration*, given the frequently observed poor quality of probability assessments (see Section 3.3). One approach focuses on the elicitation process and puts a lot of emphasis on training the experts in making probabilistic judgments (e.g. by making them aware of potential biases and of the underlying mechanisms leading to biases); in this case, the actual elicitation of judgments usually takes the form of direct analyst–expert interaction [37,43]. The other approach involves quantifying the experts' calibration and using this quantification in the aggregation of the judgments (see Section 4.4).

4.4 Processing and analysis

Processing involves the evaluation and combination (aggregation) of judgments. In Section 4.3, three types of aggregation are distinguished: behavioral aggregation (structured or unstructured), judgmental aggregation and mathematical aggregation. The latter two are necessary when the actual elicitation leads to individual experts' assessments and when the eventual analysis requires aggregated distributions; they are performed by the analyst. Since judgmental aggregation is considered to be objectionable, this subsection focuses

on methods for mathematical aggregation. Three main categories of methods for mathematical aggregation exist:

(1) Weighted averaging. This method involves assigning weights to the experts and applying these weights to the assessments given by the experts. The most widely known variants of weighted averaging are the linear opinion pool, based upon arithmetical averaging, and the logarithmic opinion pool, based upon geometrical averaging; see [9,16,44] for further discussions. The weights can be obtained in a variety of ways: from self-ratings, colleague ratings, ratings by analyst, etc. Recently, Cooke has developed a theory of weights which are based on expert performance and which reward both good calibration and high information [9,45–47]. The so-called classical model which applies this theory requires the formulation of calibration variables, of which the true values are known (or will become known) to the analyst but not to the experts. Calibration variables should resemble the actual variables of interest as much as possible.

(2) Bayesian models. These models require that the analyst supplies prior probability distributions; processing then involves updating these distributions via Bayes' theorem. This theorem essentially provides a mechanism for updating a particular state of knowledge when new information, e.g. expert judgments, becomes available. Bayesian models which require judgmental input from the analyst or decision maker have been proposed by Mosleh and Apostolakis [48,49], whereas Mendel and Sheridan [50] propose a model which makes use of calibration variables; see [9] for further discussions.

(3) Paired comparisons models. In these models, developed within the area of psychological scaling, experts are asked to compare objects in pairs and to indicate, for each pair, their preference for one of the objects concerning the attribute under investigation; the attribute may be "probability of occurrence". This is done for a number of objects, and usually with a relatively large number of experts. The responses of the experts are processed in order to pick up the underlying trend in the comparisons and obtain values for the objects considered. This processing is performed in two steps: (a) using modelling assumptions, scale values for the attribute in question are derived, and (b) using reference values, the scale values are transformed into absolute values. Various models are available for this processing; see [9,45,47] for further discussions.

Finally, *analysis* involves a critical evaluation of the results obtained from aggregating the experts' assessments. Preferably, this evaluation should be followed by extensive feedback to the experts who participated in the process.

4.5 Documentation and communication

Documentation involves reporting both the expert judgment process as a whole and the final results. Depending upon the context, it may be appropriate to prepare an intermediate report documenting the first two ingredients (problem analysis and selection of experts).

Communication is related to the interaction between the analyst(s) and management, with the analyst(s) being responsible for the expert judgment process and management being responsible for the decision making process.

5. Practical experiences

To date little experience exists with the use of structured expert judgment in risk analysis for the chemical process industries. However, expert judgment has been used in a structured form as a source of data in many studies in related areas. Relevant examples are several risk studies of nuclear power plants [17,37,51], several studies of seismic risk in connection with nuclear safety [18,52,53], the analysis of health risks due to air pollution [19], risk analysis of spaceflight systems [54], and a number of studies in the area of reliability engineering and maintenance management [55–57]. Perhaps the most prominent area of applying expert judgment is the assessment of human error probabilities [23,58]. The experiences in all these areas are certainly of value for future applications of structured expert judgment in risk analysis for the chemical process industries. Some of the above experiences are briefly discussed below; furthermore, one application is described in more detail at the end of this section.

5.1 European experiences

An inventory of experiences in Europe has recently been made by a Project Group on Expert Judgment, in which a number of European organizations are represented, within the context of ESRRDA (the European Safety and Reliability Research and Development Association). The results of this work are documented in an ESRRDA report [11]. Altogether, 15 experiences were identified; the report contains one-page summaries of all 15 experiences, and more detailed descriptions of six of these. This subsection presents short descriptions of experiences which were obtained in five of the organizations represented in the project group.

In 1982, the Gesellschaft für Reaktorsicherheit in the Federal Republic of Germany conducted a survey of expert opinion within the framework of the risk-oriented analysis for the German fast breeder reactor at Kalkar [17]. The objective of this survey was to obtain a probability distribution for the work energy release caused by an unprotected loss of flow accident. Experts from 18 organizations (in five countries) involved in fast breeder reactor safety participated in the survey. Elicitation was done by means of a comprehensive ques-

tionnaire. The questions were grouped into five categories, and allowed for consistency checks; the experts were also asked to quantitatively assess on a given scale both their own and the other experts' familiarity with the subject matter of each of the five categories of questions. The complete questionnaire is included in [17]. The answers were processed in two steps. Firstly, individual cumulative distribution functions for the work energy release were derived from the answers in the individual questionnaires. Next, aggregation was performed by weighted averaging, where the weights were derived from the self- and colleague-ratings given by the experts. It was concluded that there exists substantial variation between experts. Among the recommendations are more opportunity for interaction between analysts and experts, inclusion of questions concerning the potential for dependencies between uncertainties, and presenting not only the aggregated result but also the individual results of such surveys [11].

The Safety and Reliability Directorate of AEA Technology in England has been involved in describing and assessing techniques for the quantification of human error probabilities [58]. One of the applications concerned a loss of coolant nuclear accident. The objective of the exercise was to assess human error probabilities associated with tasks which have to be performed in order to maintain core cooling after such an accident. Most of the six techniques which were applied are designed specifically to quantify human error. Altogether, six experts participated in the exercise: one manager, four plant operators and one ergonomist. In applying one of the techniques, task probabilities were derived with a group consensus method involving three experts. The experts were first asked to give their own probability estimates; the individual experts' estimates were then discussed among the participants until consensus values were obtained for each task probability. The recommendations include the use of more structured processes for describing the problems to be quantified and for elicitation of the human error probabilities [11].

At the Koninklijke/Shell Laboratory Amsterdam in the Netherlands, expert judgment has been used in the context of maintenance optimization. The objective of this application was to obtain component lifetime distributions required for a decision support system which optimizes preventive maintenance. Altogether, 15 experts were consulted: maintenance technicians and supervisors, all with several years of experience with the equipment in question. Two different elicitation methods were used, with at most five experts per component. In a first phase, elicitation was done by means of two subsequent questionnaires. The second questionnaire was necessary in order to resolve inconsistencies which appeared in the responses to the first questionnaire. It was concluded that the use of a questionnaire is not advisable, since people are not motivated to give answers and accordingly the reliability of the data may be low. Thus, an analyst guiding the elicitation is necessary to explain the ques-

tions and ensure a good procedure. Consequently, in a second phase, a PC program for the elicitation was used, in order to automate the analysis and give feedback to the experts. The use of the program turned out to be a success. The experts were enthusiastic about receiving feedback, and the time required for elicitation was reduced considerably. In both phases, aggregation was performed by engineering judgment. It was recommended to develop formal methods, in particular, for combining and updating of expert judgment [11].

Four case studies were performed as part of an extensive project for the Dutch government, carried out jointly by Delft University of Technology and TNO in the Netherlands [59]. After an extensive literature review [5,8], three model-types were developed and/or made operational: the classical model, the Mendel–Sheridan Bayesian model and paired comparisons models [9,45–47]; see also the discussion in Section 4.4. The four case studies (see [60] for a short summary) can be characterized as follows:

(1) A case study at European Space Research and Technology Centre (ESTEC), discussed in more detail in Section 5.4, concerned failure rates of basic events in the fault tree of a spaceflight propulsion system [54,61]. Four propulsion and reliability experts were consulted. Aggregation was performed with the classical model and the Mendel–Sheridan Bayesian model.

(2) A case study at N.V. Nederlandse Gasunie concerned failure rates of two different components of gas pressure regulators [55]. The experts were maintenance personnel: 21 mechanics and 6 supervisors. Aggregation was performed with the paired comparisons models.

(3) A case study at XYZ concerned contamination causes and frequencies in a pilot plant-scale fermentation process [56]. Altogether, 11 experts were consulted: one from the engineering department, the others from pilot plant personnel and laboratory personnel. Aggregation was performed with the paired comparisons models.

(4) A case study at DSM Limburg B.V. in the Netherlands concerned relative contributions of failure causes of flanged connections in a chemical process plant [57]. Altogether, 14 experts participated: operators, mechanics, maintenance engineers, mechanical engineers and their chiefs. Aggregation was performed with both the paired comparisons models and the classical model.

In all four cases, elicitation was done by means of questionnaires; an analyst was always present during the elicitation process, and the experts were interviewed individually, whenever possible. The paired comparisons models appeared to provide an effective tool for "consensus building" with regard to the ranking of objects; transformation of scale values to absolute values must be approached with reservation. The classical model seemed promising for producing absolute values, whereas the Mendel–Sheridan Bayesian model only seemed to work well when applied to individual experts. The recommendations include further application to "real world" problems, and the development of software support [59].

5.2 Experiences in the USA

Mosleh et al. [7] present a critical review of the elicitation and use of expert opinions in probabilistic risk assessment of nuclear power plants. They describe four case studies, one of which is concerned with the assessment of seismic hazard rates. This particular case study focuses on a recent seismic hazard study which was performed by the Electric Power Research Institute (EPRI), Palo Alto, CA in the U.S.A. [53]. The EPRI seismic hazard study used six independent teams of experts. Within each team, all relevant disciplines of knowledge were represented. Elicitation and consensus aggregation happened within the various teams which were allowed to use their own decompositions. There was no interaction between the teams, and final aggregation of the results provided by the different teams was done mathematically, using equal weights, in order to yield a distribution for the likelihood of seismic activity. Thus, the benefits of mathematical aggregation were combined with the use of multidisciplinary teams. The results of the EPRI study are reported to indicate that the variation among teams is a significant contributor to the overall uncertainty. Mosleh et al. are very positive about both the use of expert-defined decompositions and the multiple-team approach which preserves independence among teams.

NUREG-1150 [62] can be considered as a major update of the *Reactor Safety Study* [15]. One of its purposes is to present a picture of current nuclear reactor risks in the U.S.A. The analysis involves complete probabilistic risk assessments of five nuclear power plants. The expert judgment methodology which was used in the first draft of NUREG-1150 drew substantial criticism. A revised methodology was developed and applied [63]; Hora and Iman [37] present a short summary. The revised methodology involved a multiple-panel approach in which each panel studied a particular problem area, and consisted of a ten-step process, which was implemented in a three-meeting format. The first three steps (selection of issues, selection of experts and preparation of issue statements) were performed prior to the first meeting. The first and second meeting were devoted to step 4 (elicitation training) and step 5 (presentation of issues), respectively. Step 6, preparation of analyses by the experts, was performed between the second and third meeting. The third meeting was devoted to steps 7 and 8, discussion of analyses and the actual elicitation. The latter was done by means of individual assessment meetings between each expert and a team of two analysts: one normative analyst, being expert in the field of probability assessment, and one substantive analyst, being expert in the particular problem area. Step 9, recomposition and aggregation, and step 10, review by the panel of experts, were performed after the third meeting. Recomposition of the probability distributions given by the individual experts led to their probability distributions for the quantities in question; aggregation of the individual experts' recomposed distributions was done by simple averaging of probabilities. Hora and Iman describe an example of an issue, where

the resulting aggregated distribution is shown to capture the diversity of viewpoints and the inherent uncertainty. In conclusion, they report that the process is believed to be successful by those involved.

5.3 Investigation of expert judgment method applications

Recently, expert judgment method applications in research and consultancy organizations and in industry were investigated as part of a project for the European Space Agency [64]. This investigation was carried out by selecting a number of users of expert judgment methods, both in Europe and in the U.S.A., and by interviewing these users according to a standardized interview format. This interview format consisted of two parts. The first part was directed to getting an understanding of the various aspects which are associated with the actual use of expert judgment, such as the ways in which experts are selected and the methods used for eliciting their judgments and for processing and evaluating the responses. The second part was concerned with more specific issues, such as the requirements for using expert judgment and the possibilities for evaluating the quality of judgments. Interviews were conducted with sixteen individuals from consultancy, industry and university; most participants have been involved with applications in risk assessment and related fields, such as maintenance and reliability analysis. Among the conclusions are a need for more formalized procedures for selecting/screening experts and the importance of traceability of the expert judgment data flow. There was no firm consensus regarding the reporting of expert judgment data so as to render a full scientific review possible: the question of anonymity divided opinions. The recommendations include the development of formalized procedures for selection of experts, for elicitation and processing of judgments (addressing calibration), for recording of information and for feedback to experts, and the development of formalized procedures for defining degrees of access to expert judgment data.

5.4 Example: application at ESTEC

This subsection describes in more detail an application at the European Space Research and Technology Centre in Noordwijk, Netherlands. It is considered to be appropriate for more detailed discussion, since it is concerned with fault tree quantification. The application in question was performed by Cooke, and is reported by Cooke [54] and Preyssl and Cooke [61]; this subsection is derived from [54].

As part of the risk analysis of a spaceflight propulsion system, a fault tree analysis was performed, in order to determine the frequency distribution of the event "loss of life as a consequence of system failure". The failure rates of 35 basic events in the fault tree had to be assessed. To this end, four experts were selected for consultation, who all had a high degree of technical expertise and mathematical sophistication. In addition to the actual variables of interest, 13

calibration variables were defined: variables which resemble the variables of interest as much as possible and of which the true values are known to the analysts but not to the experts.

The experts were interviewed individually. They were asked to give both a "best (i.e. median) estimate" and a "degree of confidence in the best estimate" for all 48 variables, using one assessment form per variable. The assessments could be given on either a quantitative or a qualitative scale; it was explained that only the quantitative scale would be used in the risk analysis. "Degree of confidence" was interpreted probabilistically as "degree of surprise" at finding the true value at least a factor of 10 higher than the median assessment. Assuming the uncertainty distributions for the failure rates to be log–normal, the degree of surprise can be used to determine the error factor, that is the factor by which the median must be multiplied (divided) to determine the 95% (5%) confidence bound. During the elicitation, a chart relating degrees of surprise to error factors was available to the experts. Each elicitation session lasted about an hour. At least one analyst was present at all sessions.

Aggregation was performed by weighted averaging, for which the classical model was used. The Mendel–Sheridan Bayesian model was also used, but this only seemed to work well when applied to individual experts. Thus, the discussion here is limited to the classical model.

Using the classical model, the experts' weights were derived from their assessments of the 13 calibration variables. With these weights, distributions for the 35 variables of interest were determined for the decision maker: the weighted combination of the experts. Feeding these distributions into the fault tree led to the median and 90% confidence bounds of the frequency distribution of the top event for the decision maker (see Table 1). Also shown are the median and 90% confidence bounds resulting from feeding the individual experts' assessments into the fault tree. The median assessment of the decision maker agrees with that of expert 3, but the decision maker's confidence bounds are narrower by an order of magnitude than those of expert 3. If the experts had input their

TABLE 1

Results of fault tree quantification using the decision maker's and the experts' assessments (classical model)

Expert	Top event		
	5%	Median	95%
Decison maker	9×10^{-6}	1×10^{-4}	1×10^{-2}
1	5×10^{-6}	8×10^{-5}	3×10^{-3}
2	1×10^{-7}	2×10^{-6}	3×10^{-5}
3	9×10^{-7}	1×10^{-4}	1×10^{-2}
4	4×10^{-5}	2×10^{-4}	1×10^{-3}

TABLE 2

Results of quantifying a previous fault tree using previous data and the decison maker's assessments (classical model)

	Top event			
Previous data			60% Confidence 5×10^{-6}	
	5%	50%	60%	95%
Decision maker	4×10^{-6}	3×10^{-5}	6×10^{-5}	5×10^{-4}

distributions into the fault tree individually, the resulting median assessments would span 2 orders of magnitude and the confidence bounds would span 5 orders of magnitude. Other results of the classical model are presented in [54].

The results of the above application of the classical model were compared with the results of a previous study of the same system, which used a simpler fault tree and different expert assessments (see Table 2). The "60% confidence" attached to the result 5×10^{-6} was interpreted as meaning that 5×10^{-6} is the 60% quantile of the uncertainty distribution for the top event. Feeding the decision maker's distributions of the above exercise into the simpler fault tree led to a 60% quantile of 6×10^{-5}.

The main conclusions which were drawn from this application are the following [54]:

(1) The classical model proved easy to apply and led to meaningful results.

(2) The definition of a sufficient number of meaningful calibration variables proved much easier than initially expected.

(3) Time constraints precluded a preliminary training session, which was felt to be unfortunate.

(4) The time required for data collection was limited. The method was highly appreciated by the experts who preferred graphic input to giving numerical assessments; this considerably speeds up the elicitation process. Giving both qualitative and quantitative scales also speeds up the elicitation.

(5) The presence of an analyst during all elicitation sessions is absolutely essential, whereas it is also essential that the experts be interviewed individually.

6. Discussion

Since the quantitative phases of risk analysis are, and will remain to be, characterized by a lack of data (or by insufficient/inappropriate data), the judgments of experts will remain a necessary and inevitable source of data in performing such analyses. The recognition of the need for a structured use of this data source has led to many methodological developments and applica-

tions. Nevertheless, as Mosleh et al. [7] conclude, there still exists a lot of reliance on the common sense of the substantive experts involved in the analyses, and this certainly holds true for risk analysis in the chemical process industries. Two main obstacles are considered to be responsible for this situation, and these are concerned with effort/costs and validation. The effort and costs involved in applying structured expert judgment are still relatively large, and, to date, there still exists insufficient proof of the "goodness" of the results of structured expert judgment processes in real-world applications. Consequently, future work should focus on the development of more cost-effective techniques and on the validation of expert judgment techniques for all aspects of a structured expert judgment process: problem analysis, selection of experts, elicitation of judgments (including training of experts), processing and analysis, and documentation and communication. Moreover, the focus should be on applications on real-world problems: it is emphasized by the ESRRDA Project Group on Expert Judgment that only by *applying* formalized methods for using expert judgment the value of a structured approach will be proved and the necessary development will receive the appropriate stimuli [11]. An example of a recent application is a pilot-scale expert judgment study on parameters in atmospheric dispersion models, organized by the Commission of the European Communities.

More specifically, the following points are of interest in future research and development (see also the discussion in [65]):

(1) Problem analysis. Problem decomposition, in particular the use of an expert-defined decomposition, can be considered to be an effective strategy for eliciting expert judgment. More research is needed to determine the effectiveness and optimum level of decomposition in general [7].

(2) Selection of experts. Guidelines for both the identification and the choice of experts are urgently needed, in particular for situations in which the geographical spread of the experts is large.

(3) Elicitation of judgments. Apart from the need for software tools which should make the process more cost effective, key issues of interest are the impact of training, the elicitation design, and the treatment of dependencies in elicitation.

(4) Processing and analysis. Key issues of interest are the method of aggregation, the integration of objective and subjective information sources (both during and after the process), and the treatment of dependencies. The question of how to "deal with" calibration should also be addressed: emphasis on training, emphasis on quantifying calibration, or a combination of these?

(5) Documentation and communication. Since traceability of data is considered to be an important issue and the association of experts' names to assessments appears to be controversial [64], key issues of interest are the degree of access to expert judgment data and the accountability of experts.

In conclusion, it should be noted that the variety of procedures, which this paper demonstrates to exist, is not only related with differences in methodological and philosophical viewpoints, but also with the variety of problem characteristics which appears in practical situations. Important characteristics are the degree of complexity, the availability of experts, the number of assessments which have to be made, and the availability of time and resources. Thus, different techniques may be appropriate in different situations. This observation underlines the need for further development in the context of practical situations.

Acknowledgments

This paper has benefited greatly from a recent review by the ESRRDA Project Group on Expert Judgment [11], and in particular from the contributions of Roger Cooke and Simon French to that review. The author is also grateful to Dr C. Preyssl of the European Space Agency (ESA) for permission to discuss the investigation of expert judgment method applications carried out under contract with ESA.

References

1 J. Suokas and R. Kakko, On the problems and future of safety and risk analysis, J. Hazardous Mater., 21 (1989) 105–124.
2 A. Mosleh, Hidden sources of uncertainty: Judgment in the collection and analysis of data, Nucl. Eng. Des., 93 (1986) 187–198.
3 E. Hofer, On surveys of expert opinion, Nucl. Eng. Des., 93 (1986) 153–160.
4 J.F.J. van Steen, Expert opinion use for probability assessment in safety studies: Main topics and elements of an application-oriented research program, Eur. J. Operation. Res., 32 (1987) 225–230.
5 J.F.J. van Steen, Expert opinion in probabilistic safety assessment, In: G.P. Libberton (Ed.), 10th Advances in Reliability Technology Symposium, Elsevier Applied Science, London, 1988, pp. 13–26.
6 C.A. Clarotti and D.V. Lindley (Eds.), Accelerated Life Testing and Experts' Opinions in Reliability, Elsevier North-Holland, Amsterdam, 1988.
7 A. Mosleh, V.M. Bier and G. Apostolakis, A critique of current practice for the use of expert opinions in probabilistic risk assessment, Reliability Eng. Syst. Saf., 20 (1988) 63–85.
8 J.F.J. van Steen and P.D. Oortman Gerlings, Expert Opinions in Safety Studies, Vol. 2: Literature Survey Report, Delft University of Technology/TNO, Delft/Apeldoorn, 1989.

9 R.M. Cooke, Experts in Uncertainty: Expert Opinion and Subjective Probability in Science, Department of Mathematics, Delft University of Technology, Delft, 1989 (appearing from Oxford University Press).

10 O. Svenson, On expert judgments in safety analyses in the process industries, Reliability Eng. Syst. Saf., 25 (1989) 219–256.

11 ESRRDA Project Group "Expert Judgment", Expert Judgment in Risk and Reliability Analysis: Experiences and Perspective, ESRRDA Report No. 2, European Safety and Reliability Research and Development Organization, Commission of the European Communities, Ispra, 1990.

12 D.V. Lindley, Scoring rules and the inevitability of probability, International Statistical Review, 50 (1982) 1–26.

13 S. French, Decision Theory: An Introduction to the Mathematics of Rationality, Ellis Horwood, Chichester, 1986.

14 G.E. Apostolakis, F.R. Farmer and R.W. van Otterloo (Eds.), The interpretation of probability in probabilistic safety assessments, Reliability Eng. Syst. Saf., 23 (1988) 247–320.

15 U.S. NRC, Reactor Safety Study (WASH-1400), Report NUREG 75/014, U.S. Nuclear Regulatory Commission, Washington, DC, 1975.

16 S. French, Group consensus probability distributions: A critical survey, In: J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith (Eds.), Bayesian Statistics 2, Elsevier North-Holland, Amsterdam, 1985, pp. 183–201.

17 E. Hofer, V. Javeri, H. Löffler and D.F. Struwe, A survey of expert opinion and its probabilistic evaluation for specific aspects of the SNR-300 study, Nucl. Technol., 68 (1985) 180–225.

18 D. Okrent, A survey of expert opinion on low probability earthquakes, Ann. Nucl. Energy, 2 (1975) 601–614.

19 M.G. Morgan, S.C. Morris, M. Henrion, D.A.L. Amaral and W.R. Rish, Technical uncertainty in quantitative policy analysis: A sulfur air pollution example, Risk Anal., 4 (1984) 201–216.

20 CEC, Systems Reliability Benchmark Exercise, Nuclear Science and Technology Reports EUR 10696 EN/I and EUR 10696 EN/II, Commission of the European Communities, Luxembourg, 1986.

21 A. Amendola, Uncertainties in systems reliability modelling: Insights gained through European benchmark exercises, Nucl. Eng. Des., 93 (1986) 215–225.

22 R.L. Brune, M. Weinstein and M.E. Fitzwater, Peer Review Study of the Draft Handbook for Human Reliability Analysis, Report SAND82-7056, Sandia National Laboratories, Albuquerque, NM, 1983.

23 A.D. Swain and H.E. Guttmann, Handbook of Human Reliability Analysis, with Emphasis on Nuclear Power Plant Applications, Report NUREG/CR-1278, U.S. Nuclear Regulatory Commission, Washington, DC, 1983.

24 G. Apostolakis and S. Kaplan, Pitfalls in risk calculations, Reliability Eng., 2 (1981) 135–145.

25 R.M. Cooke and R. Waij, Monte Carlo sampling for generalized knowledge dependence with application to human reliability. Risk Anal., 6 (1986) 335–343.

26 R.L. Winkler and A.H. Murphy, "Good" probability assessors, J. Appl. Meteorol., 7 (1968) 751–758.

27 M.G. Morgan, M. Henrion and S.C. Morris, Expert Judgments for Policy Analysis, Report BNL 51358, Brookhaven National Laboratory, Upton, NY, 1979.

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

29 S. Lichtenstein, B. Fischhoff and L.D. Phillips, Calibration of probabilities: The state of the art to 1980, In: D. Kahneman, P. Slovic and A. Tversky (Eds.), Judgment under Uncertainty: Heuristics and Biases, Cambridge University Press, Cambridge, 1982, pp. 306–334

30 D. Kahneman, P. Slovic and A. Tversky (Eds.), Judgment under Uncertainty: Heuristics and Biases, Cambridge University Press, Cambridge, 1982.
31 R.L. Winkler and A.H. Murphy, Experiments in the laboratory and the real world, Organization. Behav. Human Perform., 10 (1973) 252–270.
32 G. Apostolakis, S. Kaplan, B.J. Garrick and R.J. Duphily, Data specialization for plant specific risk studies, Nucl. Eng. Des., 56 (1980) 321–329.
33 IEEE, Guide to the Collection and Presentation of Electrical, Electronic and Sensing Component Reliability Data for Nuclear-Power Generating Stations (IEEE Std.-500), Institute of Electrical and Electronics Engineers, New York, NY, 1977.
34 J. Minarick and C. Kukielka, Precursors to Potential Severe Core Damage Accidents: 1969–1979: A Status Report, Report NUREG/CR-2497, U.S. Nuclear Regulatory Commission, Washington, DC, 1982.
35 W. Cottrell and J. Minarick, Precursors to Potential Severe Core Damage Accidents: 1980–1982: A Status Report, Report NUREG/CR-3591, U.S. Nuclear Regulatory Commission, Washington, DC, 1984.
36 E.R. Snaith, The Correlation between the Predicted and Observed Reliabilities of Components, Equipment and Systems, Report NCSR R 18, National Centre of Systems Reliability, U.K. Atomic Energy Authority, Warrington, 1981.
37 S.C. Hora and R.L. Iman, Expert opinion in risk analysis: The NUREG-1150 methodology, Nucl. Sci. Eng., 102 (1989) 323–331.
38 H.V. Ravinder, D.N. Kleinmutz and J.S. Dyer, The reliability of subjective probabilities obtained through decomposition, Manag. Sci., 34 (1988) 186–199.
39 H.A. Linstone and M. Turoff (Eds.), The Delphi Method: Techniques and Applications, Addison-Wesley, Reading, MA, 1975.
40 S. Kaplan, "Expert Information" versus "Expert Opinions": Another Approach to the Problem of Eliciting/Combining/Using Expert Opinion in PRA, Pickard, Lowe and Garrick, Inc., Newport Beach, CA, 1988.
41 C.S. Spetzler and C-A.S. Staël von Holstein, Probability encoding in decision analysis, Management Science, 22 (1975) 340–358.
42 T.S. Wallsten and D.V. Budescu, Encoding subjective probabilities: A psychological and psychometric review, Manag. Sci., 29 (1983) 151–174.
43 M.W. Merkhofer, Quantifying judgmental uncertainty: Methodology, experiences and insights, IEEE Trans. Syst. Man Cybernet., 17 (1987) 741–752.
44 C. Genest and J. Zidek, Combining probability distributions: A critique and an annotated bibliography, Statist. Sci., 1 (1986) 114–148.
45 R.M. Cooke, J.F.J. van Steen, M.F. Stobbelaar and M. Mendel, Expert Opinions in Safety Studies, Vol. 3: Model Description Report, Delft University of Technology/TNO, Delft/Apeldoorn, 1989.
46 R.M. Cooke, Expert Opinions in Safety Studies, Vol. 4: A Theory of Weights for Combining Expert Opinion, Delft University of Technology/TNO, Delft/Apeldoorn, 1989.
47 J.F.J. van Steen and R.M. Cooke, Expert opinions as data source: Methods and experiences, In: V. Colombari (Ed.), Reliability Data Collection and Use in Risk and Availability Assessment, Springer-Verlag, Berlin, 1989, pp. 262–285.
48 A. Mosleh and G. Apostolakis, Models for the use of expert opinions, In: R.A. Waller and V.T. Covello (Eds.), Low-Probability/High-Consequence Risk Analysis, Plenum Press, New York, NY, pp. 107–124.
49 A. Mosleh and G. Apostolakis, The assessment of probability distributions from expert opinions with an application to seismic fragility curves, Risk Anal., 6 (1986) 447–461.
50 M.B. Mendel and T.B. Sheridan, Optimal Estimation Using Human Experts, Department of Mechanical Engineering, Massachusetts Inst. Technol., Cambridge, MA, 1987.

51 Steam Explosion Review Group, A Review of the Current Understanding of the Potential for Containment Failure from In-Vessel Steam Explosions, Report NUREG/CR-1116, U.S. Nuclear Regulatory Commission, Washington, DC, 1985.

52 D.L. Bernreuter, J.B. Savy, R.W. Mensing and D.H. Chung, Seismic Hazard Characterization of the Eastern United States: Methodology and Interim Results for Ten Sites, Report NUREG/CR-3756, U.S. Nuclear Regulatory Commission, Washington, DC, 1984.

53 EPRI, Seismic Hazard Methodology for the Central and Eastern United States, Vol. 1: Methodology, Report NP-4726, Electric Power Research Institute, Palo Alto, CA, 1986.

54 R.M. Cooke, Expert Opinions in Safety Studies, Vol. 5 (Case Report 2): ESTEC Case Study, Delft University of Technology/TNO, Delft/Apeldoorn, 1989.

55 M.F. Stobbelaar and J.F.J. van Steen, Expert Opinions in Safety Studies, Vol. 5 (Case Report 1): Gasunie Case Study, Delft University of Technology/TNO, Delft/Apeldoorn, 1989.

56 P.D. Oortman Gerlings, Expert Opinions in Safety Studies, Vol. 5 (Case Report 3): XYZ Case Study, Delft University of Technology/TNO, Delft/Apeldoorn, 1988.

57 M.F. Stobbelaar, R.M. Cooke and J.F.J. van Steen, Expert Opinions in Safety Studies, Vol. 5 (Case Report 4): DSM Case Study, Delft University of Technology/TNO, Delft/Apeldoorn, 1989.

58 Human Factors in Reliability Group, Human Reliability Assessors Guide, Report RTS 88/95Q, SRD, AEA Technology, Warrington, 1988.

59 L.H.J. Goossens, R.M. Cooke and J.F.J. van Steen, Expert Opinions in Safety Studies, Vol. 1: Final Report, Delft University of Technology/TNO, Delft/Apeldoorn, 1989.

60 J.F.J. van Steen, L.H.J. Goossens and R.M. Cooke, Protocols for expert opinion use in risk analysis, in: Proc. 6th Int. Symp. on Loss Prevention and Safety Promotion in the Process Industries, Oslo, Norway, European Federation of Chemical Engineering/Norwegian Society of Chartered Engineers, 1989, Vol. II, pp. 42.1–42.20.

61 C. Preyssl and R.M. Cooke, Expert judgment: Subjective and objective data for risk analysis of spaceflight systems, In: Proc. of the PSA '89 Int. Topical Meeting on Probability, Reliability and Safety Assessment, Pittsburgh, PA, ANS/ENS, 1989.

62 U.S. NRC, Reactor Risk Reference Document (Draft), Report NUREG/CR-1150, U.S. Nuclear Regulatory Commission, Washington, DC, 1987.

63 T.A. Wheeler, S.C. Hora, W.R. Cramond and S.D. Unwin, Analysis of Core Damage Frequency from Internal Events: Expert Judgment Elicitation, Report NUREG/CR-4550, Vol. 2, U.S. Nuclear Regulatory Commission, Washington, DC, 1989.

64 J.F.J. van Steen, R.M. Cooke and S. French, Investigation of Expert Judgment Method Applications, TNO/Delft University of Technology/University of Leeds, Apeldoorn/Delft/Leeds, 1990, Report prepared under contract with the European Space Agency and obtainable from ESTEC, Noordwijk, Netherlands.

65 C.E. Elderkin and G.N. Kelly (Eds.), Proc. CEC/DOE Workshop on Uncertainty Analysis, Report PNL-SA-18372/CONF-8911 195, Pacific Northwest Laboratory, Richland, WA, 1990.