Expert Judgement in Risk Assessment

Kelvin Leung
Simona Verga
CSS OR Team

DRDC CORA TM 2007-57
December 2007

Defence R&D Canada
Centre for Operational Research & Analysis

DRDC Centre for Security Studies

National Defence Défense nationale

Canada

Expert Judgement in Risk Assessment

Kelvin Leung
Simona Verga
CSS OR Team

Defence R&D Canada – CORA

Technical Memorandum
DRDC CORA TM 2007-57
December 2007

Principal Author

Kelvin Leung and Simona Verga

Centre for Security Science Operational Research Team

Approved by

R.M.H. Burton

Acting Section Head Joint & Common Operational Research

Approved for release by

Jocelyn Tremblay

Chief Scientist, DRDC CORA

The information contained here in has been derived and determined through best practice and adherence to the highest levels of ethical, scientific and engineering investigative principles. The reported results, their interpretation, and any opinions expressed therein, remain those of the authors and do not represent, or otherwise reflect, any official opinion or position of DND or the Government of Canada.

Defence R&D Canada – Centre for Operational Research and Analysis (CORA)

Abstract

Decision and risk analysis models often require both qualitative and quantitative assessments of uncertain events; in many cases, expert knowledge is essentially the only source of good information. Over the last decade, uncertainty analysis has become an increasingly important part of operations research models. The growing use of risk assessment in government and corporate planning and operations has also increased the role of expert judgement in providing information for decision making.

Elicitation of experts' opinions is frequently used to support decision making in many different areas, from forecasting in the financial world to assessing the risk of terrorist attacks in the national security domain. The use of expert judgements has provoked questions related to the practice of utilizing experts' opinions and to the accuracy of the obtained results. This work reviews some approaches for eliciting and aggregating expert judgements as inputs into the risk assessment process, and looks at methods of assessing the degree of confidence associated with these subjective inputs, as well as confidence in the overall process.

The research synthesized in this report outlines the elicitation process and highlights both its statistical and psychological perspectives. It looks at ways to evaluate the accuracy of elicitation; it presents techniques for the aggregation of probability distributions from multiple experts; and it summarizes a conceptual framework for the quality verification of risk assessment. Two examples of the application of formal elicitation in the nuclear industry and a business study are also discussed in the Appendix.

Résumé

Les modèles d’analyse des décisions et des risques exigent souvent des évaluations qualitatives et quantitatives d’événements incertains; dans de nombreux cas, la connaissance experte est essentiellement la seule source d’information valable. Au cours de la dernière décennie, l’analyse de l’incertitude a pris de plus en plus d’importance dans les modèles de recherche opérationnelle. Le recours croissant à l’évaluation des risques dans la planification et les activités gouvernementales et d’entreprise a également intensifié le rôle du jugement expert dans la prestation de l’information pour la prise de décision.

On fait souvent appel aux opinions d’experts pour appuyer la prise de décision dans divers domaines, de la prévision dans les milieux financiers à l’évaluation du risque d’attentat terroriste dans le contexte de la sécurité nationale. Le recours à des jugements experts soulève des questions concernant cette pratique et l’exactitude des résultats obtenus. Le présent document passe en revue quelques approches employées pour obtenir et regrouper des jugements experts dans le cadre du processus d’évaluation des risques, et il examine des méthodes permettant d’évaluer le degré de confiance lié à ces commentaires subjectifs ainsi que la confiance dans le processus global.

La recherche synthétisée dans le présent rapport décrit le processus d’obtention de jugements experts et fait ressortir ses perspectives statistique et psychologique. Elle considère des façons d’évaluer l’exactitude des opinions obtenues, présente des techniques pour regrouper les distributions de probabilités provenant de multiples experts et résume un cadre conceptuel pour vérifier la qualité de l’évaluation des risques. Deux exemples de l’application d’une méthode d’obtention formelle dans l’industrie nucléaire et une étude d’entreprise sont également traités dans l’annexe.

Executive summary

Expert Judgement in Risk Assessment:

Kelvin Leung, Simona Verga; DRDC CORA TM 2007-57; Defence R&D Canada – CORA; December 2007.

Introduction

This report reviews approaches for eliciting and aggregating expert judgements as inputs into the risk assessment process, and looks at methods for assessing the degree of confidence associated with these subjective inputs, as well as confidence in the overall process.

Elicitation of experts' opinions is frequently used to support decision making in many different areas, from forecasting in the financial world to assessing the risk of terrorist attacks in the national security domain. The use of expert judgements has provoked questions related to the practices of utilizing experts' opinions and to the accuracy of the obtained results; however, there are situations where little knowledge exists to base the assessment on, and, often, these situations are ones for which risk assessment is more imperative. In such cases expert judgement is the only source of good information.

When major decisions are to be made in the presence of substantial uncertainty, such as when planning for extreme events (dam collapse, earthquakes or terrorist attacks, to give just a few quite distinct examples), there rarely exist ways to calculate "objective" probabilities and expert judgement is deemed essential to minimize and characterize the uncertainty. Even for domains where data is available and there are grounds for objective analysis, aspects such as the interpretation of results or questions about acceptable risk levels underline normative, value-judgement perspectives that are unavoidable in risk-based decision making.

Results

The report presents a review of the various roles that expert judgement plays in risk assessment/management; it also outlines and comments on important stages, roles and models in the elicitation process.

A sound elicitation should be preceded by a thorough preparation. Identifying variables of interest, expert selection, motivation and training are critical steps towards a successful elicitation with accurate results. Also, a well-designed process has built-in flexibility and clear roles and rules, provides for structured interaction and incorporates feedback. Experts should understand what is expected of them and how their judgement will be used. Reviewing available evidence with the experts will provide a basis to draw upon when making judgements. Sensitivity analysis also plays an important role in validating the outcome.

The elicitation of subjective probabilities has been presented in this paper from both a psychological and a statistical perspective. Research in psychology shows that the human mind's ability to accurately judge probabilities is limited by biases and by the tendency to use heuristics. The facilitator should recognize the potential for inadvertently introducing biases into the

elicitation process and set up training procedures that explicitly encourage experts to think analytically, reviewing the most common biases and the reasons for their occurrence, and practice elicitation runs.

From a statistician’s perspective, the uncertainty related to the accuracy of elicited probabilities is an integral part of the analysis of expert judgement. Calibration curves are used to assess the agreement between the expert’s judgement and reality, and scoring rules measure how well the elicited probabilities express the expert’s underlying opinion. For improved accuracy, the experts should be provided with feedback and with the opportunity to review and improve their judgement. The facilitator should check that a set of subjective probabilities is coherent, or consistent with the laws of probability (i.e. the probabilities associated with all possible independent outcomes of an event sum to one).

Aggregation of probability distributions from multiple experts reduces random variation in a set of judgements. Aggregation methods fall under two general categories: behavioural aggregation and mathematical aggregation. In behavioural aggregation, an interaction is created between the groups of experts, through which a single distribution is elicited from the group as a whole. In mathematical aggregation, a single distribution is elicited from each expert individually and independently of the others and then the resulting distributions are mathematically combined into a single distribution. The combined distribution can be considered consensus in most circumstances and it is more feasible to use in further analysis. The use of one type of aggregation over the other is dictated by factors such as: the type of variables of interest; the type of available information; the experts’ background and experience; and the design and structure of the elicitation protocol. No single method is best in all circumstances, and an overall aggregation process could involve both mathematical and behavioural aspects.

To assess the confidence in the overall elicitation process, the report reviews briefly a conceptual framework for the quality verification of risk assessment. In essence, the qualification process is an independent review process that verifies certain quality characteristics in the risk assessment, such us completeness, credibility, transparency and fairness. This process aims to consolidate the decision-maker’s confidence in the results and recommendations of the risk assessment. The quality verification results may influence the decision-maker to either accept the risk assessment results or request refinements in the risk assessment based on the recommendations.

Significance of Results

This work represents an important part of the research on methodologies and models that the Centre for Security Science (CSS) OR team is doing to enhance the risk “toolbox” housed within the CSS Risk Portfolio. CSS is a joint endeavour with the Department of Public Safety Canada, extending Science and Technology development services, which DRDC traditionally provides in support of the Canadian Forces, to address national public safety and security objectives. Within CSS, a Risk Portfolio has been established in response to growing interest in the risk field from across government and defence. The vision is to develop a risk resource centre to support the community with threat, vulnerability and risk assessments, gap analysis, foresight and future security visioning and other related activities and products.

Kelvin Leung; Simona Verga; 2007; Expert Judgement in Risk Assessment; TM 2007-57 DRDC – Center for Operational Research and Analysis.

Sommaire

Expert Judgement in Risk Assessment:

Kelvin Leung, Simona Verga; DRDC CORA TM 2007-57; R & D pour la défense Canada – CORA; Décembre 2007.

Introduction ou contexte:

Le présent rapport passe en revue des approches employées pour obtenir et regrouper des jugements experts dans le cadre du processus d'évaluation des risques, et il examine des méthodes permettant d'évaluer le degré de confiance lié à ces commentaires subjectifs ainsi que la confiance dans le processus global.

On fait souvent appel aux opinions d'experts pour appuyer la prise de décision dans divers domaines, de la prévision dans les milieux financiers à l'évaluation du risque d'attentat terroriste dans le contexte de la sécurité nationale. Le recours à des jugements experts soulève des questions concernant cette pratique et l'exactitude des résultats obtenus. Toutefois, dans certaines situations, il existe peu d'information sur laquelle fonder l'évaluation, et c'est souvent dans ces occasions qu'une évaluation des risques est plus impérative. En pareils cas, le jugement expert constitue la seule source d'information valable.

Lorsque des décisions importantes doivent être prises en présence d'une incertitude considérable, par exemple lorsqu'il faut établir des plans en vue d'événements extrêmes (effondrement de barrage, tremblement de terre ou attentat terroriste, pour ne donner que quelques exemples assez distincts), il existe rarement des façons de calculer des probabilités « objectives », et le jugement expert est considéré comme essentiel pour réduire au minimum l'incertitude et la caractériser. Même dans les domaines où des données sont disponibles et où il y a des raisons d'effectuer une analyse objective, des aspects comme l'interprétation des résultats ou des questions touchant les niveaux de risque acceptables soulignent des perspectives normatives basées sur un jugement de valeur qui sont inévitables dans la prise de décision axée sur les risques.

Résultats:

Le rapport examine les divers rôles que le jugement expert joue dans l'évaluation et la gestion des risques; il traite également des étapes, rôles et modèles importants dans le processus d'obtention de jugements experts.

Une solide méthode d'obtention devrait être précédée d'une préparation approfondie. La détermination des variables d'intérêt, la sélection des experts, la motivation et la formation sont des étapes cruciales en vue d'une obtention réussie et de résultats exacts. Par ailleurs, un processus bien conçu offre une certaine souplesse, indique clairement les rôles et les règles, prévoit une interaction structurée et intègre la rétroaction. Les experts devraient comprendre ce qu'on attend d'eux et comment leur jugement sera utilisé. Le fait d'examiner les éléments de preuve disponibles en compagnie des experts servira de base pour formuler des jugements. L'analyse de la sensibilité joue aussi un rôle important dans la validation du résultat.

Nous avons présenté l’obtention de probabilités subjectives à la fois d’un point de vue psychologique et d’un point de vue statistique. La recherche en psychologie montre que la capacité de l’esprit humain de juger les probabilités avec exactitude est limitée par les préjugés et la tendance à recourir à des heuristiques. Le facilitateur devrait reconnaître la possibilité d’introduire par inadvertance des préjugés dans le processus d’obtention et établir des procédures de formation qui encouragent explicitement les experts à réfléchir de manière analytique, en examinant les préjugés les plus courants et les raisons de leur manifestation, puis effectuer des exercices d’obtention de jugements experts.

Du point de vue du statisticien, l’incertitude liée à l’exactitude des probabilités obtenues fait partie intégrante de l’analyse du jugement expert. Des courbes d’étalonnage sont utilisées pour évaluer l’accord entre le jugement de l’expert et la réalité, et des règles de cotation permettent de déterminer dans quelle mesure les probabilités obtenues expriment l’opinion sous-jacente de l’expert. Pour augmenter l’exactitude, il faudrait fournir une rétroaction aux experts et leur donner l’occasion de réviser et d’améliorer leur jugement. Le facilitateur devrait vérifier qu’une série de probabilités subjectives est cohérente ou conforme aux lois de la probabilité (c.-à-d. que la somme des probabilités associées à tous les résultats indépendants possibles d’un événement est un).

Le regroupement des distributions de probabilités provenant de multiples experts réduit la variation aléatoire dans une série de jugements. Les méthodes de regroupement entrent dans deux grandes catégories : le regroupement comportemental et le regroupement mathématique. Dans la première, une interaction est créée entre les groupes d’experts, interaction par laquelle une seule distribution est obtenue de l’ensemble du groupe. Dans le cas du regroupement mathématique, une seule distribution est obtenue de chaque expert séparément et indépendamment des autres, puis les distributions résultantes sont mathématiquement regroupées en une seule distribution. Celle-ci peut être considérée comme un consensus dans la plupart des cas et se prête davantage aux analyses ultérieures. L’utilisation d’un type de regroupement plutôt que l’autre est dictée par des facteurs tels que le type de variable d’intérêt, le genre d’information disponible, la formation et l’expérience des experts, ainsi que la conception et la structure du protocole d’obtention. Aucune méthode n’est meilleure qu’une autre dans toutes les circonstances, et un processus global de regroupement pourrait présenter à la fois des aspects mathématiques et des aspects comportementaux.

Afin d’évaluer la confiance dans le processus global d’obtention, nous examinons brièvement un cadre conceptuel pour vérifier la qualité de l’évaluation des risques. Essentiellement, le processus de qualification est un processus d’examen indépendant qui vérifie certaines caractéristiques de qualité dans l’évaluation des risques, comme l’intégralité, la crédibilité, la transparence et l’équité. Ce processus vise à renforcer la confiance du décideur dans les résultats et les recommandations de l’évaluation des risques. Les résultats de la vérification de la qualité peuvent amener le décideur à accepter les résultats de l’évaluation des risques ou à demander que cette dernière soit peaufinée en fonction des recommandations.

Importance:

Ce travail représente une partie importante de la recherche sur les méthodes et modèles que l’équipe de RO du Centre des sciences pour la sécurité (CSS) effectue afin d’améliorer la « boîte à outils » sur les risques qui se trouve dans le portefeuille de risques du CSS. Le CSS est une

initiative menée conjointement avec le ministère de la Sécurité publique du Canada, dans le prolongement des services de développement des sciences et de la technologie, que RDDC assure traditionnellement à l'appui des Forces canadiennes, afin d'atteindre les objectifs nationaux en matière de sécurité publique. Un portefeuille de risques a été établi au CSS en réponse à l'intérêt croissant pour le domaine des risques dans l'ensemble du gouvernement et des milieux de la défense. Le but consiste à mettre sur pied un centre de ressources sur les risques pour appuyer la collectivité dans le cadre des évaluations de la menace, de la vulnérabilité et des risques, de l'analyse des écarts, de l'établissement des prévisions et de la vision future en matière de sécurité, ainsi que d'autres activités et produits connexes.

This page intentionally left blank.

Table of contents

List of figures

List of tables

Acknowledgements

The authors wish to acknowledge extensive discussions with Dr. Allan Douglas through which they have gained important insights. The authors are also very grateful to Dr. Paul Chouinard for his valuable comments and support of the present work.

This page intentionally left blank.

1 Introduction

The ever-increasing call for “risk-based decision making” in essentially all aspects of our modern society, including the government setting, sometimes challenges approaches based on concepts, tools and methods that have been designed for more “traditional” areas like systems engineering, construction of physical infrastructures, or project management. And while “risk” as a measure of the probability and the severity of adverse effects is conceptually simple, quantification efforts may lead to confusion and misuse if improperly attempted.

One reason for the difficulty is the use of “probability”, which is a mathematical construct, intangible, yet omnipresent in risk-based decision making. The uncertainty that surrounds the measure of probability in risk assessment is particularly difficult to assess for rare and extreme events; however, situations where there is little knowledge to base the assessment on are also ones for which risk assessment is more imperative. This is clearly true when trying to assess public security risks.

Probability enters the risk “formula” in a number of ways. There is the probability that describes the likelihood of occurrence of a risk event, and there are the probabilities that describe the likelihood of different possible outcomes, should that event occur. “Objective” probabilities are derived on the basis of historical records, statistical analysis, observations, and experimentation. However, there are situations when the “objective” data is sparse and experimentation is impractical, and one must make use of “subjective” probabilities or probabilities that are based on expert judgement. In this paper, we will review ways to generate probabilities based on expert evidence.

1.1 Background

The Centre for Security Science (CSS) is the newest among the Defence Research and Development Canada (DRDC) research centres. CSS has been set up as a joint endeavour with the Department of Public Safety Canada, extending S&T services, which DRDC traditionally provides in support of the Canadian Forces, to address national public safety and security objectives. The Centre's capabilities lie in leading and administering research, development, testing and evaluation of technologies, identifying future trends and threats, as well as a network of national and international S&T partners within the public safety and security communities. Within CSS, a Risk Portfolio has been established in response to growing interest in the risk field from across government and defence. The vision is to develop a risk resource centre to support the community with threat, vulnerability and risk assessments, gap analysis, foresight and future security visioning and other related activities and products. One of the more significant projects supported within the CSS Risk Portfolio is an All-Hazards Risk Assessment, a significant undertaking for the Centre but one that has the potential of benefiting the entire safety and security community, and the various levels of government.

In 2006, the Intelligence Advisory Coordination Committee (IACC), an interdepartmental coordinating body for the Canadian intelligence assessment community chaired by the head of the International Assessment Staff (IAS) Secretariat of the Privy Council Office (PCO), has set up an

Intelligence Experts Group (IEG) on Domestic Security, to facilitate the sharing of intelligence on domestic threats and produce assessments. The group included participants from 15 federal departments and agencies under the interim leadership of the Department of National Defence (DND) and the Royal Canadian Mounted Police (RCMP). One assessment requested by the IACC was an All-Hazards Risk Assessment (AHRA) that could potentially inform senior decision-makers in the federal government in the areas of funding, learning and development, resource allocation and threat assessment.

Given its mandate, its interest in strengthening collaboration and investing in capabilities in the public security domain, as well as the centre's interest in enhancing its own risk assessment capabilities, CSS saw an opportunity in supporting the AHRA initiative and volunteered to take it on as a project, to explore methodology and coordinate the effort.

In Canada, all levels of Government share the responsibility to protect Canadians and the Canadian society. Within each jurisdiction, the governments' public safety and security functions are shared among many departments and agencies. Hence, being prepared at the national level depends on synchronized efforts among federal, provincial, territorial and municipal partners. In order to ensure that adequate programs that enhance the safety and security of the whole population are put in place, there is a need to have a sound understanding of what kind of threats the society might face. Even singular events may evolve into emergencies that escalate across jurisdictions and organizational boundaries. When it comes to multiple events, the lack of a coherent picture on the relative severity of risks associated with threats/hazards of different types makes the matter of adequate assessment and management the more complex. Such an approach is, nonetheless, important.

When major decisions are to be made in the presence of substantial uncertainty, such as when planning for extreme events (dam collapse, earthquakes or terrorist attacks, to give just a few quite distinct examples), expert judgement is deemed essential to minimize and characterize the uncertainty. For these types of events, there rarely exist ways to calculate "objective" probabilities, and experimentation is quite obviously out of the question. But even for domains where data is available and there are grounds for objective analysis, aspects such as the interpretation of results or questions about acceptable risk levels underline normative, value-judgement perspectives that are unavoidable in risk-based decision making. To paraphrase the title of an article published in the *Risk Analysis* Journal, there is no escape from expert judgement in risk-based decision-making [1]. For this reason, it is important, both to support the work on methodology for the AHRA project, but also as a contribution to the risk "toolbox" housed within the CSS Risk Portfolio, to put together a review of the various roles that expert judgement plays in risk assessment/management as well as to outline the important stages of a sound elicitation process.

1.2 Aim

The aim of this paper is to present a synthesis of models and techniques for eliciting and aggregating expert judgements, together with methods to evaluate the accuracy of elicitation and the quality of the risk assessment, in order to assist CSS and other public security partners with essential tools to conduct sound risk assessments.

1.3 Outline

Following a general introduction on risk-based decision making in the public security domain and a quick background on the CSS Risk Portfolio and the AHRA collaborative project, Section 2 presents fundamental concepts on the elicitation of expert judgement in risk assessment; we will review roles in the elicitation process, criteria for expert selection and essential steps for a successful elicitation. Section 3 talks about limitations on the accuracy of experts' predictions due to heuristics and biases and offers both a psychological and a statistical perspective on them. In Section 4, methods to evaluate the quality of the elicitation process are reviewed; first we discuss how calibration measures the agreement between the expert's judgement and reality, then we present scoring rules that gauge how well the elicited probabilities express the expert's underlying opinion. To improve the accuracy of risk estimation, Section 5 looks at multiple experts and reviews aggregation methods as a way to reduce random variation in a set of judgements, including a discussion on the relative merits and shortcoming of the various approaches. Section 6 presents briefly a conceptual framework aimed at using expert judgement for the quality verification of risk assessment. We end with a summary of the main points and provide recommendations on best ways to engage experts in the risk assessment process.

2 Fundamental Concepts in Expert Judgement Elicitation

2.1 Elicitation and Expert Judgements

Elicitation is the process of extracting expert knowledge about one or more uncertain quantities and formulating this information as a probability distribution. In Bayesian statistics, elicitation is the basis for formulating the prior distribution; the posterior distribution is then calculated via Bayes' Theorem.

Expert judgements are the expression of inferential opinions, based on knowledge and experience. An expert is a person who is recognized and qualified with special knowledge or skills and with relevant experience in a particular domain. Criteria for defining expertise include peer assessment, demonstrable ability, and length of experience. Cognitive scientists [2] estimate the time required to acquire expertise within a professional domain to be on the order of ten years.

Expert judgement provides valuable information when the data are sparse, nonexistent, limited, difficult, or costly to obtain, and open to differing interpretations. It is particularly useful to provide estimates on new, rare, complex, or poorly understood phenomena. These estimates can include failure rates, incidence rates, or weighting factors for combining data sources. Expert judgement is also used to integrate heterogeneous information, determine the state of knowledge in a problem, and document the information in a data or knowledge base.

To illustrate the range of application, a number of examples are included where formal elicitation has been used, and two case studies – in the nuclear industry and a business study – are discussed in the Appendix at the end of the report.

2.2 Roles in the Elicitation Process

Rosqvist [3] suggests five distinct roles involved in the elicitation process. These include the decision-maker, the facilitator, the normative expert, the domain experts, and the stakeholders:

1. **Decision-maker** – presents views, status of the decision-making process, and the objective of the outcome; identifies and selects stakeholders; defines resources; and provides decision criteria;

2. **Facilitator** – selects and interviews experts, describes the case, gives comments, accepts reports, and explains conclusions to the decision-maker;

3. **Normative Expert** – trains the domain experts, elicits judgements, reports, draws conclusions based on decision criteria, and combines judgements in the case of quantitative judgements;

4. **Domain Experts** – analyze the issues that they are familiar with, and provide requested qualitative or quantitative judgements; and

5. **Stakeholders** – affected by the decision; give feedback; affirm the scope and completeness of issues.

Elicitation is best conducted as a face-to-face interaction between the experts and the facilitator [2]. It requires a careful, structured dialogue between them since experts are not expected to have expertise in probability. Having a well-planned and structured interaction will minimize the experts' biases.

2.3 Model for the Elicitation Process

Existing literature identifies certain critical elements that lead to a good elicitation process. O'Hagan et al. [2] suggest a model process that goes through five essential stages, underlined below. The actual elicitation comes at the end, while its success depends on the foundations built in the preceding stages.

1. **Background and Preparation**

   The first step is to identify variables for which expert assessment is needed. At this stage, the normative expert and the facilitator gain enough substantive expertise to be able to converse with the experts about their field to the required level of competency. This stage also includes planning the elicitation session and preparing any required documents such as background information for the experts.

2. **Identify and Recruit Experts**

   Hora and von Winterfeldt [4] recommend the use of a nomination system that includes public interest groups as well as professional organizations in order to gain a balanced perspective. They suggest six criteria when selecting the experts:

   (i) Tangible evidence of expertise;

   (ii) Reputation;

   (iii) Availability and willingness to participate;

   (iv) Understanding of the general problem area;

   (v) Impartiality; and

   (vi) Lack of an economic or personal stake in the potential findings.

3. **Motivate and Train the Experts**

   It is advisable to explain to the experts why their judgements are required and how their opinions will be used. For effective elicitation, the experts should have some understanding of the probability concept. The training of experts consists of three parts:

   (i) Training on probability and probability distributions;

(ii) Information about the most common judgement heuristics and biases (see Section 2), including advice about how to overcome them; and

(iii) Practice elicitations, particularly using examples where the true answer is known, but unlikely to be known by any of the experts. An example is the distance between two towns [5].

4. Structuring and Decomposition

The quantities that are to be elicited must be defined precisely, including a specification of their units of measurement. It is also important to review with the experts the available evidence upon which they will draw while making judgements about these quantities.

5. The Elicitation

Garthwaite et al. [6] identify an iterative process for elicitation:

(i) Elicit specific summaries of the expert's distribution for the desired variable;

(ii) Fit a probability distribution to those summaries; and

(iii) Assess adequacy: if adequate stop, if inadequate repeat the process, asking the expert to make adjustments.

2.4 Expert Judgement in Risk Assessment

Subjective judgements by domain experts can be both *qualitative* and *quantitative*. *Quantitative* forms can be expressed in numerical value of probabilities, ratings, odds, uncertainty estimates, weighting factors, physical quantities of interest (e.g. costs, time, length, weight, etc). *Qualitative* forms can be textual descriptions of the expert's assumptions in reaching an estimate, reasons for selecting or eliminating certain data or information from the analysis, and natural language statements of physical quantities of interest.

Expert judgements are required in most steps of risk assessment: hazard/threat identification, risk estimation, risk evaluation, and analysis of options. Specific judgements required during each phase of risk assessment are synthesized in Table 1, adapted from Rosqvist [3].

Table 1. Expert judgements in each step of risk assessment phase

Risk Assessment Phase	Decision-maker	Facilitator	Normative Expert	Domain Expert	Stakeholder
Scope Definition	Provides scope in general terms, decision criteria	---	---	---	---
Hazard Identification	---	Explains the motive, surveys stakeholder feedback	Method of brainstorming and expert panel	Qualitative judgements and opinion on priorities	Qualitative judgements and opinion on priorities, affirms completeness
Risk Estimation	---	---	Expert judgement protocol	Quantitative judgements on quantities	---
Risk Evaluation	---	---	Conclusions based on decision criteria	---	---
Analysis of Options	---	Provides conditions for system, options, surveys stakeholder feedback	Method of brainstorming and expert panel	Qualitative judgements and opinion on priorities	Qualitative judgements and opinion on priorities, affirms completeness

There is a large and widely dispersed literature on expert judgement elicitation in both psychology and statistics. Research findings from psychology suggest that there are limitations to how the human mind assesses probabilities or makes predictions about the future value of some quantity. People do not judge uncertainties well. To underline this point, some psychological research and theory will be discussed in the next section. From a statistician's perspective, this uncertainty is an integral part of the analysis of expert judgement, and mathematically rigorous methods have been developed for aggregating differing experts' responses, quantifying the accuracy of experts' predictions, and formulating models using experts' opinions. These methods will be discussed in detail in the last sections.

3 Psychological Aspects – Heuristics and Biases

Research shows that limitations of human memory and information processing capacity often lead to subjective probabilities that are poorly calibrated or internally inconsistent, even when assessed by experts [7]. People cannot be guaranteed to act as rational agents who follow the prescriptions of probability and decision theory. In many situations where experts are asked to provide judgements under uncertainty, they are liable to choose from a selection of easy-to-use strategies which are referred to as heuristics. These strategies can be effective, especially when time or information is scarce, but they do not always lead to accurate assessments.

Heuristics and biases may lead to inconsistency between the experts' system knowledge and their assessment of uncertainty. They may also introduce a disparity between the perceived uncertainty and the probability figure which is eventually obtained from experts. Experts showing consistent biases, across a range of knowledge areas, indicates poor elicitation.

Psychologists have described the heuristics that they believe to be responsible for these biases in terms of tendencies to process information quickly. A formal elicitation process should seek to minimize these biases in the expert judgement it obtains. Daneshkhah [8] summarizes the cognitive factors which affect subjective probability judgement and expert opinion during the elicitation process. These prominent heuristics include *availability*, *representativeness*, and *anchor-and-adjustment*.

3.1 The Availability Heuristic

The availability heuristic is an intuitive strategy used to judge the probability of an event or the frequency of class membership according to how easily events or instances of a class come to mind. Judged probabilities depend on the ease with which experts can recall similar events or instances. Events that have occurred recently or have particular personal significance will also come to mind more easily. For example, one might judge the probability of having a car accident higher by recalling an instance when a friend has recently had a car accident. In these types of cases, relative frequencies of well-publicized events (e.g. tornadoes) tend to be overestimated whereas relative frequencies of more mundane events (e.g. common influenza) are often underestimated [9]. In other words, people overestimate the probabilities of events similar to ones that they have experienced or read about, and underestimate the probabilities of less familiar events.

A common bias associated with this heuristic is *illusory correlation*, erroneous belief that two uncorrelated events have a statistical association. People misreport the frequency of co-occurrences of pairs of events that they have observed. This impacts directly on subjective probability judgement, as experts are often asked to estimate joint or conditional probabilities that depend on the correlations between events. Prior beliefs about reasons why two variables should correlate may blind experts to the data that indicate variables are in fact uncorrelated [10].

3.2 The Representativeness Heuristic

The representativeness heuristic is an intuitive strategy used when assessing the likelihood of an event's occurrence by the similarity of that occurrence to stereotypes of similar occurrences [8]. In other words, probability is assessed by how "representative" outcome A is of model B. People tend to intuitively evaluate the conditional probability P(A|B) by assessing the similarity between A and B. The more A is similar to B, the more likely people think A belongs to B. The problem with this assessment is that similarity is symmetric whereas conditional probabilities are not. There are factors affecting probability judgement that do not influence similarity or representativeness. For example, specific outcomes can be representative although unlikely, such as when diagnostic outcomes are representative of a disease, yet are improbable.

Several biases that might be attributed to the representativeness heuristic are listed below; they include *conjunction fallacy*, base rate neglect, insensitivity to sample size, confusion of the inverse and insufficiently regressive predictions.

(i) Conjunction fallacy

This bias refers to the incoherence in probability assessment that occurs when the conjunction of to events is judged more probable than either event separately.

Consider the following example [11]:

Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice and also participated in anti-nuclear demonstrations. Please check off the most likely alternative:

(1) Linda is a bank teller; or

(2) Linda is a bank teller and is active in the feminist movement.

Gigerenzer, G. and Hoffrage [12] show that most people find the second option more likely than the first one, simply because the description is representative of a feminist type. However, it cannot be more likely that Linda is both a bank teller and a feminist than that she is a bank teller, as the conjunction of two events can never be more probable than either event separately.

(ii) Base rate neglect

This bias corresponds to people's tendency to assess subjective probability by the similarity of one event to another. They ignore or under-weigh the prior probability of an event when they focus on recent identifiable results. For example, an event known to be rare in the physical world will be assigned a reasonably high probability based on imperfect evidence.

(iii) Insensitivity to sample size

The size of a sample greatly affects the likelihood of obtaining certain results. This type of bias is introduced when people ignore sample size and only use superficial similarity measures, ignoring the fact that larger samples are less likely than smaller samples to deviate from the mean.

(iv) Confusion of the inverse

People misunderstand conditional probabilities like P(A|B) with P(B|A). For example, this bias is introduced when doctors confuse test sensitivity P(positive test | disease) with P(disease | positive test).

(v) Insufficiently regressive predictions

People make predictions as if they translate one scale into another scale, placing each new value at a point of equivalent extremity to its position on the original scale. For example, people may think being tall as representative of being heavy.

These biases can create an unbounded probability problem, as people tend to overestimate each probability in a set of exhaustive and mutually exclusive schemes, so that the estimated sum of all probabilities becomes greater than one [13].

3.3 The Anchor-and-adjustment Heuristic

The anchor-and-adjustment heuristic is an intuitive strategy used when quantitative judgements are made by making simple adjustments from an initial starting value. When experts are asked to estimate an uncertain quantity, they anchor on what seems to them to be the most likely value of the quantity, and underestimate the variability or uncertainty of their estimate. For example, when people are asked to give an interval that contains 98 percent of the observed data, they tend to anchor their assessment on the median as an "initial estimate", and adjust up and down insufficiently. This usually leads to ranges that cover approximately 70 percent of the data [8].

This heuristic has important implications for the nature and ordering of questions when eliciting any numerical quantity. For example, subtle anchoring effects occur when considering the pruning bias in "fault trees" [14]. Here is an example of experiment: mechanics are asked to consider reasons a car might fail to start, and to assign probabilities to the various reasons. One group of participants considers seven categories (branches of the fault tree) while the other group considers a "pruned fault tree" where the final four categories in the seven-branch tree are subsumed into a single "all other problems" category. Responses in these two groups were mismatched. The "all other problems" category in the pruned tree condition results in lower probability estimates than the sum of its logical equivalents in the un-pruned condition.

Fox and Clemen [15] show that probabilities are "*partition dependent*", which means that assessed probabilities are biased toward a uniform distribution (equal allocation) over all events into which the relevant state space is partitioned. They surmise that experts anchor their judgments with an *ignorance prior* distribution that assigns equal probabilities for each event in the specified partition, and then adjust those probabilities insufficiently to reflect their beliefs about how the likelihoods of the events differ. In this way, judgements are systematically affected by different anchors.

Fox and Clemen [15] find that bias decreases with increased domain knowledge. Participants with greater substantive expertise show less partition dependence, especially if they consider multiple partitions of the state space. However, experts may lack sufficient knowledge to completely overcome the bias. Partition dependence seems to be quite robust to varying levels of procedural expertise. Top experts in decision analysis may be susceptible to this bias.

For a more detailed review of the heuristics and biases that influence the accuracy of experts' judgment of probabilities, see O'Hagan et al. [16].

3.4 Discussion

The psychology of judgement under uncertainty plays a major role in behavioural decision-making. O'Hagan et al. [16] suggest that the facilitator and the experts should be aware of the biases that can occur in judgement under uncertainty and the reasons for their occurrence. The potential for inadvertently introducing biases into the elicitation process should also be recognized, such as anchoring effects induced by particular orders of questions. It is helpful to introduce procedures that explicitly encourage experts to think analytically. For instance, the facilitator should consider some training and aids for the experts to reduce or avoid biases. In addition, it is necessary to assess the coherence of the set of assessments provided by the experts themselves, especially when eliciting assessments for a series of mutually exclusive events.

4 Evaluating Elicitation

The aim of the elicitation is to formulate the experts' knowledge faithfully in probabilistic form, so any evaluation of the elicitation should measure the extent to which this aim has been met. A good elicitation should accurately reflect both the assessor's opinion and the reality.

4.1 Calibration Curve

Calibration is a measure of the quality of subjective judgements elicited from experts. It is defined as the faithfulness of generated probabilities, in that events that are assigned a probability of *p* should occur with a relative frequency of *p* [17]. We can draw a calibration curve graphically in order to evaluate the accuracy of the elicitation.

Calibration curves provide a graphical representation to examine accuracy over a range of confidence levels. On a calibration curve graph, the horizontal axis represents the predicted value and the vertical axis represents the true value. *Subjective probability judgement* (X) is plotted against *observed relative frequency* (Y). A near diagonal curve indicates a good calibration.

Overconfidence occurs when the observed relative frequency is lower than the mean subjective probability, or when the elicited probabilities are extreme near either end (0 or 1) and the distributions have little dispersion about the mean. Past experiments [18] show that overconfidence is the most widespread bias in assessing probability. On the other hand, *underconfidenc*e occurs when observed relative frequency is higher than the mean subjective probability. In general, underconfidence occurs less frequently then overconfidence [19].

There are two forms of overconfidence: *over-prediction* and *over-extremity*. *Over-prediction* is the tendency to assign probabilities that are consistently too high, while *over-extremity* is the tendency to assign probabilities that are too extreme (i.e. too close to either 0 or 1). Under-estimation and under-extremity can be defined similarly for underconfidence, but are far less common. Over-extremity is the most frequent type of overconfidence; in this case, the calibration line will be more shallow then the diagonal and will cross it. For overconfident curves one can look at the degree of *discrimination,* which measures the ability to identify correctly when an event is more, or less, likely to occur. The better the discrimination exhibited by a set of probability judgements, the closer to the diagonal the calibration curve will be. In other words, a horizontal calibration curve indicates poor discrimination [16].

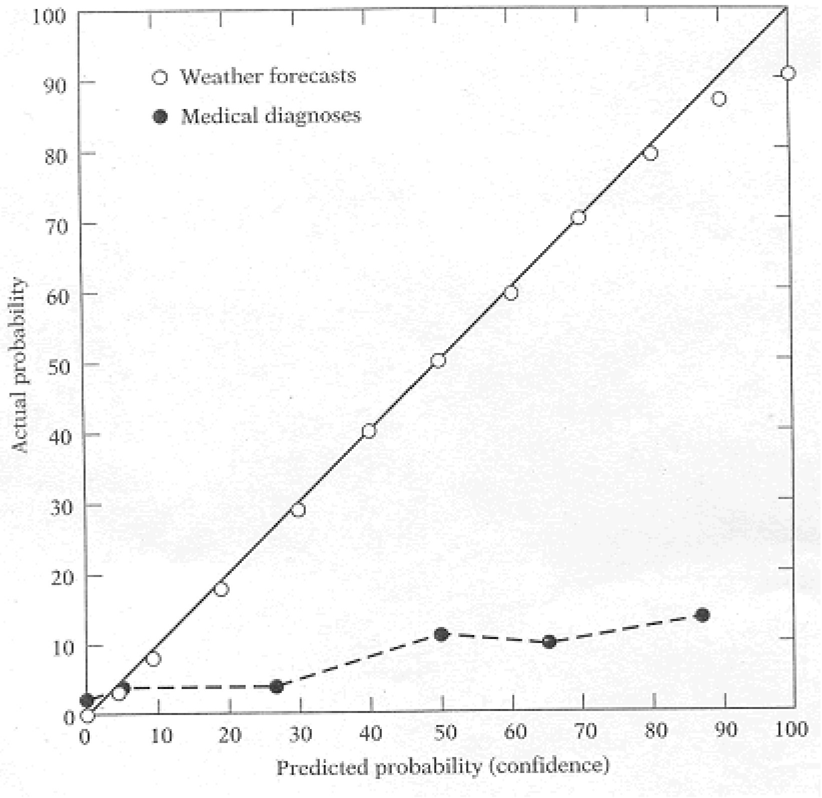

Figure 1. Calibration curves for weather forecasts and medical diagnoses (taken from Plous [20])

Figure 1 contains two calibration curves for weather forecasters' predictions of precipitation (hollow circles) and physicians' diagnoses of pneumonia (filled circles). The predictions of weather forecasters are almost perfectly calibrated; on average, their predictions closely match the weather. In contrast, the physicians are poorly calibrated; most of their predictions lie below the diagonal, indicating substantial overconfidence (i.e., unwarranted certainty that patients have pneumonia). The physicians’ set of probability assessments also exhibits poor discrimination, resulting in an almost horizontal calibration curve. The data on weather forecasters comes from a meteorology report by Murphy and Winkler [21], and the data on physicians comes from a study by Christensen-Szalanski and Bushyhead [22].

4.2 Improving Calibration of Subjective Probabilities

Subjective probabilities can be well calibrated, but in many cases they are not. This highlights the importance of training. In previous study [17], calibration improved after training the experts in probability elicitation via testing, practice, and feedback.

Simple averages of elicited probabilities work well in terms balancing the elicitation effort with the accuracy of predictions for decision and risk analysis. Hora [17] demonstrates that an equally weighted linear combination of probabilities elicited from experts who are “overconfident” can improve calibration. However, it is shown by theoretical arguments that combining well-calibrated predictions of individual experts via a linear rule results in reducing calibration. On the other hand, improvement in calibration for overconfident groups of experts is rapid as the size of the group increases from a single expert to five or six, but there is only modest improvement from increasing the number of experts beyond that point.

4.3 Poor Calibration

There are two different causes for failure of an expert to provide well-calibrated assessments. *Poor elicitation* results when experts’ beliefs are actually well calibrated to reality, but the elicited probabilities fail to express those beliefs accurately. On the other hand, experts may express their beliefs accurately, but those beliefs agree poorly with reality. In this case, poor calibration may reflect *inaccurate knowledge* rather than poor elicitation. If an expert’s probabilities are not well calibrated, then poor elicitation, inaccurate knowledge and poor appreciation of the expert’s knowledge may all be present. It is not possible to say which is more responsible for mis-calibration.

4.4 Scoring Rules

Although calibration measures the agreement between the expert’s stated probabilities and reality, it does not compare the elicited probabilities with the expert’s underlying opinions. We cannot only use calibration to evaluate the success of elicitation [16].

O’Hagan et al. [16] develop the use of proper scoring rules to compare the expert’s performance against reality. A scoring rule is termed proper if it encourages experts to record their true beliefs, because any departure from the expert’s personal probability results in a diminution of his/her own average as he/she sees it [23]. Poor scores inevitably confound poor elicitation with inaccurate knowledge. These scores also provide an incentive for experts to record their opinions well and to help train experts to quantify their opinions accurately.

Probability score measures the discrepancy between one expert’s assessment and reality, so it should be viewed as a penalty rather than a reward to the expert. High scores indicate a high discrepancy while low scores indicate a low discrepancy. Hence a score of 0 indicates faultless prediction of the outcome of every event.

4.4.1 Scoring Rules for Discrete Probability Distributions

First, a single event that has n possible outcomes is considered.

Suppose the event E takes exactly one of the n outcomes, $O_1, \ldots\ldots, O_n$

Let p_i be the probability that the expert states he/she attaches to O_i ($i = 1, \ldots., n$); p_i satisfies the usual laws of probability, so each p_i must be non-negative and $\sum_{i=1}^{n} p_i = 1$.

Let d_i = 1 if O_i occurs and 0 if O_i does not occur (exactly one of $d_1,\ldots\ldots,d_n$ is non-zero).

The probability score is defined bellow:

Probability score: $$PS = \sum_{i=1}^{n} \left(p_i - d_i \right)^2$$

If a set of events $E_1,\ldots\ldots,E_m$, as opposed to a single event E, is considered, useful comparisons between sets of probability judgements can be made by calculating the *mean probability score* $\overline{PS}$, simply the arithmetic mean of the expert's probability scores for each of a set of events with the same set of possible outcomes. Comparison is often made with the performance of a "uniform judge" who always states the same probability over a set of probability judgement. For example, an expert who assesses each outcome of a set of dichotomous events to be 0.5 would obtain a $\overline{PS}$ of 0.25.

Various schemes have been proposed that decompose measures of overall accuracy (i.e., decompose $\overline{PS}$) into meaningful sub-components. One purpose of such decomposition is to try to explain how and why experts achieve the levels of accuracy that their probability scores indicate. Also, the score can be given as feedback to help train the assessors, but information given by probability scores may be too coarse to inform the expert precisely on how to improve his/her judgement. The most widely used decomposition is Murphy's decomposition [24].

Suppose each event in the set $E_1,\ldots\ldots,E_m$ has the same set of possible outcomes $O_1,\ldots\ldots,O_n$.

The probability assessments for this outcome are partitioned into T categories.

In what follows the notation j will be used as an event index (j = 1,….,m), the notation i will be used as an outcome index (i = 1,….,n), and the notation k will be used as a category index (k = 1,….,T).

Murphy's decomposition partitions $\overline{PS}$ into three components:

Mean Probability score: $$\overline{PS} = \sum_{i=1}^{n} \hat{\text{var}}\left(d_i\right) + \sum_{i=1}^{n} \text{calibration}\left(i\right) - \sum_{i=1}^{n} \text{resolution}\left(i\right),$$

where $\hat{\text{var}}\left(d_i\right)$, $\text{calibration}\left(i\right)$ and $\text{resolution}\left(i\right)$ are defined for the i^{th} outcome as follows:

Let $\overline{d_i}$ be the proportion of the m events for which outcome i occurred (e.g. the proportion of patients with illness i) : $\overline{d_i} = 1/m \sum_{j=1}^{m} \left\{d_i\right\}^j$, with $\left\{d_i\right\}^j$ being a zero-one indicator that denotes whether outcome i has occurred for event j. Then one can set

$$\hat{\mathrm{var}}\left(d_i\right) = \overline{d_i}\left(1 - \overline{d_i}\right),\ i = 1,\ldots\ldots,n$$

(If $\left\{d_i\right\}^j$ is a zero-one indicator, then $\left[\left\{d_i\right\}^j\right]^2 = \left\{d_i\right\}^j$ and one can show that $\hat{\mathrm{var}}\left(d_i\right)$ as defined above is the maximum likelihood estimate of the variance of d_i; see [16]).

The probability assessments for this outcome are partitioned into categories according to their values. (e.g. 0 to 0.1, 0.1 to 0.2, …., 0.9 to 1)

Let T be the number of categories
Let $\overline{p}_i^k$ be the average of the probability assessments in the k^{th} category
Let $n^k{}_i$ be the number of times the assessments for this outcome fell in the k^{th} category
Let $\overline{d}_i^k$ be the proportion of times outcome i occurred when the probability assessment for this outcome was in the k^{th} category

$$\mathrm{calibration}\left(i\right) \ = \ 1/n \ \sum_{k=1}^{T} n^k{}_i \left(\ \overline{p}_i^k \ - \ \overline{d}_i^k \ \right)^2$$

$$\mathrm{resolution}\left(i\right) \ = \ 1/n \sum_{k=1}^{T} n^k{}_i \left(\overline{d}_i^k \ - \ \overline{d}_i \ \right)^2$$

The first component, $\hat{\mathrm{var}}\left(d_i\right)$, discussed in a previous paragraph, is determined purely by the outcomes and not by the probability assessments. The second component, $\mathrm{calibration}\left(i\right)$, indicates how well the probability assessments related to reality for outcome *i*. The third component, $\mathrm{resolution}\left(i\right)$, indicates whether the categories discriminate between the occurrence/non-occurrence of outcome *i*, and it differs from the first two terms in that a larger value improves $\overline{PS}$.

The logarithmic and spherical scoring rules are other proper scoring rules that have been proposed:

Logarithmic score:	$LS = \ln\left(\sum_{i=1}^{n} p_i d_i\right)$
Spherical score:	$SS = \left(\sum_{i=1}^{n} p_i d_i\right) \Big/ \left(\sum_{i=1}^{n} p_i{}^2\right)^{1/2}$

On the basis of the frequency of use, the *probability scoring rule* (also called quadratic score) is the preferred scoring rule for judging subjectively assessed discrete distributions (Example: meteorology – weather forecasting). In principle, various decompositions of the probability scoring rule can be used to focus on different features of a set of assessments, but, in practice, their efficacy is unclear.

4.4.2 Scoring Rules for Continuous Probability Distributions

Let θ denote the quantity of interest (the value of θ must be known to the facilitator or become known).

Let *f(θ)* be the expert's subjective probability density function for the value that θ takes, and *F(θ)* be the expert's cumulative distribution function.

Let θ* denote its actual value.

Scoring rules for continuous distributions are defined bellow:

Linear score = $f\left(\theta *\right)$

Quadratic score = $2f(\theta *) - \int_{-\infty}^{\infty} \left\{ f(\theta) \right\}^2 \, \mathrm{d}\theta$

Logarithmic score = $\ln\left[f(\theta *) \right]$

Spherical score = $f\left(\theta *\right) / \left(\int_{-\infty}^{\infty} \left\{ f(\theta) \right\}^2 \, \mathrm{d}\theta \right)^{1/2}$

Ranked Probability Score = $\int_{-\infty}^{\theta *} \left\{ F(\theta) \right\}^2 \, \mathrm{d}\theta + \int_{\theta *}^{\infty} \left\{ 1 - F(\theta) \right\}^2 \, \mathrm{d}\theta$

On the basis of the frequency of use, the *logarithmic scoring rule* is the preferred scoring rule for continuous distributions (Examples: estimating the time a specific operation will take, the average change in blood pressure that will be observed in a particular drug trial, or the future change in the weight of a person who has recently developed diabetes). Existing literature suggests that scoring rules are used less frequently with continuous distributions.

4.5 Coherence, Feedback, Over-fitting

A set of probability statements is *coherent* if they are collectively consistent with the laws of probability. One of the most common incoherent cases occurs when the sum of the probabilities associated with all possible independent outcomes of an event is either more or less than one. There has been only limited empirical success at relating coherence to better calibration [16].

The most natural way of improving the accuracy with which a subjective distribution represents an expert's opinion is through *feedback.* It is very important that feedback is given on training exercise to help the experts learn how to make better probability assessments. One approach is to elicit the probability of an event and then give the experts feedback by telling them whether the event occurred and perhaps give scores for their assessments. The following is a common approach for assessing a probability distribution [16]:

1. Elicit sufficient assessments to determine the probability distribution;

2. Use this distribution to estimate other quantities that should correspond to the expert's judgement if the assessed probability distribution adequately represents his/her opinion;

3. *Feedback*: Inform the expert about these estimated quantities (perhaps using interactive graphics) and ask if they represent his/her opinion; and

4. If they do not represent his/her opinion, ask him/her either to give estimates of the quantities that do or to revise some of his/her earlier assessments.

Steps 2 to 4 are repeated until a probability distribution is determined that seems to represent the experts' opinions adequately.

In addition, feedback encourages experts to think more carefully about their assessments and, among other benefits, it can highlight apparent errors, thus providing an opportunity for the expert to correct them, which seems the most sensible way of dealing with assessments that are out of line. Note that interactive software is almost essential for the effective use of feedback (graphical feedback).

One problem with feedback is that experts may be too willing to accept that a proposed value is representative of their opinions, since this is the first option, and they avoid having to revise the value. Moreover, if they revise it, then the proposed value may act as an anchor and lead to their adjustment being insufficient (see Section 2).

To avoid these issues, *over-fitting* should be coupled with feedback. The experts are asked to make more assessments than are necessary for their subjective probability distribution to be estimated. Then a distribution is fitted to these assessments through some form of reconciliation. The differences (residuals) between the assessments and those given by the fitted distribution would be calculated and fed back to the expert, with large residuals highlighted. The experts are then asked to modify assessments and a probability distribution is refitted. In this case, large assessment errors are noticed and corrected. This process is repeated until a distribution is found that is close to all their revised assessments.

5 Expert Judgement Protocol in Risk Estimation – Multiple Experts

Judgements under uncertainty are demonstrably subject to random variation. Often, there exist approaches that are effective in reducing this random variation in a set of judgements, thereby increasing the accuracy of the set of judgements. Aggregating judgements from multiple experts is one of the approaches.

Winkler et al. [25] explain that the reason why assessments of multiple experts should be aggregated is that a combined distribution can produce a better appraisal than the individual distribution. It is in accord with the psychological perspective that "several heads are better than one head" and with the statistical fact that "a sample mean is better than one observation". This combined distribution can be considered as consensus and it is more practical and feasible to use a single distribution as a representative of several distributions for further analysis.

The expert judgement protocol in *risk estimation* is to aggregate elicited probability distributions of an unknown quantity and to control cognitive biases inherent in eliciting probabilities. Aggregation may be achieved in two ways: by *behavioural aggregation;* and by *mathematical aggregation* [26].

Behavioural aggregation is to create an interaction between the groups of experts, through which a single distribution is elicited from the group as a whole. On the other hand, *mathematical aggregation* is to elicit one distribution from each expert individually and independently of the others and then mathematically combine the resulting distributions into a single distribution.

5.1 Behavioural Aggregation

Behavioural aggregation approaches require experts themselves to produce the consensus distribution. The purpose is to generate agreement among the experts by various ways of interaction in order to share and exchange knowledge, information, and interpretations between the experts. Methods include face-to-face group meetings via discussion and debate, interaction by computer, or sharing information in some other way.

Having experienced analysts serve as facilitators for the group interaction can improve the process. For example, *decision conferencing* [27] is a particular approach to structuring group discussion for decision-making and group probability assessment. The role of facilitators is to control the process and structure of group interaction.

There are some more specific and controlled forms of interaction between the experts. These include the Delphi method, Nominal Group Techniques, and Kaplan's approach.

1. Delphi Method

The Delphi method [28] is one of the oldest approaches to structured group judgements. Experts first make individual judgements after which judgements are shared anonymously. They may then revise their probabilities, and the process is repeated for a few rounds. The final probability distributions are aggregated mathematically. (see Section 5.2)

2. Nominal Group Technique

The Nominal Group Technique [29] is a variant of Delphi in which experts first assess judgements individually and then present to other group members. Group discussion follows with the assistance of a facilitator, and then experts may revise their probabilities. The final probability distributions may require mathematical aggregation as well.

3. Kaplan's Approach

Kaplan [30] emphasizes that group elicitation should focus on combining the experts' evidence rather than their opinion. The experts are first invited to present and discuss their evidence in a group meeting. The facilitator then leads the experts through discussions to determine the consensus body of evidence. Subsequently, the facilitator proposes a probability distribution, conditioned on the consensus. He or she must obtain assurance from the experts that evidence has been interpreted correctly in arriving at the probability distribution.

Reagan-Cirincione [31] finds that small group elicitations can outperform their best individual member if the elicitation combines these three features: impartial facilitation that is responsive to the potential for biases in the group interaction; a well-designed protocol that involves careful structuring and decomposition of the elicitation task; and continuous feedback via computer technology, outlining the implications of the experts' judgements.

To choose the right approach in behavioural aggregation, one should consider the type of interaction (face-to-face, via computer, anonymous); the nature of interaction (e.g. sharing information); the possibility of individual reassessment after interaction; and the role of the assessment team (e.g. facilitators). If these behavioural methods do not succeed in obtaining a single consensus distribution for all experts, then a mathematical aggregation may be applied.

5.2 Mathematical Aggregation

In mathematical aggregation, experts' individual probability distributions are aggregated by the assessment team after elicitation. The purpose is to specify the performance of the experts and combine or adjust the individual probability values or distributions into one single value or distribution. The most common methods include Bayesian methods, opinion pooling, and Cooke's method.

5.2.1 Bayesian Methods

Bayesian methods were first proposed by Morris [32]; full review of his work is provided by Clemen and Winkler [26]. Suppose each of n experts is asked individually to assess some unknown quantity θ, so that a distribution $f_i(\theta)$ is elicited from expert i. The decision-maker first begins with his own prior distribution $f(\theta)$ for θ and defines the likelihood of the experts' judgements as the posterior distribution *f(θ|D)*, where $D = \{f_1(\theta), \ldots.., f_n(\theta)\}$ is the set of the experts' elicited distributions.

For updating prior belief to posterior belief according to Bayes' Theorem, *f(θ|D)* is proportional to *f(θ)* multiplied by the likelihood term *f(D|θ)*. Data is in the form of a finite number of percentiles or full probability distributions.

In practice, this approach is complex and difficult to implement. The decision-maker needs to conduct a very sophisticated elicitation exercise on his or her prior distribution *f(θ)* and construct *f(D|θ)*, which formulates the prior beliefs about what the experts are going to tell him or her, conditional on the true value of *θ*. One example of the application of this approach is in weather forecasting.

5.2.2 Opinion Pooling

Opinion pooling is simple and widely used in practice, especially linear opinion pool. A *consensus distribution f(θ)* is obtained as some function of the individual distributions $\{f_1(\theta), \ldots.., f_n(\theta)\}$.

1. *Linear opinion pool*: weighted average of the individual distributions with weights w_i summing to 1

$$f(\theta) = \sum_{i=1}^{n} w_i \, f_i(\theta)$$

   Simple average (equal-weighted) is $w_i = 1/n$ (for *n* experts); otherwise choose weights depending on the expertise of the experts.

2. *Logarithmic opinion pool*: weighted geometric mean of the *n* individual distributions

$$f(\theta) = k \prod_{i=1}^{n} f_i(\theta)^{w_i}$$

   where *k* is a normalizing constant that ensures that *f(θ)* integrates to 1

Assuming there are two experts, the logarithmic pool implies stronger information than that given by either expert separately, whereas the linear pool represents less knowledge than either expert alone. Logarithmic opinion pool is more appropriate if both experts were held to be good judges of *θ* on the basis of the information available to them. On the other hand, linear opinion pool is more appropriate when the decision-maker's beliefs encompass the full range of values of *θ* that either expert considers reasonable. In general, linear opinion pool has been widely used in practice, while logarithmic opinion pool has been largely ignored, perhaps due to unrealistically strong aggregated beliefs.

5.2.3 Cooke's Method

Cooke's method [33] is a more complex mathematical aggregation. Cooke proposes that, in choosing weights for an opinion pool, it is desirable to weight highly those experts who are perceived to have more expertise. Cooke advocates assigning the weights on the basis of the performance of each expert in a separate elicitation exercise utilizing *seeding variables* (quantities that are chosen from the same subject area as the uncertain quantity of interest). His method is summarized by O'Hagan et al. [16] and outlined below.

The true values of the seeding variables are known to the facilitator, but not to the experts, and the experts are then asked to give various quantiles for these variables (typically 5th, 50th, and 95th percentiles). Experts are down-weighted for poor performance relative to others in two aspects: if their probabilities are poorly calibrated or if their distributions are wide so that they give little information.

The weight for expert j is proportional to the product of a calibration component (C_j) and an information component (K_j). These components are based on the idea of the Kullback-Leibler (K-L) distance between two discrete probability distributions:

Let $\mathbf{p} = \{p_{1,} p_{2,\ldots\ldots,} p_m\}$ and $\mathbf{q} = \{q_{1,} q_{2,\ldots\ldots,} q_m\}$ be two probability distributions for a discrete random variable taking m possible values. Then the K-L distance between them is:

$$I(\mathbf{p},\ \mathbf{q}) = \sum_{i=1}^{m} p_i \ln\left(p_i / q_i\right).$$

The distance is 0 if $\mathbf{p} = \mathbf{q}$, otherwise it is positive. The elicited quantiles define regions with fixed elicited probabilities $\mathbf{q}$. For example, if expert j provides the 5th, 50th, and 95th percentiles for each seeding variable, then $\mathbf{q}$ = {0.05, 0.45, 0.45, 0.05}. The calibration component C_j in Cooke's weighting system is based on $I(\mathbf{p}_j,\ \mathbf{q}_j)$, where $\mathbf{p}_j$ is the proportion of true values of the seeding variables that fall in each of the four regions elicited from expert j. Cooke notes that if n seeding variables are used, then $2n\ I(\mathbf{p}_j,\ \mathbf{q}_j)$ has approximately a chi-square distribution if the expert is perfectly calibrated, and he defines C_j to be the probability that such a chi-square random variable would exceed the observed value of $2n\ I(\mathbf{p}_j,\ \mathbf{q}_j)$. Poor calibration will lead to a small value of this weight component.

The information component K_j is defined by comparing the expert's distribution for each seeding variable with a uniform distribution (or a log-uniform distribution, where appropriate). A more informative distribution will be far from uniform, placing concentrations of probabilities on relatively short ranges. So K_j is the average value, over the n seeding variables, of the K-L distance between the expert's distribution and a uniform distribution, therefore giving more weight to more informative experts.

Cooke and Goossens [34] provide empirical evidence to show that this method does improve the overall performance of the elicitation. The use of seeding variables also indicates the likely performance of the elicitation for the unknown variable of interest. Cooke and Goossens [35] also report very positive results with the method using the elicitation of quantiles for seeding variables to construct weights. The performance-weighted average generally calibrates as well as or better than either the best individual expert or the equal-weighted average on the seeding variables.

To choose the right approach in mathematical aggregation, one should consider the type of information available (full or partial probability distributions); the individuals performing the aggregation of probabilities; the degree of modeling to be undertaken (e.g. risk assessment team); the form of the combination rule (e.g. weighted average); the specification of parameters for the combination rule (e.g. weights); and the consideration of simple vs. complex rules (e.g. simple averages vs. complex models).

5.3 Discussion

The debate over mathematical versus behavioural aggregation has yet to be resolved, with no evidence and conclusion to favour one method over the other. In particular, the simple average of distributions (equal-weighted linear opinion pool) from a number of experts provides a simple, robust, general method for aggregating expert knowledge. However, in the area of mathematical aggregation, the more complex aggregation like Cooke's method [33] has the potential to extract more information through unequal weighting or group synergy, but its success depends on having a well-structured and very thorough elicitation protocol. Experts should be chosen to have a broad range of substantive knowledge without undue overlap [16].

In the area of behavioural aggregation, group elicitation has even more potential, since it can bring better synthesis and analysis of knowledge through the group interaction. It can be successful if embedded in a well-structured process that manages the interaction through careful facilitation and strong feedback. Success is essentially dependent on the abilities of the facilitator, who must:

a. encourage the sharing of knowledge (as opposed to opinions);

b. encourage the recognition of expertise;

c. encourage the integration of feedback;

d. avoid domination of the group by shared knowledge or over-strong opinions;

e. avoid the biases found in individual assessments; and

f. avoid the tendency of groups towards overconfidence.

Clemen and Winkler [26] conclude that an overall aggregation process could involve both mathematical and behavioural aspects, and no single process is best in all circumstances. An understanding of the advantages and disadvantages of both methods is the key to design the best combination process. A *combined probability aggregation approach* can ideally be viewed as representing a summary of the current state of expert opinion regarding the uncertainty of interest. Such an approach would have the following steps:

1. Each expert gives an estimate on a parametric value in the form of percentiles (5%- and 95%-percentiles are elicited and marked on generic logarithmic scale) ;

2. Experts discuss reasons for a possible deviation in the percentile values;

3. Experts try to agree on common percentile values;

4. If consensus not achieved, mathematical aggregation is performed with the consent of the experts; and

5. If consent not provided, more information related to the parameter has to be acquired.

There are six important issues that should be considered: the selection of experts, flexibility and process design, the role of modeling versus rules, the role of interaction, the role of sensitivity analysis, and the role of the risk-assessment team. Details of each are provided by Clemen and Winkler [26].

6 Conceptual Framework for Quality Verification of Risk Assessment

Rosqvist [3] develops a conceptual framework for the *qualification*, or *quality verification*, of risk assessment that is built around concepts such as quality characteristics, independent peer review and related tasks (quality verification tasks), and ensuring accountability. A definition is proposed, based on Rosqvist and Tuominen [36]: *the qualification of risk assessment is an independent review process consolidating the decision-maker's confidence in the results and recommendations of the risk assessment*.

In essence, the qualification process verifies certain quality characteristics in the risk assessment. The following are a set of four methodological qualities that are deemed relevant in risk informed decision-making: *completeness, credibility, transparency* and *fairness*. Each of these qualities is implied in at least one of the *quality criteria* that should be met for each step in the risk assessment process. Rosqvist [3] argues that the qualification criteria should be defined such that:

a) verifying them entails a binary judgement of "yes" or "no";

b) they are mutually exclusive;

c) they are unambiguous supporting consensus judgements; and

d) their fulfillment is a necessary condition for qualification.

It is important to note that when defining specific quality criteria one needs to relate them to concrete risk assessment tasks, in order to satisfy the points a) – c) above. This is illustrated in Table 2 (adapted from Rosqvist [3]), which synthesizes the four methodological quality characteristics and the associated qualification criteria in each step of risk assessment.

Although not part of the risk assessment, but rather a part of risk management, the 6th point is added to Table 2 to convey the purpose of the qualification exercise.

The verification of the fulfillment of a qualification criterion is undoubtedly judgemental. It depends on the verifier's experience on risk assessments. If consensus cannot be achieved in terms of "yes" or "no", it may imply that the risk assessment document is inadequate and more details should be provided.

During the qualification process, the inputs are the documented results related to each step of risk assessment, while the outputs are the qualification results and recommendations as judged by the assessment group. The decision-maker then makes decisions based on these risk assessment results and the quality verification results. He or she may either accept the risk assessment results or request refinements in the risk assessment based on the recommendations.

Table 2. *Qualification criteria and implied quality characteristics for each step in the risk assessment*

	RISK ASSESSMENT STEP	QUALITY CHARACTERISTICS	QUALIFICATION CRITERIA
1	Scope Definition	Transparency	verifies that the stakeholders have been informed about adopted decision rules and criteria
2	Hazard Identification	Completeness	verifies that the stakeholders' and the domain experts' feedback on the completeness of hazard identification process is adequately surveyed
3	Risk Estimation	Credibility	verifies that sensitivity studies, based on parameter uncertainty, are adequate; verifies that model uncertainty and direction of bias of risk model is adequately addressed
4	Risk Evaluation	Credibility	verifies that the conclusions drawn as based on the decision rules are consistent
5	Analysis of Options	Completeness	verifies that the stakeholders' and the domain experts' feedback on the completeness of risk control options is adequately surveyed
6	Recommendation for the decision-maker	Qualified Risk Assessment	A peer review has been conducted

Project funding is an important issue related to the interaction between the risk assessment and the qualification process. The allocated project resources as mandated by the decision-maker should be considered. If these resources affect or steer the qualification process, the assessment team may not be able to continue the verification.

7 Summary and Recommendations

This report presents a review of approaches for eliciting and aggregating expert judgements in the various stages of the risk assessment process. Expert judgement is deemed essential to minimizing and characterizing the substantial uncertainty around making certain decisions, particularly when historical data is sparse or unavailable.

To ensure a sound process, the actual elicitation should be preceded by a thorough preparation. Identifying variables of interest, expert selection, motivation and training are critical steps towards a successful elicitation with accurate results. Another essential prerequisite of a successful endeavour is a well-designed process with built-in flexibility and clear roles and rules, which provides for structured interaction and incorporates feedback. Experts should understand what is expected of them and how their judgement will be used. Reviewing available evidence with the experts will provide a basis to draw upon when making judgements. Sensitivity analysis also plays an important role in validating the outcome.

Expert judgements in risk assessment can be qualitative – such as stating assumptions, defining criteria for data selection, textual description of physical quantities – and/or quantitative – mostly in the form of probabilities, but also as ratings, odds, weighting factors, numerical estimates of physical quantities. The elicitation of subjective probabilities has been presented in this paper from both a psychological and a statistical perspective. Research in psychology shows that the human mind's ability to accurately judge probabilities is limited by biases and by the tendency to use heuristics. The most common biases and heuristics have been discussed in the paper, including ways to evaluate and suggestions on how to minimize their effect. The facilitator should recognize the potential for inadvertently introducing biases into the elicitation process, such as anchoring effects induced by a particular ordering of questions. Procedures such as training, reviewing the most common biases and the reasons for their occurrence, and practice elicitation runs are helpful because they explicitly encourage experts to think analytically.

From a statistician's perspective, the uncertainty related to the accuracy of elicited probabilities is an integral part of the analysis of expert judgement. The report has presented discussions on a number of mathematically rigorous methods for quantifying the accuracy of experts' predictions and formulating models using experts' opinions. Calibration curves are used to assess the agreement between the expert's judgement and reality, and scoring rules measure how well the elicited probabilities express the expert's underlying opinion. For improved accuracy, the experts should be provided with feedback and with the opportunity to review and improve their judgement. One of the checks that should be done on a set of subjective probabilities is coherence, or the degree to which they are consistent with the laws of probability (i.e. the probabilities associated with all possible independent outcomes of an event sum to one).

Also presented in the report are techniques for the aggregation of probability distributions from multiple experts, as a way to reduce random variation in a set of judgements. Aggregation methods fall under two general categories: behavioural aggregation and mathematical aggregation. In behavioural aggregation, an interaction is created between the groups of experts, through which a single distribution is elicited from the group as a whole. In mathematical aggregation, a single distribution is elicited from each expert individually and independently of the others and the resulting distributions are mathematically combined into a single distribution.

The combined distribution can be considered consensus and it is more practical to use in further analysis. In the area of mathematical aggregation, the simple average of distributions (equal-weighted linear opinion pool) from a number of experts provides a simple, robust, general method for aggregating expert knowledge, though a more complex aggregation like Cooke's method [33] has the potential to extract more information under certain conditions. In the area of behavioural aggregation, group elicitation has great potential, because it can bring better synthesis and analysis of knowledge through the group interaction. A successful process is essentially dependent on the abilities of the facilitator. The use of one type of aggregation over the other is dictated by factors such as the type of variables of interest; the type of available information; the experts' background and experience; and the design and structure of the elicitation protocol. No single method is best in all circumstances, and an overall aggregation process could involve both mathematical and behavioural aspects.

In addition to assessing the confidence associated with the experts' opinions, the report has reviewed methods for assessing the confidence in the overall elicitation process. The report ends with a brief presentation of a conceptual framework for the quality verification of risk assessment. In essence, the qualification process is an independent review process that verifies certain quality characteristics in the risk assessment, such as completeness, credibility, transparency and fairness. This process aims to consolidate the decision-maker's confidence in the results and recommendations of the risk assessment. The quality verification results may influence the decision-maker to either accept the risk assessment results or request refinements in the risk assessment based on the recommendations.

References

1. Claycamp, H.G. (2006) Rapid Benefit-Risk Assessments: No Escape from Expert Judgements in Risk Management. *Risk Analysis*, **26**, 147-156.

2. O’Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., P.H., G., Jenkinson, D.J., Oakley, J.E. and Rakow, T. (2006) *Uncertain Judgements: Eliciting Experts’ Probabilities*. John Wiley & Sons, Ltd.

3. Rosqvist, T. (2003) *On the Use of Expert Judgement in the Qualification of Risk Assessment*. VTT Publications 507, Espoo, Finland.

4. Hora, S.C. and von Winterfeldt, D. (1997) Nuclear Waste and Future Societies: A Look into the Deep Future. *Technological Forecasting and Social Change*, **56**, 155-170.

5. O'Hagan, A. (1998) Eliciting Expert Beliefs in Substantial Practical Applications. *The Statistician*, **47**, 21-35.

6. Garthwaite, P.H., Kadane, J.B. and O'Hagan, A. (2005) Statistical Methods for Eliciting Probability Distributions. *Journal of the American Statistical Association*, **100**, 680-701.

7. Gilovich, T., Griffin, D. and Kahneman, D. (2002) *Heuristics and Biases: The Psychology of Intuitive Judgement*. Cambridge University Press, Cambridge, UK.

8. Daneshkhah, A. (2004) Psychological Aspects Influencing Elicitation of Subjective Probability. *Bayesian Elicitation of Experts' Probabilities*. University of Sheffield.

9. Slovic, P., Fischhoff, B. and Lichtenstein, S. (1979) Rating the Risks: The Structure of Expert and Lay Perceptions. *Environment*, **21**, 36-39.

10. Gilovich, T. (1991) *How we know what isn't so: The fallibility of human reasoning in everyday life*. The Free Press, New York.

11. Tversky, A. and Kahneman, D. (1983) Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgement. *Psychological Review*, **90**, 293-315.

12. Gigerenzer, G. and Hoffrage, U. (1995) How to improve Bayesian reasoning without instruction: frequency formats. *Psychological Review* **102**, 684-704.

13. Anderson, J.L. (1998) The Interface of Bayesian Statistics and Cognitive Psychology. *Conservation Ecology*, **2**.

14. Fischhoff, B., Slovic, P. and Lichtenstein, S. (1978) Fault Trees: Sensitivity of Estimated Failure Probabilities to Problem Representation. *Journal of Experimental Psychology: Human Perception and Performance*, **4**, 330-344.

15. Fox, C.R. and R.T., C. (2005) Subjective Probability Assessment in Decision Analysis: Partition Dependence and Bias toward the Ignorance Prior. *Management Science*, **51**, 1417-1432.

16. O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., P.H., G., Jenkinson, D.J., Oakley, J.E. and Rakow, T. (2006) *Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley & Sons, Ltd.

17. Hora, S.C. (2004) Probability Judgements for Continuous Quantities: Linear Combinations and Calibration. *Management Science*, **50**, 597-604.

18. Lichtenstein, S., Fischhoff, B. and Phillips, L.D. (1982) Calibration of probabilities: The state of art to 1980. In Kahneman, D., Slovic, P. and Tversky, A. (eds.), *Judgement under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, pp. 306-334.

19. Griffin, D. and Brenner, L. (2004) Perspective on Probability Judgement Calibration. In Koehler, D. and Harvey, N. (eds.), *Blackwell Handbook of Judgement and Decision Making*. Blackwell Pub., Oxford, UK.

20. Plous, S. (1993) *The Psychology of Judgement and Decision Making*. McGraw-Hill.

21. Murphy, A.H. and Winkler, R.L. (1984) Probability Forecasting in Meteorology. *Journal of the American Statistical Association*, **79**, 489-500.

22. Christensen-Szalanski, J.J. and Bushyhead, J.B. (1981) Physicians' Use of Probabilistic Information in a Real Clinical Setting. *Journal of Experimental Psychology: Human Perception and Performance*, **7**, 928-935.

23. de Finetti, B. (1964) Foresight: Its logical laws, its subjective sources. In Kyburg, H.E.J. and Smokler, H.E. (eds.), *Studies in subjective probability*. John Wiley & Sons, New York, pp. 53-118.

24. Murphy, A.H. (1973) Hedging and skill scores for probability forecasts. *Journal of Applied Meteorology*, **12**, 215-223.

25. Winkler, R.L., Hora, S.C. and Baca, R.G. (1992) The Quality of Expert Judgement Elicitations. Center for Nuclear Waste Regulatory Analyses, San Antonio, TX.

26. Clemen, R.T. and Winkler, R.L. (1999) Combining Probability Distributions from Experts in Risk Analysis. *Risk Analysis*, **19**, 187-203.

27. Phillips, L.D. and Phillips, M.C. (1990) Facilitated Work Groups: Theory and Practice. London School of Economics and Political Science.

28. Rowe, G. and Wright, G. (1999) The Delphi Technique as a Forecasting Tool: Issues and Analysis (with discussion). *International Journal of Forecasting*, **15**, 353-381.

29. Delbecq, A.L., Van de Ven, A.H. and Gustafson, D.H. (1975) *Group techniques for program planning*. Scott Foresman Co., Glenview, IL.

30. Kaplan, S. (1992) “Expert Information” vs. “Expert Opinions”: Another Approach to the Problem of Eliciting/Combing/Using Expert Knowledge in PRA. *Reliability Engineering and System Safety*, **35**, 61-72.

31. Reagan-Cirincione, P. (1994) Improving the Accuracy of Group Judgement: A Process Intervention Combing Group Facilitation, Social Judgement Analysis, and Information Technology. *Organizational Behavior and Human Decision Processes*, **58**, 246-270.

32. Morris, P.A. (1974) Decision Analysis Expert Use. *Management Science*, **20**, 1233-1241.

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

34. Cooke, R.M. and Goossens, L.H. (2000) Procedures Guide for Structured Expert Judgement in Accident Consequence Modelling. *Radiation Protention Dosimetry*, **90**, 303-309.

35. Cooke, R.M. and Goossens, L.H. (2006) TU Delft Expert Judgement Data Base. Delft University Technology, Delft.

36. Rosqvist, T. and Tuominen, R. (2004) Qualification of Formal Safety Assessment – An Exploratory Study. *Safety Science*, **42**, 99-120.

37. De Wispelare, A.R., Herren, L.T. and Clemen, R.T. (1995) The Use of Probability Elicitation in the High-level Nuclear Waste Regulation Program. *International Jounral of Forecasting*, **11**, 5-24.

38. Gustafson, D.H., Sainfort, F., Eichler, M., Adams, L., Bisognano, M. and Steudel, H. (2003) Developing and Testing a Model to Predict Outcomes of Organizational Change. *Health Services Research*, **38**, 751-776.

This page intentionally left blank.

Annex A

There are many case studies in which researchers have sought to elicit subjective experts' opinions. The following two examples of elicitation come from the nuclear industry and a business study.

A.1 Nuclear Industry

De Wispelare et al. [37] discussed the use of probability elicitation in the high-level nuclear waste regulation program. They evaluated the mechanics of eliciting expert judgements to explore techniques for generating and aggregating probabilistic judgements of future conditions at the proposed high-level waste repository at Yucca Mountain, Nevada. An actual elicitation was conducted as an aid to these evaluations. Distributions of the change in temperature were elicited during seven discrete periods of time in the vicinity. Future climate in the vicinity was selected as the topic for elicitation. Personnel from the Nuclear Regulatory Commission (NRC) and Center for Nuclear Waste Regulatory Analyses (CNWRA) defined the climatic parameters of interest in conjunction with a panel of five expert climatologists.

The formal expert elicitation procedure consisted of 11 steps: (i) determine the objectives and goals of the study; (ii) recruit the experts; (iii) identify issues and information needs; (iv) provide initial data to the experts; (v) conduct the elicitation training session; (vi) discuss and refine the issues; (vii) provide a multi-week study period; (viii) conduct the elicitations; (ix) provide post-elicitation feedback to the experts; (x) aggregate the experts' judgements; and (xi) document the process.

The elicitation process included a period of a few weeks where the selected five experts were given the opportunity to do some work on the data and read relevant literature before their opinions were sought. The elicitations were done using the variable interval method, with between five and nine quantiles being elicited from each expert for each time period. The cumulative distribution functions and probability density functions of the distributions were then produced.

After the individual elicitations, a group session was conducted to explore aggregation of the experts' opinions by both behavioral and mathematical methods. Four exercises are conducted in an attempt to gain consensus:

i. giving the experts their five individual distributions;

ii. giving them mathematically aggregated distributions using four different methods;

iii. offering a mixture of the two; and

iv. allowing discussion of substantive issues where there was disagreement between the experts.

In the first exercise, consensus was reached quite easily, but it proved more difficult to attain in the remaining three exercises. The experts felt that they needed to have a number of distribution curves in order to represent the range of opinions for the second method. For the third exercise, a consensus distribution could not be agreed upon, but a solution was proposed that involved taking points from the three different distribution curves. Similarly, in the fourth method, the experts could not agree on the size of the effects, with three separate curves being agreed upon in the end.

One of the goals of this study was to demonstrate that expert elicitation can be utilized. The five experts reported little difficulty in representing their judgements as probability distributions. They felt that the training session was essential to acquaint them with the process. They were also comfortable with generating and evaluating cumulative distribution functions.

Another objective was to explore the aggregation of individual expert judgements. Once experts have publicly provided their individual analyses, they rarely admitted a change. The experts agreed to modifications and consensus distributions only on the pretext of establishing a group opinion. They indicated that they did not actually change their minds about their own stance.

A.2 Business Study

Gustafson et al. [38] used a Bayesian network model and employed a full elicitation process to predict the outcomes of organizational change. For the model development, experts' subjective assessments were elicited using an integrative group process. A panel of theoretical and practical experts and literature in organizational change were used to identify factors predicting success and failure of health care improvement projects.

In this study, experts' subjective assessment data were collected from 221 health care improvement projects in the United States, Canada, and the Netherlands in between 1996 and 2000. Experts were first nominated based on their knowledge of organizational change, recognition by their peers, and ability to work effectively in a group. Those receiving the most nominations were interviewed by phone and asked to identify factors.

The authors also reviewed the literature to identify definitions of success, factors, and levels of performance. After combining the interview and literature data into a non-duplicative taxonomy of more than one hundred factors and possible measures, the authors conducted a face-to-face meeting with the panelists to agree on the definitions, factors, and ways of measuring each factor. They also estimated the parameters of the Bayesian model.

To develop the Bayesian model, the panel of experts estimated the probability of implementation success using subjective estimates of likelihood ratios and prior odds. Firstly, the authors elicited ideas from the experts about what implementation success of the organizational change would mean to that individual. Secondly, they investigated which factors each expert felt would best predict the success or failure of the implementation. Once these two inputs have been collated, the conditional independence between factors was elicited. After that, experts' opinions were elicited in the form of likelihood ratios by responding to the following questions:

"Think about two healthcare improvement projects. One was successfully implemented and the other was not. Which one is more likely to have the following characteristics? How much more likely? (a lot, somewhat, barely)"

In this way, a factor level is specified, such as “a staff that hates the current situation and believes that change is essential.” Both verbal and numerical responses were obtained. A subsequent retrospective empirical analysis of change efforts in 198 health care organizations was performed to validate this model.

Gustafson et al. [38] concluded that the subjective Bayesian model is effective in predicting the outcome of actual improvement projects. Additional prospective evaluations, as well as testing the impact of this model as an intervention, are warranted.

A.3 Other Examples

Cooke [33] and O’Hagan et al. [16] reviewed many developments over the years as attempts have been made to use expert judgements in various settings. To illustrate the diversity of the areas where expert elicitation has found application (from nuclear engineering, aerospace, various types of forecasting, to military intelligence and seismic risk), we provide some examples below:

- Defence (e.g. assess the risk of terrorist attacks)
- Medicine (e.g. diagnosis and treatment decisions, clinical trials, survival analysis, clinical psychology)
- Veterinary Science (e.g. animal disease diagnosis system)
- Agriculture (e.g. farmers assess rice crop yields)
- Meteorology (e.g. weather forecasting)
- Economics (e.g. opinions about future weekly earnings, effects of discount sales in the retail industry)
- Finance (e.g. interest rates and inflation forecasts, stock prices forecast)
- Environment, Conservation, and Ecology (e.g. risk from toxic chemicals, radioactive groundwater contamination and soil lead contamination)
- Public Health (e.g. level of personal exposure to benzene)
- Psychology (e.g. probability that a patient’s depression status after a treatment)
- Engineering (e.g. structural safety)
- Law (e.g. make legal judgements)
- Sport (e.g. predict football scores over a series of weeks)
- Archaeology (e.g. prior information in archaeology and chronology building)
- Game Theory (e.g. elicit probabilities of the other players in a game)
- Demography (e.g. explore the Iraqi Kurdish population)
- Emergency Services (e.g. firefighting)
- Maps (e.g. pixel classification for thematic maps)
- Maritime Safety

- Maintenance Management
- Software Reliability

List of symbols/abbreviations/acronyms/initialisms

AHRA	All-Hazards Risk Assessment
CSS	Centre for Security Science
CORA	Centre for Operational Research and Analysis
DND	Department of National Defence
DRDC	Defence Research & Development Canada
IACC	Intelligence Assessment Coordination Committee
IAS	International Assessment Staff
IEG	Intelligence Experts Group
PCO	Privy Council Office
RCMP	Royal Canadian Mounted Police
R&D	Research and Development
S&T	Science and Technology

Distribution list

Document No.: DRDC CORA TM 2007-57

	LIST PART 1: Internal Distribution by Centre
1	NORAD OR
1	MARLANT N02OR
1	MARPAC N02OR
1	DGLCD OR Team – Fred Cameron
1	RCMP – Wendy Nicol
1	PS – Mike MacDonald
1	CDA – Yves Goulet
6	SPORT
1	DGFDA
2	CFEC: Cmdt/Dr. M. Dixson
1	CFMWC
1	DOS SJS
1	ADM S&T
1	DRDC CapDEM TDP PM
2	JCDS21 PM/Dr. K. Wheaton
1	DG CORA
1	CS CORA
1	SH J&C & Teams & SH Strategic Analysis & Teams
1	SH Maritime, Land, Air & Teams
1	DG DRDC CSS – DG, D PSTP & D CRTI
1	CORA Library (1 PDF)
1	DRDKIM (1 PDF)
1	NDHQ Library (1 PDF)
6	Royal Military College – Dean of Arts + 1 PDF
2	Authors (1 HC & 1 PDF)
6	Spares
44	TOTAL LIST PART 1

	LIST PART 2: External Distribution by DRDKIM
1	Library and Archives Canada
70	TOTAL LIST PART 2

45 TOTAL COPIES REQUIRED

DOCUMENT CONTROL DATA

(Security classification of title, body of abstract and indexing annotation must be entered when the overall document is classified)

1. ORIGINATOR (The name and address of the organization preparing the document. Organizations for whom the document was prepared, e.g. Centre sponsoring a contractor's report, or tasking agency, are entered in section 8.)

Defence R&D Canada Centre for Operational Research & Analysis
National Defence Headquarters, 6 CBS
101 Colonel By Drive
Ottawa, ON, K1A 0K2 Canada

2. SECURITY CLASSIFICATION (Overall security classification of the document including special warning terms if applicable.)

UNCLASSIFIED

3. TITLE (The complete document title as indicated on the title page. Its classification should be indicated by the appropriate abbreviation (S, C or U) in parentheses after the title.)

Expert Judgement in Risk Assessment:

4. AUTHORS (last name, followed by initials – ranks, titles, etc. not to be used)

Leung, K.; Verga, S..

5. DATE OF PUBLICATION (Month and year of publication of document.)

December 2007

6a. NO. OF PAGES (Total containing information, including Annexes, Appendices, etc.)

50

6b. NO. OF REFS (Total cited in document.)

32

7. DESCRIPTIVE NOTES (The category of the document, e.g. technical report, technical note or memorandum. If appropriate, enter the type of report, e.g. interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered.)

Technical Memorandum

8. SPONSORING ACTIVITY (The name of the department project office or laboratory sponsoring the research and development – include address.)

Defence R&D Canada Centre for Operational Research & Analysis
National Defence Headquarters
101 Colonel By Drive
Ottawa, ON, K1A 0K2 Canada

9a. PROJECT OR GRANT NO. (If appropriate, the applicable research and development project or grant number under which the document was written. Please specify whether project or grant.)

9b. CONTRACT NO. (If appropriate, the applicable number under which the document was written.)

10a. ORIGINATOR'S DOCUMENT NUMBER (The official document number by which the document is identified by the originating activity. This number must be unique to this document.)

DRDC CORA TM 2007-57

10b. OTHER DOCUMENT NO(s). (Any other numbers which may be assigned this document either by the originator or by the sponsor.)

11. DOCUMENT AVAILABILITY (Any limitations on further dissemination of the document, other than those imposed by security classification.)

Unrestricted access

12. DOCUMENT ANNOUNCEMENT (Any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availability (11). However, where further distribution (beyond the audience specified in (11) is possible, a wider announcement audience may be selected.))

13. ABSTRACT (A brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is highly desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), (R), or (U). It is not necessary to include here abstracts in both official languages unless the text is bilingual.)

Decision and risk analysis models often require both qualitative and quantitative assessments of uncertain events; in many cases, expert knowledge is essentially the only source of good information. Over the last decade, uncertainty analysis has become an increasingly important part of operations research models. The growing use of risk assessment in government and corporate planning and operations has also increased the role of expert judgement in providing information for decision making.

Elicitation of experts' opinions is frequently used to support decision making in many different areas, from forecasting in the financial world to assessing the risk of terrorist attacks in the national security domain. The use of expert judgements has provoked questions related to the practice of utilizing experts' opinions and to the accuracy of the obtained results. This work reviews some approaches for eliciting and aggregating expert judgements as inputs into the risk assessment process, and looks at methods of assessing the degree of confidence associated with these subjective inputs, as well as confidence in the overall process.

The research synthesized in this report outlines the elicitation process and highlights both its statistical and psychological perspectives. It looks at ways to evaluate the accuracy of elicitation; it presents techniques for the aggregation of probability distributions from multiple experts; and it summarizes a conceptual framework for the quality verification of risk assessment. Two examples of the application of formal elicitation in the nuclear industry and a business study are also discussed in the Appendix.

14. KEYWORDS, DESCRIPTORS or IDENTIFIERS (Technically meaningful terms or short phrases that characterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus, e.g. Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus identified. If it is not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title.)

Risk Assessment, risk analysis, expert judgement elicitation, subjective probabilities

DEFENCE DÉFENSE

DRDC CORA

www.drdc-rddc.gc.ca