International Series in
Operations Research & Management Science

Anca M. Hanea
Gabriela F. Nane
Tim Bedford
Simon French *Editors*

Expert Judgement in Risk and Decision Analysis

Springer

International Series in Operations Research & Management Science

Volume 293

More information about this series at http://www.springer.com/series/6161

Anca M. Hanea · Gabriela F. Nane ·
Tim Bedford · Simon French
Editors

Expert Judgement in Risk and Decision Analysis

Springer

Editors
Anca M. Hanea
Center of Excellence for Biosecurity
Risk Analysis (CEBRA)
University of Melbourne
Melbourne, VIC, Australia

Tim Bedford
Department of Management Science
University of Strathclyde
Glasgow, UK

Gabriela F. Nane
Delft Institute of Applied Mathematics
Delft University of Technology
Delft, The Netherlands

Simon French
Department of Statistics
University of Warwick
Coventry, UK

ISSN 0884-8289 ISSN 2214-7934 (electronic)
International Series in Operations Research & Management Science
ISBN 978-3-030-46473-8 ISBN 978-3-030-46474-5 (eBook)
https://doi.org/10.1007/978-3-030-46474-5

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

In front of the reader lies a book on Structured Expert Judgement (SEJ). A close look at the contents of the chapters shows not only a large variety of applications in practice, but also new developments and directions in expert judgement theories. We should not forget that this SEJ modelling has been developed over the past 40 years. Many scientists have worked with the Classical Model of Structured Expert Judgement, and also with other modelling approaches as, for instance, Bayesian techniques.

But what is most interesting is that one particular scientist at the Delft University of Technology in the Netherlands is actually leading the SEJ technique: Prof. Dr. Roger M. Cooke. And he has been for almost 40 years. In the early years, Roger developed expert judgement methodologies from a more philosophical point of view. In 1991, he published his book on the subject called "Experts in Uncertainty".

In practical scientific and engineering contexts, certainty is achieved through observation, and uncertainty is that which is removed by observation. Quantitative studies must be provided with a mathematical representation of probability. The experts' assessments properly cast in probability distributions, and combining experts' judgements enhances rational consensus. Rational consensus is the carrier of the performance-based expert judgement studies, by which the experts' assessments are combined using different performance weights for each expert. In the last decade of the twentieth century, Roger developed a procedure guide for applying SEJ in practical applications. EUR 18820 is the report of the large European Community Research Project of the EU Nuclear Science and Technology Division.

From the year 2000 up till now, efforts were used to develop the Classical Model further, and to experience with potential predictive issues, such as volcano eruptions, seismic events, and long-term effects of, for instance, cancers. I wish Roger Cooke and his enthusiastic co-workers a productive and enjoyable next period. And remember, if anything goes wrong, it might be caused by the last beer.

Delft, The Netherlands Louis Goossens

Contents

Part III Process, Procedures and Education

Part IV Applications

Chapter 1
Introduction and Overview of Structured Expert Judgement

Simon French, Anca M. Hanea, Tim Bedford, and Gabriela F. Nane

1.1 Background

Although we live in a data-rich age, it is not true that we have or ever will have sufficient data to evaluate all potential future events, risks or opportunities. Some will have novel or unexplored characteristics: for instance, the medium- and long-term socio-economic effects of Brexit, the effects of plastic waste in the aqueous environment or a future pandemic. In such cases, there are simply too few data on which solely to base useful quantitative assessments of risk and we need to look to experts for guidance. Of course, any risk or decision analysis relies on expert judgement to some extent. Data necessarily refer to events or entities in the past or immediate present, so a judgement is needed that they are relevant to the prediction of any risk or opportunity in the future. Expert judgements are also needed to select appropriate models and analytic methods, to interpret the output of an analysis and to assess whether it provides sufficient guidance to implement risk management strategies or make a decision. Although such topics will be touched on in the following chapters, they are not the prime focus of this collection of readings. Rather our concern is with the use of expert judgement to provide quantitative probabilistic assessments of key uncertainties in an analysis when empirical data are unavailable, incomplete, uninformative or conflicting.

S. French (✉)
University of Warwick, Coventry, England
e-mail: simon.french.50@gmail.com

A. M. Hanea
University of Melbourne, Melbourne, Australia

T. Bedford
University of Strathclyde, Glasgow, Scotland

G. F. Nane
Delft University of Technology, Delft, Netherlands

A. M. Hanea et al. (eds.), *Expert Judgement in Risk and Decision Analysis*,
International Series in Operations Research & Management Science 293,
https://doi.org/10.1007/978-3-030-46474-5_1

Theoretical studies of how to coherently and systematically use expert judgement go back at least to the 1960s and arguably a century or two before that (French 1985; Cooke 1991). Practical studies really began in the second half of the 1980s: see Chap. 9. Despite such a long history and the fact that expert judgement is routinely required to inform critically important decisions across many domains, too often it is obtained by canvasing the judgement of one expert alone, or by using ill-informed and inappropriate elicitation and aggregation methods. Instead, best practice requires the use of panels of experts, structured elicitation protocols and aggregation methods that recognise the complexity of human judgements.

Some organisations, e.g. the European Food Safety Authority (EFSA), have established full protocols for drawing expert judgement into their analyses and working practices in a structured and explicit manner (EFSA 2014). More organisations need to follow EFSA's lead, but currently there is undoubted growth in the number of risk and decision studies that use such methods.

We shall emphasise the importance of expert judgement studies being *structured* and *explicit*. It is easy to ask anyone, expert or not, for their judgement and, by and large, they will give it. But it is not so easy to do this in a way that encourages a thoughtful, auditable and relevant answer that is not affected or biased in some way by the giver's psychology. Moreover, when more experts are asked, seldom will they give the same answer. So how should we combine them? Should we give them 'equal weight' in some sense or perhaps give greater importance to those with either acknowledged or assessed expertise? Should we allow interaction and provide feedback? Whatever we do, those words 'structured' and 'explicit' tell us that we should do so in a careful, auditable, fully reported way. As scientists, we are well versed in how to report data and analyses from empirical studies so that they are clear, open to peer review and allow repetition in validation studies. How do we report the process of gathering the relevant judgements of experts so that they can form some or all the evidence in an analysis?

Those, in brief, are the topics that we shall be surveying in this overview chapter. The readings that follow will flesh out many of the topics through theoretical and methodological discussions and case studies. In the next section, we set the context a little further, by categorising different contexts in which structured expert judgement (SEJ) may be used. Section 1.3 discusses how judgements should be elicited from experts. As we have suggested, simply asking them risks answers biased by potentially flawed thinking. The judgements of several experts can be aggregated in several ways. They can be elicited individually and then combined by some mathematical process or they can be elicited consensually from the group through qualitative discussion. These are the topics of Sects. 1.4 and 1.5, respectively. In Sect. 1.6, we consider how SEJ studies should be reported. Although we believe that SEJ is now a mature technology that can be—and indeed has been—applied in many complex risk and decision analyses, there are still many areas requiring further research and development. We indicate some of these in Sect. 1.7. Finally, in Sect. 1.8, we give an overview of the following chapters.

1.2 Contexts

Experts may be consulted for their advice on risks and uncertainties in a number of contexts. French (1985) introduced three broad categories: the *expert problem*, the *group decision problem* and *textbook problem*, though often, individual problems reflect aspects of two or all of these.

- In the *expert problem*, a group of experts are asked for advice from a problem-owner or decision-maker, external to the group. Responsibility and accountability for the potential consequences rest with that person and the experts are free from those, and the many pressures that might bias their judgements. In this context, the emphasis is on the decision-maker learning from the experts.
- In the *group decision problem*, the group itself is jointly responsible and accountable for the decision. They are their own experts and are both experts and decision-makers. They have responsibility and accountability for the decision. The group may, and indeed probably will, wish that their actions appear rational, fair and democratic. Thus, they may wish to combine their judgements in some formal structured way; but in voting, each will surely wish to be guided by their own personal views having heard the opinions of the others.
- In the *textbook problem,* a group of experts may simply be consulted for their judgements for others to use in as yet undefined circumstances. Here, there is no predefined risk or decision problem, but many potential ones that are as yet only vaguely perceived. An example here is the reports of the Intergovernmental Panel on Climate Change.

The careful distinctions between the roles of expert and decision-maker implicit in the above descriptions are important. Whatever the context, the experts are asked for their opinion on the *science* of the situation: either how likely something is to happen or their subjective estimate of an unknown quantity. They are not asked for *value* or *preference* judgements. This reflects recognised practice in the relationship between science advisors and decision-makers, the scientific culture of being evidence-led, and also a technical perspective from the theory of rational decision-making in which uncertainties and value judgements are separate components of decision-making under uncertainty.

The expert and group decision problems have at their heart a specific risk or decision problem; the third does not. This specific focus provides a structure against which possible approaches can be judged—which does not imply that their resolution needs to be similar. What may be appropriate to one problem may be less suited to the other. For instance, in the expert problem, it seems entirely appropriate for the decision-maker to process the experts' judgements if she believes that they may be biased in some way, i.e., poorly calibrated. In the case of the group decision problem, equity arguments would suggest that the group members should be able to vote according to their best beliefs even if others believe them to be poorly calibrated. Similarly, arguments drawing on democratic principles may suggest that all experts should be treated equally in the group decision problem, whereas in the expert problem, it may not be reasonable to assume that all experts are equally knowledgeable and, hence, the decision-maker might weigh them differently.

Both the expert and the group decision problems have been well explored with many theoretical and methodological contributions over the years (see, e.g. Clemen and Winkler 1999; Cooke and Goossens 2000; O'Hagan et al. 2006; Hora 2008; Burgman 2015; Dias et al. 2018). However, little has been written on the textbook problem, perhaps because its lack of structure makes it more difficult to address. Developments in web-based public participation, stakeholder engagement and deliberative democracy, however, are giving the topic some importance (French 2012), as different groups seek to draw on previous SEJ studies to provide evidence in a different context.

Cooke (1991) argues from a methodological point of view that in all the above contexts, the goal of an SEJ elicitation is to enhance rational consensus which is attainable if problem owners commit in advance to the way the experts' views are elicited and aggregated. He formulates necessary conditions for rational consensus in the form of four principles: *Scrutability/Accountability*, "all data, including experts' names and assessments, and all processing tools should be open to peer review and results must be reproducible by competent reviewers"; *Empirical Control*, "quantitative expert assessments should be subject to empirical quality controls"; *Neutrality*, "the method for combining and evaluating expert opinion should encourage experts to state their true opinions, and must not bias results"; *Fairness*, "experts should not be prejudged, prior to processing the results of their assessments".

These principles are operationalised in the Classical Model for SEJ which will be discussed in more detail in Sect. 1.4.3 and in Chap. 10 of this book. Even though sometimes criticised (e.g. French 2011), no alternative principles have been formulated; hence, they remain important guidelines for SEJ protocols. The *Scrutability/Accountability* principle is essential in how SEJ studies should be reported, and it will be further discussed in Sect. 1.6.

Throughout we assume that the experts are asked to quantify their uncertainty using probabilities, or numerical estimates corresponding to quantiles of probability distributions. We recognise that others have proposed different formalisms for encoding uncertainty numerically, but no other methodology has the power, axiomatic and empirical validity of probability theory. We do not assume, however, that experts are only asked for their numerical estimates. A good elicitation also gathers the experts' reasoning behind their statements and reports this too. Such qualitative material is as important to sound risk and decision analyses as their quantitative judgements, providing, for instance, a qualitative commentary on the validity of the models used.

Finally, while we have spoken of risks and decisions, we emphasise that SEJ methods are also important in Bayesian statistical inference in developing informative prior distributions for analyses (French and Rios Insua 2000).

1.3 Elicitation

To use SEJ, it is inevitable that the analyst asks the experts for their probabilities. That is a far more skilled task than at first might seem. The problem is that when anyone—experts included—is asked such a question, they may respond based on very superficial thinking. Behavioural scientists and psychologists have investigated how people frame and respond to questions relating to uncertainty, since Ward Edwards (1954) asked whether real people were as rational in their behaviour as economic and Bayesian theories of expected utility would suggest. Evidence quickly accumulated that in general they were not. Their behaviour was governed by many heuristic patterns of thought that could lead to judgements and actions that were systematically biased away from the assumptions underlying theories of rationality. Empirical behavioural studies identified many 'heuristics and biases', and this area of research became known under that title (Kahneman and Tversky 1974). Nowadays, one distinguishes System 1 Thinking and System 2 Thinking. The former refers to simple, fast, heuristic patterns of thought on the borders of consciousness; the latter to more conscious, slower, explicit and auditable analysis, one that can be tested against and corrected to be consistent with some norms of rationality (Kahneman 2011). A caveat: our description suggests a dichotomy between these two systems of thinking, but they may represent two ends of a scale with many forms of thinking between the two, moving from the subconscious to the conscious. Indeed, there is much debate within the psychological and behavioural sciences about the precise details of such systems of thinking (Evans and Stanovich 2013). The potential for subconscious patterns of thought to lead to irrationalities is unquestioned, however.

The heuristics of System 1 Thinking may lead to biased responses from the experts during elicitation of their probabilities. For instance, the *availability* heuristic leads individuals to overestimate the probability of events that are easily recalled because of horrific consequences. The *anchoring* heuristic suggests that if a question contains a potentially relevant number, the responder will anchor on that, giving a numerical reply biased towards it. Overconfidence is considered "the most significant of the cognitive biases" (Kahneman 2011) and can play a paramount role in expert's ability of quantifying uncertainty. We do not survey and summarise the many heuristics and biases that need to be taken into account during elicitation: there is a large literature doing precisely that (e.g. Kahneman and Tversky 2000; Gigerenzer 2002; Bazerman 2006; Kahneman 2011). Nor do we provide guidance on the forms of questioning that nudge experts into more System 2 forms of thinking and thus reduce the potential biases in their stated probabilities (see, e.g. for surveys: Wright and Ayton 1987; O'Hagan et al. 2006; Hora 2007). What we would emphasise is that elicitation is a skill that needs to be acquired from training and guided by mentoring; it is not easily developed simply by reading a textbook.

We would also note the importance of developing a detailed elicitation protocol *before* embarking on any elicitations from experts to ensure that all are treated in the same way. Elicitation protocols are as important in expert judgement studies as experimental designs are in empirical research.

1.4 Mathematical Aggregation

1.4.1 Introduction

Taking a rather simplistic view, there are two ways of producing a combined judgement from a group of experts. Firstly, we could ask each to give their assessments and then take their numerical arguments and combine them by some mathematical process. Secondly, we could ask them to discuss the uncertainties and provide a consensual numerical group judgement, thus aggregating their individual opinions behaviourally. Here, we discuss the former approach, leaving the latter to the next section. We confine attention in this section to opinion pools, Cooke's Classical Model and Bayesian approaches. Arguably, these span all the mathematical aggregation approaches though proponents of some approaches may prefer different terminologies. For wider reviews, see, i.e., French (1985), Genest and Zidek (1986) and Jacobs (1995).

Note that we do not discuss mathematical approaches to the group decision-making problem here, since that would lead us to review game theory, adversarial risk analysis and social choice literatures all of which have large and continually growing literatures (Osborne 2003; French et al. 2009; Banks et al. 2015; Sen 2017).

1.4.2 Opinion Pools

Opinion pools take a pragmatic approach. Assuming that a group of experts have each provided assessments for a number of different possible outcomes (typically we take outcomes that are exclusive and exhaustive, that is, one, and only one, of them has to occur), they simply average the individual expert's assessments. Intuitively if not conceptually, these approaches take the experts' judgements as probabilities in their own right. The process may use a weighted arithmetic or weighted geometric mean or perhaps something rather more general:

$$P_{DM} = \sum_{e=1}^{\mathrm{E}} w_e P_e \quad \text{or} \quad P_{DM} = \prod_{e=1}^{\mathrm{E}} P_e^{w_e} \quad \text{or} \quad P_{DM} = \phi(P_1, P_2, \ldots, P_E)$$

The P_e are the experts' probabilities indexed over the E experts, the numbers w_e are weights adding to 1 and the ϕ function denotes a general mathematical formula. The combined probability is subscripted DM for decision-maker. In their 'vanilla' form, the weights in an opinion pool are often simply given by the decision-maker or analyst based on their judgements of the experts' relative expertise, seldom with any operational meaning being offered for the concept of 'relative expertise'. Alternatively, the weights may be taken as equal, perhaps on some Laplacian Principle of Indifference, or of equity, or, even, on the basis that all the experts are paid the same.

A suggestion that the weights might be defined as some measure of the reputation and influence of the experts on social networks has been investigated but not found useful (Cooke et al. 2008).

There have been many attempts to investigate axiomatic justifications of opinion pools, requiring such properties as follows:

- *Marginalisation.* Suppose that the experts are asked for a joint distribution over (X, Y) say. Then the same result should be obtained for the marginal distribution over X whether the experts' distributions are marginalised before forming the combination or the combination formed and the result marginalised (McConway 1981).
- *Independence Preservation.* If all experts agree that (X, Y) are independent variables, the variables should remain independent in the combined distribution (French 1987; Genest and Wagner 1987).
- *External Bayesianity.* If relevant data become available, then the same result should be obtained whether the experts update their individual distribution through Bayes Theorem before they are combined or their distributions combined and the result then updated (Madansky 1964; Faria and Smith 1997).

If, for instance, one insists that any opinion pool should satisfy marginalisation, then one is limited essentially to weighted arithmetical, i.e., linear, pools (McConway 1981). But if one adds in other requirements such as the other two above or further ones, then impossibility results quickly accumulate showing that no pool can simultaneously satisfy all the requirements (French 1985).

It can be argued (see, e.g. Bedford and Cooke 2001) that independence preservation is less important than marginalisation, and hence that it makes most sense to adopt the linear opinion pool.

The equally weighted linear opinion pool is extensively used in applications though the justification for doing so is seldom discussed in any detail. Possible arguments are appeals to fairness between experts and the lack of any reason to deviate from equal weights. However, we have argued above that fairness/equity arguments should not be applied to expert problems, and Cooke's approach described below provides reasons to deviate from equal weights.

1.4.3 Cooke's Classical Model

Cooke developed the *Classical* Model to combining expert judgement (Cooke 1991; see also Part II of this book). In this, the weights are defined empirically on the basis of the experts' relative performance on a calibration set of variables. This is in accordance with the *Empirical Control* principle for rational consensus. This principle is the one that justifies the collection of calibration data so that the quality of each expert's input can be assessed and their judgements weighted accordingly. Here, the experts do not know the true values of the calibration quantities, but the analyst does. Comparing the experts' answers with the true values over the calibration set allows Cooke to form measures of the calibration and informativeness of each expert, and from these he constructs *performance-based* weights. Cooke justifies his weights on the basis of a particular group scoring rule, which follows the *Neutrality* principle.

Cooke's Classical Model has been used in over 150 published studies. Several case studies are reported later in this book: see also Dias et al. (2018) and the special issue of *Reliability Engineering and System Safety* (2008, **93**(5)). There is also a growing database with the data from these SEJ studies available online at http://rogermcooke.net/.

1.4.4 Bayesian Approaches

Bayesian approaches differ from opinion pools in that they treat the numerical judgements as data and seek to update a prior distribution supplied by or constructed for the decision-maker using Bayesian methods. This requires that the analyst develops appropriate likelihood functions to represent the information implicit in the experts' statements. Specifically, the likelihood function needs to model:

- the experts' ability to encode their uncertainty probabilistically (Clemen and Lichtendahl 2002; O'Hagan et al. 2006; Hora 2007; Lin and Bier 2008);
- correlations that arise because of experts' shared knowledge and common professional backgrounds (Shanteau 1995; Mumpower and Stewart 1996; Wilson 2016);
- correlations between the decision-maker's own judgements and the experts' (French 1980);
- the effects of other biases arising from conflicts of interests and the general context of the elicitation (Hockey et al. 2000; Skjong and Wentworth 2001; Lichtendahl and Winkler 2007; French et al. 2009; Kahneman 2011).

Constructing such a likelihood is a far from easy task. Indeed, developing suitable likelihood functions proved an insurmountable hurdle for many years and only recently have tractable methods with reasonable likelihood functions become available (Albert et al. 2012; French and Hartley 2018; see also Chap. 5).

This does not mean that Bayesian ideas have been without influence. They have provided an analytical tool for investigating the principles of combining and using expert judgement. For instance, French (1987) used a simple Bayesian model to argue against the principle of Independence Preservation and French and Hartley (2018) used a Bayesian approach to critique the European Food Safety Authority's SEJ methodology (EFSA 2014).

1.5 Behavioural Aggregation

Behavioural approaches work with the group of experts to agree on probabilities or quantiles that they can all accept as reasonable input to the risk or decision analysis, even if they themselves would still hold to different values. Simplistically, the analyst might gather the experts together and let them come to some agreement about the numbers to put into the models (DeWispelare et al. 1995). However, it is better to use a more structured approach. The Sheffield method is a version of facilitated workshop or decision conferencing (Reagan-Cirincione 1994; Phillips 2007) in which the group is helped to agree on probabilities that an 'impartial observer' of their discussions might give (O'Hagan et al. 2006; EFSA 2014). The long established Delphi method, or rather family of methods, does not allow the experts to meet but structures a discussion in which their share and revise their opinions through several iterations arriving at an agreed set of values (Dalkey and Helmer 1963; Rowe and Wright 1999; EFSA 2014).

Many factors need to be balanced in choosing between mathematical and behavioural aggregation (Clemen and Winkler 1999). Behavioural aggregation may be affected by many group dysfunctional behaviours, though these may be countered by good facilitation. It can help share knowledge and ensure that all experts share the same precise understanding of the quantities to be elicited. However, there is a risk that behavioural aggregation can win people over, losing concerns about some issues held by a small minority in building a group consensual judgement. Thus in risk management contexts, behavioural aggregation may lose sight of some potential hazards. In planning or regulatory decisions, the explicit, auditable nature of mathematical aggregation can be an advantage in recording all the reasoning implicit in the analysis.

However, there is no reason for the two approaches to be kept separate. Hanea et al. (2018) have developed the IDEA protocol which can combine group discussion with the use of the Classical Model to form the final judgements from the individual assessments.

1.6 Reporting

SEJ studies are necessarily important inputs to a risk or decision analysis. Since such studies are expensive, they are only undertaken when the events or quantities of concern are significant in driving the output uncertainties of the analysis. So it behooves those conducting the studies to report their conduct and conclusions fully. It is somewhat surprising therefore that there is remarkably little guidance on how this should be undertaken. In contrast, the research community and scientific journals have developed and enforced a wide range of principles to govern the peer review, publication and use of empirical studies, alongside which has grown a recognition of the importance of evidence-based decision-making (Pfeffer and Sutton 2006;

Shemilt et al. 2010). The latter developments began within medicine, particularly in the Cochrane Collaboration; but the imperatives of basing decisions on evidence are now changing thinking in many domains.

As mentioned previously, Cooke suggested the *Scrutability/Accountability* principle according to which all data, including experts' names and assessments, and all processing tools should be open to peer review and results must be reproducible. This, unfortunately, is sometimes unachievable as many experts are uncomfortable about having their assessments published under their names. They prefer and expect publication under Chatham House Rules, namely, their participation in the study will be noted, but their judgements and other input will be reported anonymously.

French (2012) considered issues relating to the reporting of SEJ studies from the perspective of potential future meta-analyses that might seek to draw information from two or more. One point here is that in scientific reporting of empirical studies, one should always report the experimental design process underpinning the data collection. In SEJ the elicitation protocol serves the same purpose, though it is not reported in detail in many studies.

EFSA (2014) guidance on running SEJ studies is perhaps the most thorough to date and that does provide a lot of advice on the content of report and the responsibilities of different teams for writing sections of reports. However, the community of analysts involved in SEJ still mainly relies on their experience and personal perspectives in deciding what to include and with what details in their reporting.

1.7 Directions for Future Developments and Research

SEJ methodologies are maturing. The past decade or so has seen a steady growth in applications across many domains. Several are reported in this and an earlier sister book (Dias et al. 2018); and any literature review will easily find many more. We have a toolbox of methods and tools to call on when an SEJ study is needed. But that is not to deny the need for further research. There are several areas in which more developments, both theoretical and methodological, are needed.

Firstly, our current SEJ methods focus mainly on eliciting and aggregating expert assessments of the probability of unknown events or univariate probability distributions of unknown quantities. However, once multivariate probability distributions are needed, things become more difficult and there is still a need for research. Multivariate distributions, whether described by a full multivariate distribution function or represented by a belief network, hierarchical model or some such, require dependences, independences, correlations, etc., if they are to be defined fully. How these should be elicited and aggregated remains a research topic. Quigley et al. (2013) and Werner et al. (2017) are very relevant references; as are Chaps. 2, 4 and 7 in this volume. However, much remains to be done.

We mentioned above that Bayesian methods had not been much used in practice, though they had provided many theoretical insights into SEJ. Things are changing. Chapter 5 in this volume suggests that a new Bayesian method of grouping experts automatically using a calibration set may be achieving a similar performance to the Classical Model. While there may be no great improvement over the simpler Classical approach, the Bayesian methodology will combine seamlessly with other Bayesian models for data analysis, machine learning, risk and decision analysis in an overall analysis. This suggests that further work to improve the Bayesian modelling of the correlations between experts and between them and the decision-maker may bring overall benefits.

Experts do not have unlimited time and effort to give to the elicitation of uncertainties. Apart from the value of their time in other aspects of the analysis and elsewhere, reflected in high consultancy rates, they tire because reflective elicitation requires considerable thought and effort. Currently, much simple sensitivity analysis is used in the early stages of the overall risk and decision analysis to prioritise the uncertainties which should be elicited from experts (French 2003). However, those processes are rudimentary and conducted before the SEJ elicitations. One aspect of elicitation that is not emphasised as much as it should be is that much *qualitative* knowledge is elicited from the experts alongside their quantitative judgements. This qualitative information can be used to shape the overall modelling further, and hence may change the priorities that a sensitivity analysis would determine. Indeed, their quantitative judgements on some uncertainties may constrain possible values of others in complex risk and decision models. Thus there may be benefit in developing procedures and tools to integrate the processes of SEJ with the overall risk and decision analysis process yet more effectively.

Taking this last point further, the overall risk and decision models need building in the first place. There are a host of—arguably under-researched—methods and tools to catalyse this process, which go under various names in different disciplines, e.g. problem structuring methods, soft-OR and knowledge engineering. We might step back and look at the whole process of formulating models, identifying parameters for which there is insufficient data to quantify their values and the uncertainties, and then focusing on those as targets for SEJ. Qualitative information should flow back and forth along this as the model is shaped. We might term this process as iterating between soft and hard elicitation, i.e., between identifying model structure and parameter values. Note that in simpler models a parameter may approximate the average effect of a sub-model in a more complex model. Thus, the distinction between parameters and model structure is to some extent arbitrary and depends on the modeller's perspective. This suggests looking at the modelling process as a whole, from the point of view of eliciting knowledge from experts. If we do that we should pause and think: why do we use a variety of techniques in hard quantitative elicitation to counter possible biases arising from System 1 Thinking, yet use virtually none in eliciting knowledge to structure models? There is a need for research to look at the whole modelling and elicitation process in relation to the potential effects of System 1 Thinking and to develop mechanisms and interventions that encourage comprehensive System 2 Thinking.

Finally, we now have around 30 years of experience in SEJ applications. That is plenty of time for many of the uncertainties relating to the risks and decisions to be resolved. It would be interesting in historical research to relate the actual outcomes to the probabilities derived from SEJ used in past analyses. In short, we could step back and seek to calibrate and validate SEJ overall. There would be many difficulties, of course. The world is far more complex than the microcosmos explored in risk and decision analyses, and the ultimate outcomes may be derived more from unanticipated events and behaviours outside the models. Nonetheless, there are potentially data that might be used to provide validation for SEJ as a whole, and that would do much to reassure the decision-makers and risk owners of today of the value of our methods.

1.8 Outline of the Book

This book grew from a conference on *The State of the Art in the Use of Expert Judgement in Risk and Decision Analyses* held in Delft in July 2017. The conference marked the end of a European Co-operation in Science and Technology (COST)[1] Action, which ran from 2013 to 2017. Its main aim had been to create a multidisciplinary network of scientists and policymakers using SEJ to quantify uncertainty for evidence-based decisions, and hence improve effectiveness in the use of science knowledge by policymakers. An earlier sister volume to this had been written at the outset of the Action and had recently been published (Dias et al. 2018).

The conference also had a second purpose: to honour and celebrate the work of Roger Cooke in establishing sound SEJ processes over some four decades. Many of us had worked with him over that period and, indeed, still work with him. Thus, this volume is also a festschrift for him and a recognition of his leadership.

We have divided the chapters into four parts:

Part I. Current Research.
Part II. Cooke and the Classical Model.
Part III. Process, Procedures and Education.
Part IV. Applications.

1.8.1 Part I: Current Research

Part I gathers recent theoretical developments in the field of SEJ. Chapter 2 focuses on expert elicitation of parameters of multinomial models. It presents an extensive overview of the recent research on the topic, along with guidelines for carrying out an elicitation and supporting examples. Chapter 3 discusses whether using performance weights is beneficial and advances the random expert hypothesis. Chapter 4 considers expert elicitation for specific graphical models and discusses the importance of choosing an appropriate graphical structure. Chapter 5 considers how to model dependencies between experts' assessments within a Bayesian framework

[1] See https://www.cost.eu/actions/IS1304/ and https://www.expertsinuncertainty.net/.

and provides a performance comparison with the Classical Model. Still within a Bayesian context, yet in a preventive maintenance setting, Chap. 6 focuses on eliciting experts' lifetime distributions in order to obtain prior parameters of a Dirichlet process. Finally, Chap. 7 proposes an adversarial risk analysis approach for SEJ studies in which the main uncertainties relate to the actions of other actors, usually though not necessarily adversaries.

1.8.2 Part II: Cooke and the Classical Model

Roger Cooke's oration in 1995 on taking up his chair at the Technical University of Delft has never been formally published. It was given barely 10 years after he developed the Classical Model and showed its potential value in early applications; we are proud to publish the oration here as Chap. 8. In Chap. 9, one of us reflects back to that early decade, showing how all the basic principles of the Classical Model and the processes surrounding were laid down then. Chapter 10 provides a current overview of the theory of the Classical Model, providing a deep and comprehensive perspective on its foundations and its application. In Chap. 11, we present an interview with Roger Cooke in which he reflects on the Classical Model and the processes of SEJ.

1.8.3 Part III: Process, Procedures and Education

Part III focuses on processes and procedures for SEJ and on how experiences should be turned into lessons and guidelines that continue to shape what structured elicitation protocols are in the digital age. In Chap. 12, we report an interview with Professor Dame Anne Glover, who served as the Chief Scientific Advisor to the President of the European Commission. She reflects on the role of expert scientific advice to governments. Chapter 13 synthesises the characteristics of good elicitations by reviewing those advocated and applied. It examines the need of standardisation in mature protocols. Chapter 14 discusses the design and development of a training course for SEJ, based on two major experiences in training postgraduates, early career researchers and consultants. Chapters 15 and 16 detail specific experiences with SEJ protocols with the intention of presenting the challenges and insights collected during this journey, and the way those re-shaped what an optimal protocol may look like.

1.8.4 Part IV: Applications

There have been many applications of SEJ over the years. The database maintained by Roger Cooke reports over 100 using the Classical Model alone. The Sheffield Method which is the leading behavioural aggregation approach has had many applications though there is not a specific database of these. In the COST Action, we discussed and promoted applications in many areas including natural seismic and volcanic risks, geographical and spatio-temporal uncertainties relating to pollution and disease, uncertainties in managing public health systems, food safety and food security, project and asset management risks, and the uncertainties that arise in innovation and development. Part IV begins with some reflections from Willy Aspinall on his many experiences in applying the Classical Model in several application domains (Chap. 17). Chapter 18 also provides some related reflections on imperfect elicitation. We present several discussions and applications relating to medicines policy and management (Chap. 19), supply chain cyber risk management (Chap. 20), geopolitical risks (Chap. 21), terrorism (Chap. 22) and the risks facing businesses looking to internationalise (Chap. 23).

References

Albert, I., Donnet, S., Guihenneuc-Jouyaux, C., Low-Choy, S., Mengersen, K., & Rousseau, J., et al. (2012). Combining expert opinions in prior elicitation. 503–532.

Banks, D. L., Aliaga, J. M. R., & Insua D. R. (2015). *Adversarial risk analysis*. CRC Press.

Bazerman, M. H. (2006). *Managerial decision making*. New York: John Wiley and Sons.

Bedford, T., & Cooke, R. M. (2001). *Probabilistic risk analysis: Foundations and methods*. Cambridge: Cambridge University Press.

Burgman, M. A. (2015). *Trusting judgements: How to get the best out of experts*. Cambridge University Press.

Clemen, R. T., & Lichtendahl, K. C. (2002). *Debiasing expert overconfidence: A Bayesian calibration model*. PSAM6, San Juan, Puerto Rico.

Clemen, R. T., & Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis, 19*(2), 187–203.

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

Cooke, R. M., ElSaadany, S., & Xinzheng, H. (2008). On the performance of social network and likelihood based expert weighting schemes. *Reliability Engineering and System Safety, 93*(5), 745–756.

Cooke, R. M., & Goossens, L. H. J. (2000). Procedures guide for structured expert judgement in accident consequence modelling. *Radiation Protection Dosimetry, 90*(3), 303–309.

Dalkey, N., & Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. *Management Science, 9*(3), 458–467.

DeWispelare, A. R., Herren, L. T., & Clemen, R. T. (1995). The use of probability elicitation in the high-level nuclear waste regulation program. *International Journal of Forecasting, 11,* 5–24.

Dias, L., Morton, A., & Quigley, J., (Eds.) (2018). *Elicitation of preferences and uncertainty: Processes and procedures*. Springer.

Edwards, W. (1954). The theory of decision making. *Psychological Bulletin, 51,* 380–417.

EFSA. (2014). Guidance on expert knowledge elicitation in food and feed safety risk assessment. *EFSA Journal, 12*(6), 3734–4012.

Evans, J. S. B., Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Pyschological Sciences*, *8*(3), 223–241.

Faria, A., & Smith, J. Q. (1997). Conditionally externally Bayesian pooling operators in chain graphs. *Annals of Statistics, 25*(4), 1740–1761.

French, S. (1980). Updating of belief in the light of someone else's opinion. *Journal of the Royal Statistical Society, A143,* 43–48.

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

French, S. (1987). Conflict of belief: when advisers disagree. In P. G. Bennett (Ed.), *Analysing conflict and its resolution: Some Mathematical Contributions* (pp. 93–111). Oxford: Oxford University Press.

French, S. (2003). Modelling, making inferences and making decisions: The roles of sensitivity analysis. *TOP, 11*(2), 229–252.

French, S. (2011). Aggregating expert judgement. *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales, 105*(1), 181–206.

French, S. (2012). Expert judgment, meta-analysis, and participatory risk analysis. *Decision Analysis, 9*(2), 119–127.

French, S., & Hartley, D. (2018). Elicitation and calibration: A Bayesian perspective. In L. Dias, A. Morton & J. Quigley (Eds.), *Elicitation: The science and art of structuring judgement.* New York: Springer (in press).

French, S., Maule, A. J., & Papamichail, K. N. (2009). *Decision behaviour, analysis and support*. Cambridge: Cambridge University Press.

French, S., & Rios Insua, D. (2000). *Statistical decision theory*. London: Arnold.

Genest, C., & Wagner, C. G. (1987). Further evidence against independence preservation in expert judgement synthesis. *Aequationes mathematicae, 32,* 74–86.

Genest, C., & Zidek, J. V. (1986). Combining probability distributions: A critique and annotated bibliography. *Statistical Science, 1,* 114–148.

Gigerenzer, G. (2002). *Reckoning with risk: Learning to live with uncertainty.* Harmondsworth: Penguin Books.

Hanea, A., McBride, M., Burgman, M., & Wintle, B. (2018). Classical meets modern in the IDEA protocol for structured expert judgement. *Journal of Risk Research, 21*(4), 417–433.

Hockey, G. R. J., Maule, A. J., Clough, P. J., & Bdzola, L. (2000). Effects of negative mood on risk in everyday decision making. *Cognition and Emotion, 14,* 823–856.

Hora, S. (2007). Eliciting probabilities from experts. In W. Edwards, R. F. Miles, & D. Von Winterfeldt (Eds.), *Advances in decision analysis: From foundations to applications* (pp. 129–153). Cambridge: Cambridge University Press.

Hora, S. C. (2008). Expert judgement. In E. L. Melnick & B. S. Everitt (Eds.), *Encylcopedia of quantitative risk analysis and assessment* (pp. 667–676). Chichester: John Wiley and Sons.

Jacobs, R. A. (1995). Methods for combining experts' probability assessments. *Neural Computing, 7,* 867–888.

Kahneman, D. (2011). *Thinking, Fast and Slow*. London, Penguin: Allen Lane.

Kahneman, D., & Tversky, A. (1974). Judgement under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131.

Kahneman, D., & Tversky, A. (Eds.). (2000). *Choices, values and frames.* Cambridge: Cambridge University Press.

Lichtendahl, K. C., & Winkler, R. L. (2007). Probability elicitation, scoring rules, and competition among forecasters. *Management Science, 53*(11), 1745–1755.

Lin, S.-W., & Bier, V. M. (2008). A study of expert overconfidence. *Reliability Engineering and System Safety, 93,* 711–721.

Madansky, A. (1964). Externally Bayesian groups. RM-4141-PR, RAND.

McConway, K. (1981). Marginalisation and linear opinion pools. *Journal of the American Statistical Association, 76,* 410–414.

Mumpower, J. L., & Stewart, T. R. (1996). Expert judgement and expert disagreement. *Thinking & Reasoning, 2*(2–3), 191–211.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, R., Garthwaite, P. H., Jenkinson, D., et al. (2006). *Uncertain judgements: Eliciting experts' probabilities.* Chichester: John Wiley and Sons.

Osborne, M. J. (2003). *An introduction to game theory.* Oxford: Oxford University Press.

Pfeffer, J., & Sutton, P. J. (2006). Evidence-based management. *Harvard Business Review, 84*(1), 62–76.

Phillips, L. D. (2007). Decision conferencing. In W. Edwards, R. F. Miles, & D. von Winterfeldt (Eds.), *Advances in decision analysis: From foundations to applications* (pp. 375–399). Cambridge: Cambridge University Press.

Quigley, J., Wilson, K. J., Walls, L., & Bedford, T. J. R. A. (2013). A Bayes linear Bayes method for estimation of correlated event rates. *33*(12), 2209–2224.

Reagan-Cirincione, P. (1994). Combining group facilitation, decision modelling and information technology to improve the accuracy of group judgement Improving the accuracy of group judgment: a process intervention combining group facilitation, social judgment analysis, and information technology. *Organisational Behaviour and Human Decision Process, 58,* 246–270.

Reliability Engineering and System Safety special Issue: Expert Judgment (2008). Edited by Roger M. Cooke, *93*(5), 655–778.

Rowe, G., & Wright, G. (1999). The Delphi technique as a forecasting tool: Issues and analysis. *International Journal of Forecasting, 15,* 353–375.

Sen, A. (2017). *Collective choice and social welfare, expanded edition.* Harmondsworth: Penguin.

Shanteau, J. (1995). Expert judgment and financial decision making. In B. Green (Ed.), *Risky business: Risk behavior and risk management*. Stockholm: Stockholm University

Shemilt, I., Mugford, M., Vale, L., Marsh, K., & Donaldson, C. (Eds.). (2010). *Evidence-Based decisions and economics: Helath care, social welfare, education and criminal justice*. Oxford University Press: Oxford.

Skjong, R. & Wentworth, B. H. (2001*).* Expert judgement and risk perception. In *Proceedings of the Eleventh (2001) International Offshore and Polar Engineering Conference.* Stavanger, Norway: The International Society of Offshore and Polar Engineers.

Werner, C., Bedford, T., Cooke, R. M., Hanea, A. M., & Morales-Nápoles, O. (2017). Expert judgement for dependence in probabilistic modelling: A systematic literature review and future research directions. *European Journal of Operational Research, 258*(3), 801–819.

Wilson, K. J. (2016). An investigation of dependence in expert judgement studies with multiple experts. *International Journal of Forecasting*.

Wright, G., & Ayton, P. (1987). Eliciting and modelling expert knowledge. *Decision Support Systems, 3*(1), 13–26.

Part I
Current Research

Although structured expert judgement is a maturing field, there remain many areas where further theoretical developments are needed. The chapters in this part report on progress in some of those.

Chapter 2
Recent Advances in the Elicitation of Uncertainty Distributions from Experts for Multinomial Probabilities

Kevin J. Wilson, Fadlalla G. Elfadaly, Paul H. Garthwaite, and Jeremy E. Oakley

Abstract In this chapter, we consider the problem of the elicitation and specification of an uncertainty distribution based on expert judgements, which may be a subjective prior distribution in a Bayesian analysis, for a set of probabilities which are constrained to sum to one. A typical context for this is as a prior distribution for the probabilities in a multinomial model. The Dirichlet distribution has long been advocated as a natural way to represent the uncertainty distribution over the probabilities in this context. The relatively small number of parameters allows for specification based on relatively few elicited quantities but at the expense of a very restrictive structure. We detail recent advances in elicitation for the Dirichlet distribution and recently proposed alternative approaches, which offer greater flexibility at the expense of added complexity. In order of increasing flexibility, they are the generalised Dirichlet distribution, multivariate copulas and vines. An extension of multinomial models containing covariates is discussed.

2.1 Introduction

The specification of an uncertainty distribution from the judgements of an expert for a set of probabilities which are constrained to sum to one, typically found as the parameters in a multinomial likelihood, is a common problem in many fields.

K. J. Wilson (✉)
School of Mathematics, Statistics & Physics, Newcastle University, Newcastle upon Tyne, UK
e-mail: kevin.wilson@ncl.ac.uk

F. G. Elfadaly · P. H. Garthwaite
School of Mathematics & Statistics, Open University, Milton Keynes, UK
e-mail: Fadlalla.Elfadaly@open.ac.uk

P. H. Garthwaite
e-mail: Paul.Garthwaite@open.ac.uk

J. E. Oakley
School of Mathematics & Statistics, University of Sheffield, Sheffield, UK
e-mail: j.oakley@sheffield.ac.uk

A. M. Hanea et al. (eds.), *Expert Judgement in Risk and Decision Analysis*,
International Series in Operations Research & Management Science 293,
https://doi.org/10.1007/978-3-030-46474-5_2

The resulting distribution is typically used as a sampling distribution in a Monte Carlo simulation or as a prior distribution in a Bayesian analysis. We will use the phrase “prior distribution” to mean either of these contexts or similar contexts.

In this chapter, we are concerned with modelling parameter dependence rather than the dependence structure of multivariate data. The most widely used approach is to specify a Dirichlet distribution for the probabilities, but, while there has been much work looking at elicitation for the Dirichlet distribution, there has been less emphasis placed on the assessment of whether the Dirichlet distribution is a suitable structure to represent an expert’s beliefs. With a small number of parameters, the dependence structure the Dirichlet distribution imposes is relatively restrictive. For example, if we were to specify the mean of each probability and the variance of just one of them, the Dirichlet structure would then impose all of the other variances and the covariances between the probabilities. Each covariance is constrained to be negative.

Some alternatives to the Dirichlet distribution which relax some of these restrictions have been proposed. The Connor–Mosimann distribution has nearly twice as many parameters as the Dirichlet distribution, allowing, for example, specification of means and variances of each probability, before imposing the covariances between them. There are restrictions on the values the covariances can take in the Connor–Mosimann distribution. For example, the first probability is restricted to be negatively correlated with all of the remaining probabilities.

The Gaussian copula and vine distributions (also known as pair copula constructions) are more flexible again, with enough parameters to allow the specification of means, variances and covariances of the probabilities. They also allow both positive and negative correlations between any pair of probabilities. Vines, in addition, offer a natural structure that avoids explicit consideration of ensuring that the covariance matrix is positive definite, which must be considered in the specification of a Gaussian copula.

However, with each increase in the number of parameters to be specified in the prior distribution, comes an increase in the time and effort that is needed to quantify expert opinion about the parameters. Note that means, variances and covariances for probabilities should not be specified directly, but should be specified indirectly by asking experts about observable quantities. Elicitation and specification of parameter dependence in the general context using copulas has been considered in the literature. For a review, see Sect. 4.1.1 of Werner et al. (2017). The added complexity induced by the condition that the probabilities must sum to one means that such general methods cannot be used directly here.

In this chapter, recent advances in the elicitation, implementation and diagnostic assessment of a Dirichlet distribution in this context within the Sheffield elicitation framework (SHELF) are reviewed in Sect. 2.2. Recently proposed alternatives to the Dirichlet distribution which offer greater flexibility are detailed in Sects. 2.3 and 2.4. In particular, Sect. 2.3 discusses the Connor–Mosimann distribution, the Gaussian copula, and vine copulas as prior distributions in this context, while Sect. 2.4 discusses the specification of a prior distribution for multinomial probabilities in the presence of covariates.

The emphasis throughout the chapter is on the practical implementation of the approaches. For general information on prior elicitation, see Garthwaite et al. (2005), O'Hagan et al. (2006).

2.2 Elicitation and Diagnostics for the Dirichlet Distribution

In this section, consideration is given to how to elicit a Dirichlet distribution, and what diagnostics one might use to assess whether a Dirichlet distribution adequately represents an expert's beliefs. Methods for eliciting Dirichlet distributions have been proposed in Bunn (1978), Dickey et al. (1983), Chaloner and Duncan (1987), van Dorp and Mazzuchi (2004), Evans et al. (2017) and Zapata-Vázquez et al. (2014). Here, we present Zapata-Vázquez et al.'s method (hereafter ZBO), which can be implemented in R (R Core Team 2018), using the package `SHELF` (Oakley 2017). Arguably, their method is the simplest to use from the expert's point of view, in that the expert is asked to make judgements about univariate quantities only. Two further approaches are given in Elfadaly and Garthwaite (2013) and Elfadaly and Garthwaite (2017), who also suggest more flexible prior distributions than the Dirichlet distribution.

The `SHELF` package includes various elicitation tools designed to support the "Sheffield Elicitation Framework" (Oakley and O'Hagan 2010): a behavioural aggregation method for eliciting a distribution from multiple experts. For more discussion and general advice about the practicalities of conducting an expert elicitation session, see also O'Hagan et al. (2006), EFSA (2014) and Gosling (2018). Here, we concentrate on the technical details of eliciting the Dirichlet distribution only.

2.2.1 Notation

Suppose there is some population where each member belongs to one of k categories. The uncertain quantities of interest are the population proportions in each category, which are denoted by the vector $\mathbf{p} := (p_1, \ldots, p_k)$. If uncertainty about $\mathbf{p}$ is to be described by a Dirichlet distribution, write $\mathbf{p} \sim \text{Dirichlet}(a_1, \ldots, a_k)$ with

$$f(\mathbf{p}) = \frac{\Gamma(a_1 + \ldots + a_k)}{\Gamma(a_1) \ldots \Gamma(a_k)} \prod_{i=1}^{k} p_i^{a_i - 1}, \tag{2.1}$$

where each $p_i \in [0, 1]$ and $\sum_{i=1}^{k} p_i = 1$, and so the aim of the elicitation method is to obtain values for $a_1, \ldots, a_k$, where each $a_i > 0$, based on suitable judgements from the expert.

2.2.2 Eliciting Marginal Distributions for Each Proportion

In the ZBO method, first elicit a marginal distribution for each p_i: we suppose $p_i \sim \text{beta}(d_i, e_i)$, and then choose values of d_i and e_i given appropriate judgements from the expert about p_i. This is a univariate elicitation problem, and any suitable elicitation method could be used (see Oakley 2010, for a review). One general approach is to elicit a small number of points from the expert's cumulative distribution function, asking the expert to provide either quantiles or probabilities, and then fitting a parametric cumulative distribution function to these points (typically using a suitable numerical approach). This approach can be implemented with the `SHELF` package.

ZBO choose to elicit quartiles for each p_i. The expert is asked to provide three values, denoted by $p_{i;\,0.25}$, $p_{i;\,0.5}$ and $p_{i;\,0.75}$, that satisfy

$$Pr(p_i \leq p_{i;\,0.25}) = 0.25, \quad Pr(p_i \leq p_{i;\,0.5}) = 0.5, \quad Pr(p_i \leq p_{i;\,0.75}) = 0.75. \tag{2.2}$$

Values d_i and e_i can then be chosen to minimise

$$\left(F(p_{i;\,0.25};\, d_i, e_i) - 0.25\right)^2 + \left(F(p_{i;\,0.5};\, d_i, e_i) - 0.5\right)^2 + \left(F(p_{i;\,0.75};\, d_i, e_i) - 0.75\right)^2, \tag{2.3}$$

where $F(.;\, d_i, e_i)$ is the cumulative distribution function of the beta(d_i, e_i) distribution, and the minimisation is done numerically.

Note that other quantiles could be elicited (e.g. 0.05th, 0.5th, 0.95th), but arguments in favour of using the quartiles are that there is less risk of overconfidence: tail probabilities can be difficult to judge, and the quartiles are more easily interpretable: the expert should judge that the four intervals $[0, p_{i;\,0.25}]$, $[p_{i;\,0.25}, p_{i;\,0.5}]$, $[p_{i;\,0.5}, p_{i;\,0.75}]$ and $[p_{i;\,0.75}, 1]$ are all equally likely to contain the true value of p_i.

2.2.2.1 Illustration and Implementation with SHELF

For illustration, suppose an election is to be held between three candidates, and it is wished to elicit an expert's beliefs about the proportion of the vote each candidate will get. Define p_i to be the proportion of the vote received by candidate i, for $i = 1, 2, 3$. The expert is asked to provide her quartiles for each p_i. Suppose the values she gives are those in Table 2.1.

Table 2.1 The expert's elicited quartiles for the proportion of votes each of three candidates will get in an election

Candidate	$p_{i;\,0.25}$	$p_{i;0.5}$	$p_{i;0.75}$
1	0.40	0.45	0.50
2	0.25	0.30	0.35
3	0.20	0.25	0.30

The parameters for the fitted beta distribution can be obtained with the `SHELF` package using the command `fitdist()`:

```
library(SHELF)
myfit1 <- fitdist(vals = c(0.4, 0.45, 0.5),
                  probs = c(0.25, 0.5, 0.75),
                  lower = 0, upper = 1)
myfit1$Beta
##   shape1 shape2
## 1  20.41  24.89
```

Hence, the fitted distribution is $p_1 \sim \text{beta}(20.41, 24.89)$. Similarly, $p_2 \sim \text{beta}(11.59, 26.71)$ and $p_3 \sim \text{beta}(8.64, 25.43)$. Within the `SHELF` package, there are various facilities for providing feedback to the expert: checking whether the fitted distribution is an acceptable representation of her beliefs. These involve showing plots of the fitted distribution to the expert, and reporting other probabilities and/or quantiles implied by the fitted distribution, but this is not shown here.

2.2.3 Obtaining the Dirichlet Distribution from the Marginal Distributions

If $\mathbf{p} \sim \text{Dirichlet}(a_1, \ldots, a_k)$, then $p_i \sim \text{beta}(a_i, a - a_i)$ for $i = 1, \ldots, k$, where $a = \sum_{i=1}^k a_i$. Therefore, in theory, eliciting the marginal distribution for each p_i would be sufficient to identify the values of $(a_1, \ldots, a_k)$: having first elicited $p_i \sim \text{beta}(d_i, e_i)$, set $a_i = d_i$ for $i = 1, \ldots, k$. In practice, an expert's set of elicited marginal distributions for $p_1, \ldots, p_k$ would almost certainly *not* be consistent with any Dirichlet distribution. Obtaining a marginal beta distribution from a Dirichlet, it can be seen that the parameters a_i and $a - a_i$ in the beta distribution sum to a for all i. Hence, eliciting $p_i \sim \text{beta}(d_i, e_i)$ directly would require $d_i + e_i$ to be the same for all i, to be consistent with a Dirichlet. In the election example, this was not the case: the three sums are 45.3, 38.3 and 34.1.

Another problem is that, in general, it is unlikely that the set of elicited marginal distributions would be coherent with each other. The constraint $\sum_{i=1}^k p_i = 1$ implies that $\sum_{i=1}^k E(p_i) = 1$; it would be hard for an expert to consider means directly such that they sum to 1: each p_i is constrained to lie in [0,1], so there may be skewness in the expert's marginal distributions. Again, this problem has occurred in the example: $\sum_{i=1}^3 E(p_i) = 1.007$.

There are two possibilities here. One is that the Dirichlet distribution is simply not a suitable choice for the expert's beliefs; a more flexible distribution is needed. Alternatively, there may be a Dirichlet distribution that *does* adequately represent the expert's beliefs, with implied marginal distributions that, although different, are acceptably close to what the expert first stated. This is quite possible given that there is likely to be some imprecision and/or "rounding error" in the expert's elicited quantiles.

ZBO suggest proposing a suitable Dirichlet as follows. Defining

$$r := \sum_{i=1}^{k} \frac{d_i}{d_i + e_i}, \tag{2.4}$$

they define

$$d_i^* := \frac{d_i}{r}, \quad e_i^* := d_i + e_i - d_i^*, \tag{2.5}$$

and impose the condition for the elicited Dirichlet that $E(p_i) = d_i^*/(d_i^* + e_i^*)$ for $i = 1, \dots, k$. This ensures that the prior expectations sum to 1. For use later on, define $n_i = d_i^* + e_i^*$ and

$$v_i^* = \frac{d_i^*(n_i - d_i^*)}{n_i^2(n_i + 1)}, \tag{2.6}$$

which is the variance of p_i, if $p_i \sim \text{beta}(d_i^*, e_i^*)$.

Now for $\mathbf{p} \sim \text{Dirichlet}(a_1, \dots, a_k)$, ZBO propose setting

$$a_i = n \frac{d_i^*}{d_i^* + e_i^*}, \tag{2.7}$$

where n is a further parameter to be selected. Increasing the value of n will decrease the prior variance of each p_i given in (2.6). ZBO suggest one of three choices for n:

1. an "optimal" value, which they define as

$$n_{opt} := \arg\min_n \left[\sum_{i=1}^{k} \left(\sqrt{Var(p_i)} - \sqrt{v_i^*} \right)^2 \right] \tag{2.8}$$

$$= \left(\frac{\sum_{i=1}^{k} v_i^*(n_i + 1)}{\sum_{i=1}^{k} v_i^* \sqrt{n_i + 1}} \right)^2 - 1. \tag{2.9}$$

   The criterion they consider here is to make the marginal standard deviations from the chosen Dirichlet to be as close as possible to those directly elicited, following the "adjustment" from beta(d_i, e_i) to beta(d_i^*, e_i^*);
2. a "compromise" value, such as the mean or median of $n_1, \dots, n_k$;
3. a "conservative" value: the minimum of $n_1, \dots, n_k$.

Note that $n_1, \dots, n_k$ relate to the expert's elicited prior uncertainty about $p_1, \dots, p_k$, respectively, with a decreasing value corresponding to an increasing prior variance, hence, the notion of "compromise" and "conservative" values.

This can be implemented in SHELF with the command fitDirichlet() (having first repeated the process in Sect. 2.2.2.1 for p_2 and p_3 to create the objects myfit2 and myfit3, respectively):

```
fitDirichlet(myfit1, myfit2, myfit3,
             categories = c("candidate 1",
             "candidate 2","candidate 3"),
             n.fitted = "opt")
##
## Directly elicited beta marginal distributions:
##
##        candidate 1 candidate 2 candidate 3
## shape1     20.4000     11.6000      8.6400
## shape2     24.9000     26.7000     25.4000
## mean        0.4510      0.3030      0.2540
## sd          0.0731      0.0733      0.0735
## sum        45.3000     38.3000     34.1000
##
## Sum of elicited marginal means: 1.007
##
## Beta marginal distributions from Dirichlet fit:
##
##        candidate 1 candidate 2 candidate 3
## shape1     17.6000     11.8000      9.9100
## shape2     21.7000     27.5000     29.4000
## mean        0.4480      0.3010      0.2520
## sd          0.0783      0.0722      0.0684
## sum        39.3000     39.3000     39.3000
```

The first table provides the results from fitting a beta distribution directly to each set of judgements about p_1, p_2, p_3. The second table gives the beta marginal distributions that result from the final chosen Dirichlet (where we have used n_{opt} for the parameter n). There is a small change in the prior expectations (unlikely to be of any importance to an expert in this context), and a slight reduction in the prior standard deviation for p_3, as a consequence of imposing the Dirichlet distribution.

2.2.4 Diagnostics

One diagnostic is to compare the directly elicited marginal distributions with those obtained from the fitted Dirichlet. This can be done by inspecting the tables in the R output but can also be presented graphically. Although not shown here, the discrepancy for each p_i is fairly small and might be expected as the elicited interquartile ranges were the same in each case. To give an example where the Dirichlet does not fit well, consider increasing $p_{3;\,0.75}$ from 0.30 to 0.40: the expert has more uncertainty about how well candidate 3 might do in the election. Following the same fitting procedure, the two sets of marginal distributions (directly elicited and those implied by the Dirichlet) are compared in Fig. 2.1. This discrepancy may not be acceptable to an expert, assuming she does not revise her initial judgements.

2.2.4.1 Investigating the Conditional Distributions

In the Dirichlet distribution, $k - 1$ parameters are required to describe the proportion in each category, leaving only one parameter to describe their dependence structure.

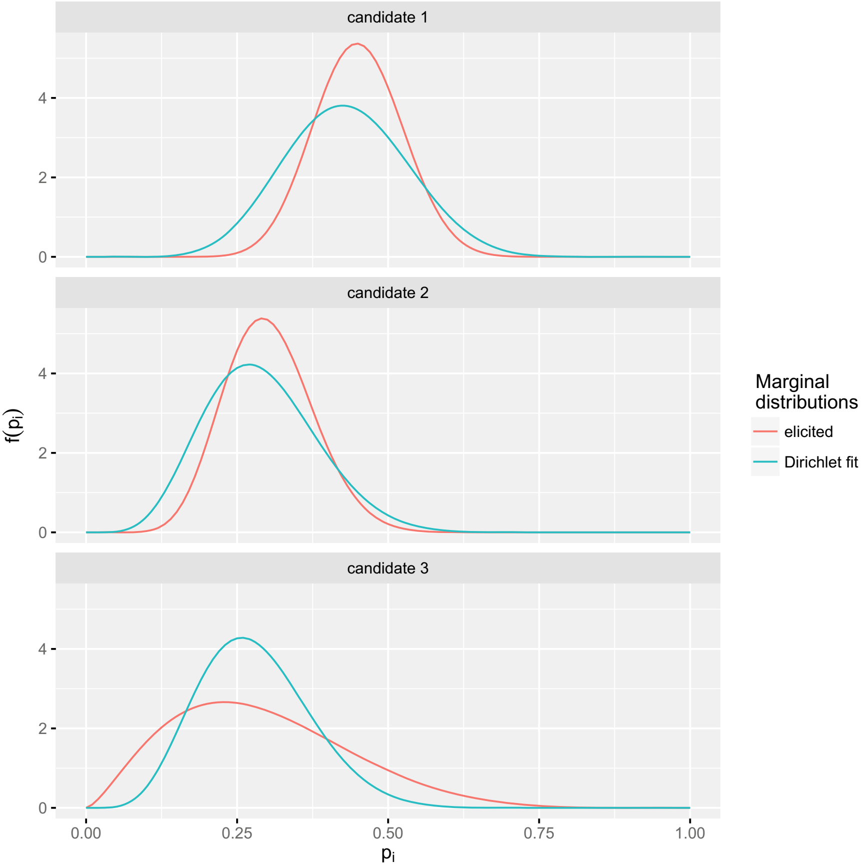


Fig. 2.1 Elicited marginal distributions, and the corresponding marginals implied by the fitted Dirichlet distribution. In this case, there is quite a large discrepancy, suggesting the Dirichlet family is not suitable for representing the expert's beliefs; it cannot adapt well to the greater uncertainty expressed about candidate 3's proportion of the vote

This will sometimes be insufficient to give a reasonable representation of the expert's opinions. The other well-known restriction of the Dirichlet prior is that the correlation between every pair of proportions must be negative, which again might not reflect an expert's opinions.

This suggests a second diagnostic procedure, where the conditional distribution of p_i, given p_j, is examined (visually) taking some hypothetical value x. The conditional distribution will be beta, scaled to the interval $[0, 1-x]$:

$$\left.\frac{p_i}{1-x}\right| p_j = x \sim \text{beta}\left(a_i, \sum_{s \neq i,j} a_s\right). \tag{2.10}$$

To illustrate this, assume the original judgements from Table 2.1. Suppose the experts were to learn that $p_2 = 0.05$: candidate 2 did considerably worse than expected. The conditional distributions (together with the fitted marginals) are plotted in Fig. 2.2. The distributions for p_1 and p_3 are both shifted upwards, with a stronger effect for p_1. This may or may not suitably represent the expert's beliefs. One can imagine a scenario where she believes that, should candidate 2 perform badly, this is likely to be a consequence of losing votes primarily to candidate 1; she might not revise her judgements about candidate 3 substantially. This sort of dependency structure is hard to model with the Dirichlet, and alternatives in the following sections may be more suitable.

In the `SHELF` package, an interactive tool is available using the command `condDirichlet()`, where the user can vary both the choice of j and x in conditioning on $p_j = x$.

2.3 Increasing the Flexibility of the Prior Distribution

In this section, three recently proposed elicitation methods for quantifying opinion about a multinomial model in a form that is more flexible than a Dirichlet distribution are described. The first method constructs a Connor–Mosimann distribution to represent expert opinion (Elfadaly and Garthwaite 2013). The Connor–Mosimann distribution has far more parameters than the Dirichlet distribution, and hence can capture a broader range of expert opinion. Like the Dirichlet distribution, it is conjugate to the multinomial distribution, and hence easily combined with data to form a posterior distribution. One regrettable constraint is that it forces all the probabilities to be negatively correlated with the probability of the first category.

The second method was proposed by Elfadaly and Garthwaite (2017). It constructs a more flexible prior than the Connor–Mosimann distribution through using a multivariate copula function: a copula represents a joint multivariate distribution in terms of one-dimensional marginal distributions. Elfadaly and Garthwaite (2017) use beta marginal distributions and a multivariate normal distribution function (a Gaussian copula) that binds these marginals and expresses the dependence structure between

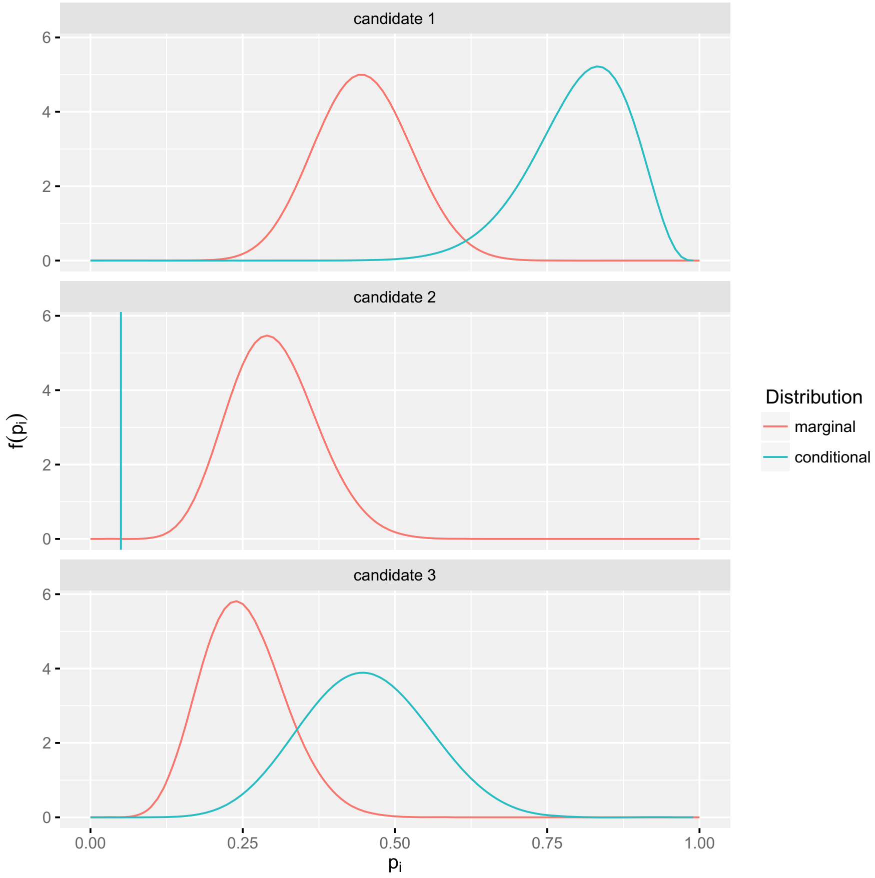


Fig. 2.2 The marginal distributions for each p_i, together with the conditional distributions of p_1 and p_3 given $p_2 = 0.05$. Here, the expert would reflect on whether she would revise her judgements in this way about p_1 and p_3, given a lower proportion of the vote than expected for candidate 2

them. The expert gives assessments that are used to determine the parameters of both the marginal beta distributions and the copula.

The Gaussian copula offers a prior structure which allows for the specification of the marginal distributions and dependencies between the parameters separately. Another prior structure which has this property is a vine (Bedford and Cooke 2002; Bedford et al. 2016). An approach to specify a prior distribution using vines in the case of multinomial probabilities is given in Wilson (2018). A vine is a graph made up of nodes and edges arranged in trees. This allows the specification of a vine distribution, in which each node in the first tree of the vine represents an unconditional variable and in each subsequent tree each nodes, which are made up of the edges of the tree above, represent conditional variables. The edges represent dependency between the adjacent nodes in the form of unconditional and conditional bivariate copulas. Thus, to specify a multivariate distribution using a vine, first specify marginal distributions for the variables and then a series of bivariate copulas representing unconditional and conditional dependence. A special case of a vine, which can be visualised, is known as a D-vine (or drawable-vine).

2.3.1 Assessing a Connor–Mosimann Prior Distribution

The density function of the Connor–Mosimann distribution has the following form (Connor and Mosimann 1969):

$$f(p_1, p_2, \ldots, p_k) = \prod_{i=1}^{k-1} \left[\frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)} \, p_i^{a_i - 1} \left(\sum_{j=i}^{k} p_j \right)^{b_{i-1} - (a_i + b_i)} \right] p_k^{b_{k-1} - 1}, \tag{2.11}$$

where $\sum p_i = 1$, $0 \le p_i \le 1$, $a_i > 0$ and $b_i > 0$ for $i = 1, \ldots, k-1$, and b_0 is arbitrary. The distribution can also be expressed in terms of $(k-1)$ independent beta variates $\theta_1, \theta_2, \ldots, \theta_{k-1}$, where $\theta_i \sim \text{beta}(a_i, b_i)$, for $i = 1, \ldots, k-1$, as follows:

$$p_1 = \theta_1, \;\; p_j = \theta_j \prod_{i=1}^{j-1} (1 - \theta_i), \;\; \text{for } j = 2, \ldots, k. \tag{2.12}$$

where, by definition, $\theta_k \equiv 1$.

The Connor–Mosimann distribution has $2(k-1)$ parameters, almost twice as many as the Dirichlet, giving it greater flexibility. There are still some constraints on the correlations between prior probabilities but these are less restrictive. It follows from (2.12) that the first prior probability, p_1, is always negatively correlated with the other prior probabilities, but any two successive probabilities p_j, p_{j+1} can be positively correlated for $j \neq 1$. The standard Dirichlet distribution is a special case of the Connor–Mosimann distribution in which $b_i = a_{i+1} + b_{i+1}$, for $i = 1, \ldots, k-2$. For this reason, the Connor–Mosimann is also called the *generalised Dirichlet distribution* and, like the Dirichlet distribution, it is a conjugate prior distribution for multinomial sampling.

Conditional assessments are the primary task in the elicitation method of Elfadaly and Garthwaite (2013) for quantifying an expert's opinion as a Connor–Mosimann distribution. Putting $p_r^* = p_r/(1 - \sum_{i=1}^{r-1} p_i)$ for $r = 2, \ldots, k-1$, (2.12) gives that the distribution of $p_r^* | p_1, \ldots, p_{r-1}$ is

$$p_r^* | p_1, \ldots, p_{r-1} \sim \text{beta}(a_r, b_r) \quad \text{for } r = 2, \ldots, k-1. \tag{2.13}$$

The following are the assessment tasks that the expert performs in the initial elicitation process (before receiving feedback).

(i) She assesses three quartiles for p_1, the probability of the first category. The quartile assessments are converted into estimates of parameters of a beta distribution, using the method given in Sect. 2.2.2. These parameter estimates are a_1 and b_1, as the prior distribution of p_1 is beta(a_1, b_1).

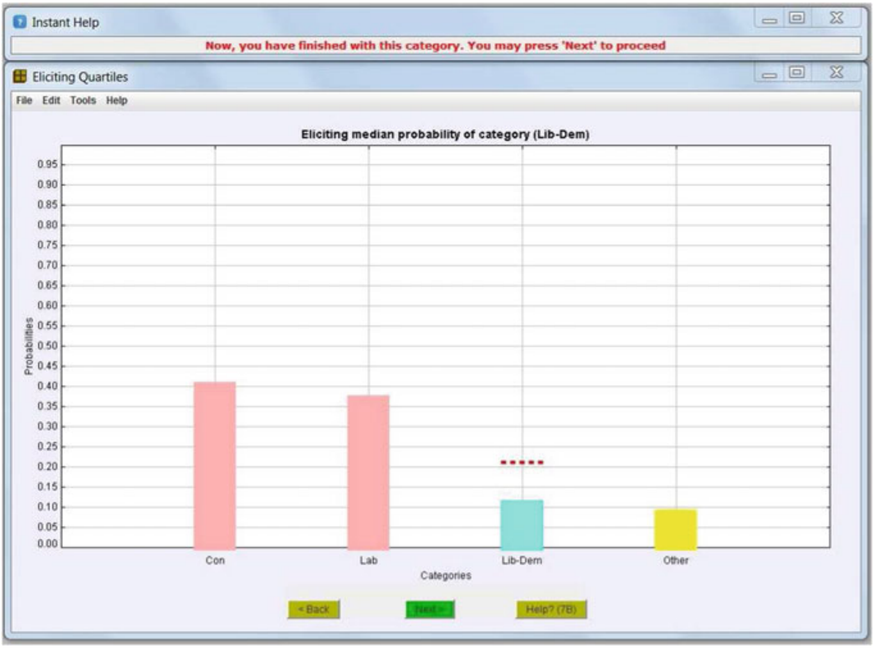
(ii) The expert is asked to assume that the median value she gave in the first step is the correct value of p_1, and she then assesses three quartiles for $p_2|p_1$. Dividing each of these quartiles by $1 - p_1$ gives the quartiles of $p_2^*|p_1$. These yield estimates of a_2 and b_2, the parameters of the beta distribution of $p_2^*|p_1$ (c.f. (2.13)).

(iii) The same process is repeated for each category except for the last one. For $r = 3, \ldots, k - 1$, the expert gives quartiles for $p_r|p_1, \ldots, p_{r-1}$. Dividing these by $1 - \sum_{i=1}^{r-1} p_i$ gives the three quartiles of $p_r^*|p_1, \ldots, p_{r-1}$, which are used to estimate a_r and b_r, the parameters of the beta distribution of $p_r^*|p_1, \ldots, p_{r-1}$.

These tasks yield estimates of the parameters a_i and b_i $(i = 1, \ldots, k - 1)$, and their specification fully determines the Connor–Mosimann prior distribution in (2.11).

Elfadaly and Garthwaite (2013) have implemented the elicitation method in freely available software that uses interactive computer graphics to question the expert. The software may be downloaded from http://statistics.open.ac.uk/elicitation. Using the software, the expert forms bar charts by clicking the computer mouse on vertical lines in figures that the computer displays, and hence she expresses her opinions about prior probabilities and conditional probabilities for different categories. Assessment of medians is illustrated in Fig. 2.3. (The context is voting behaviour in a local election.) The figure is a screenshot after the expert had assessed her median for the third category conditional on the true values for the first two categories being 0.41 and 0.38 (0.41 and 0.38 were the expert's median assessments for the first two categories). Figure 2.4 shows the conditional quartiles that the expert assessed for the third category (shown as short continuous horizontal (blue) lines), again conditional on the true values for the first two categories being 0.41 and 0.38. The dotted orange lines are suggested bounds for the conditional quartiles: if the conditional quartiles lie within these bounds (as in the figure) then the assessed conditional distribution is unimodal. This conditional distribution is displayed to the expert via a small graph, as illustrated in Fig. 2.4. Sometimes, the expert may find that the displayed shape of the conditional distribution does not form an acceptable representation of her opinions and opt to revise her quartile assessments by using the computer mouse to change their positions on the large graph.

Feedback is highly desirable in an elicitation method, especially if it helps the expert examine features of her assessed prior distribution from a fresh perspective. In the elicitation procedure, the expert has assessed quartiles that relate to *unconditional* probabilities for just the first category, so it is helpful to inform the expert of the marginal unconditional probability quartiles for all the categories. Marginal distributions of the Connor–Mosimann distribution are not directly of the beta type, but they can be adequately approximated as beta distributions (Fan 1991). The elicitation method uses a beta approximation introduced by Fan (1991) to calculate approximate quartiles of the marginal unconditional distribution for each category, and these are



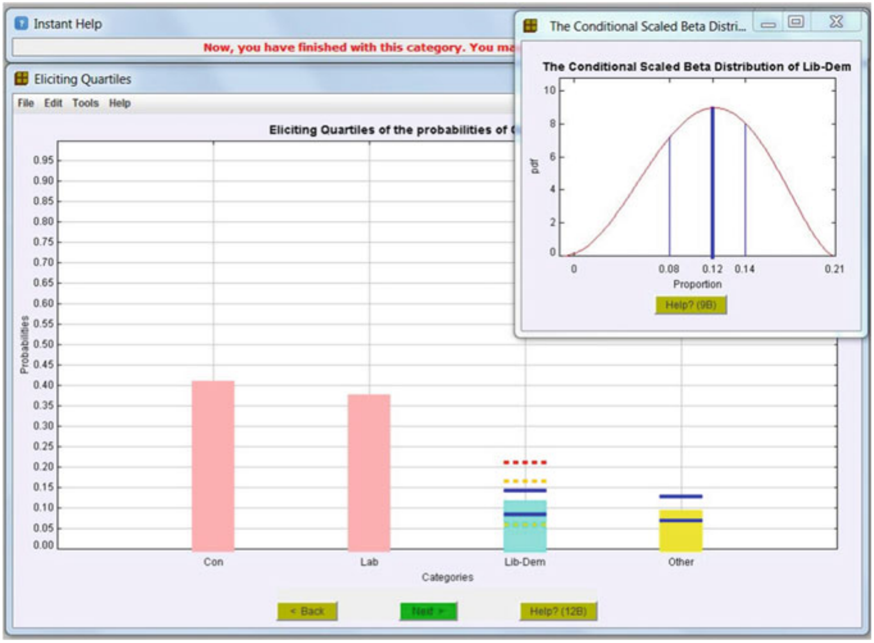
Now, you have finished with this category. You may press 'Next' to proceed

Fig. 2.3 Assessing conditional medians. The expert must assume that the true values for the first two categories are 0.41 and 0.38, as given by the two (pink) columns to the left. Conditional on this assumption, the expert assesses the conditional median of the third category, shown as the third (blue) column. An upper limit for this conditional median is given by the short dotted line. The median for the fourth category, shown as the (yellow) rightmost bar, is automatically computed and displayed once the median for the third category has been assessed

displayed on a bar chart (see Elfadaly and Garthwaite 2013 for further details). The expert is then given the opportunity to modify them and, if any changes are made, the elicitation method changes the parameter values in the prior distribution to reflect the modifications. The median and quartiles of the new marginal distribution of p_r ($r = 1, \ldots, k$) are displayed as fresh feedback, and the expert is again invited to accept or revise their values. The whole process can be repeated until the expert is satisfied with the feedback.

The assessments used to quantify opinion as a Connor–Mosimann distribution could also be used to fit a Dirichlet distribution (Elfadaly and Garthwaite 2013). In any given application, an obvious question is whether the additional flexibility of the Connor–Mosimann prior gives practical benefit or whether it would be adequate to use the simpler Dirichlet prior. As the Connor–Mosimann distribution reduces to the Dirichlet distribution when

$$b_i = a_{i+1} + b_{i+1} \quad \text{for } i = 1, \ldots, k-2, \tag{2.14}$$



Now, you have finished with this category. You ma

Fig. 2.4 Assessing conditional quartiles. The expert must assume that the true values for the first two categories are 0.41 and 0.38, as given by the two (pink) columns to the left. Conditional on this assumption, the expert assesses the two quartiles of the third category, shown as two (blue) short horizontal lines. Boundaries for these assessments are suggested by the short dotted lines. The assessed conditional probability density function (PDF) curve for the third category is shown in the small graph. The two quartiles for the fourth category, shown as (blue) short horizontal lines, are automatically computed and displayed once the quartiles for the third category have been assessed

a good diagnostic is to examine whether (2.14) is approximately satisfied in the Connor–Mosimann prior, simplifying to the Dirichlet distribution when it approximately holds.

2.3.2 Assessing a Gaussian Copula Prior Distribution

A copula is a multivariate function that represents a multivariate cumulative distribution function (CDF) in terms of one-dimensional marginal CDFs. Hence, it joins marginal distributions into a multivariate distribution that has those marginals. A copula has good flexibility because the marginal distributions can be chosen independently from the dependence structure of the copula function. For an introduction to copulas see, for example, Joe (1997), Frees and Valdez (1998) or Nelsen (1999).

The simplest and most intuitive family of copulas is the inversion copula, which has the form

$$C[G_1(x_1),\ldots,G_m(x_m)] = H_{(1,\ldots,m)}\left\{H_1^{-1}[G_1(x_1)],\ldots,H_m^{-1}[G_m(x_m)]\right\}. \quad (2.15)$$

In this equation, C is the copula function, G_i are the marginal CDFs of $X_1,\ldots,X_m$, and $H_{(1,\ldots,m)}$ is a multivariate CDF whose marginal distributions are $H_1,\ldots,H_m$. Hence, C uses $H_{(1,\ldots,m)}$ to couple the marginal functions $G_1,\ldots,G_m$ into a new multivariate distribution. The best-known example of the inversion copula in (2.15) is the Gaussian copula (Clemen and Reilly 1999). This is obtained from (2.15) by (i) choosing $H_{(1,\ldots,m)}$ as the CDF of an m-variate normal distribution with zero mean vector, unit variances and a correlation matrix $\mathbf{R}$ that reflects the desired dependence structure, and (ii) letting each H_i be the standard univariate normal CDF. For these choices, we denote $H_{(1,\ldots,m)}$ as $\Phi_{m,\mathbf{R}}$ and each H_i as ϕ. Jouini and Clemen (1996) and Clemen and Reilly (1999) consider exploiting copula functions to elicit multivariate distributions. In general, the joint distribution can be elicited by first assessing each marginal distribution, and then the dependence structure is elicited through the copula function.

The prior distribution should obey the laws of probability and so, in particular, the multinomial probabilities should satisfy the unit sum constraint, $p_1 + \cdots + p_k = 1$. To ensure that this constraint holds, Elfadaly and Garthwaite (2017) adopt the parameterisation used for the Connor–Mosimann distribution and parameterise the copula in terms of $\theta_1,\ldots,\theta_k$ as specified in (2.12). That is,

$$\theta_1 = p_1,\ \theta_i = \frac{p_i}{1-\sum_{j=1}^{i-1} p_j},\ \text{for } i = 2,\ldots,k-1,\ \text{and } \theta_k = 1. \quad (2.16)$$

The prior distribution for $p_1,\ldots,p_k$ is defined in terms of the prior distribution of $\theta_1,\ldots,\theta_{k-1}$. It is assumed that each θ_i $(i = 1,\ldots,k-1)$ has a marginal beta distribution with parameters a_i and b_i. The copula function is then $\Phi_{k-1,\mathbf{R}}\{\phi^{-1}[G_1(\theta_1)],\ldots,\phi^{-1}[G_{k-1}(\theta_{k-1})]\}$, where $G_i(\,.\,)$ is the CDF of a beta(a_i, b_i) distributed random variable, θ_i, for $i = 1,\ldots,k-1$. Differentiating this function with respect to $\theta_1,\ldots,\theta_{k-1}$ gives the probability density function of $\theta_1,\ldots,\theta_{k-1}$, which is the following multivariate density (Clemen and Reilly 1999):

$$f(\theta_1,\ldots,\theta_{k-1}|\mathbf{R}) = \frac{\prod_{j=1}^{k-1} g_j(\theta_j)}{|\mathbf{R}|^{1/2}}\ \exp\left\{-\frac{1}{2}\mathbf{w}'(\mathbf{R}^{-1} - \mathbf{I}_{k-1})\mathbf{w}\right\}, \quad (2.17)$$

where $\mathbf{w}' = (\phi^{-1}[G_1(\theta_1)],\ldots,\phi^{-1}[G_{k-1}(\theta_{k-1})])$, and $g_i(.)$ is the beta density of θ_i $(i = 1,\ldots,k-1)$. The correlation matrix $\mathbf{R}$ and the parameters of the beta densities are the quantities that must be encoded from an expert's assessments.

The elicitation method is implemented in software that uses interactive graphics to elicit an expert's prior distribution. As with the elicitation method for the Connor–Mosimann prior distribution, the expert forms bar charts to quantify her opinions.

To determine the parameters of the marginal beta distributions, the expert performs the following tasks.

(i) She assesses the median of the probability of each category. The median for the first category is the marginal unconditional median of θ_1.
(ii) For category i ($i = 2, \ldots, k - 1$), the expert is asked to assume that an observation does not fall in one of the first $i - 1$ categories, but which of the remaining categories it falls in is unknown. Given this situation, the expert is asked to assess the median of the probability of each of the categories $i, i + 1, \ldots, k$. Only the assessment for the first of these categories is used; it is the marginal unconditional median of θ_i. The other assessments are elicited to improve internal consistency; assessments of means should sum to one so the sum of the assessments of the medians should be close to one.
(iii) The expert assesses the lower and upper quartiles for the first category, which are equated to the marginal unconditional lower and upper quartile of θ_1. Then the set of assumptions that the expert made when assessing medians is repeated. Under the assumption that an observation does not fall in one of the first $i - 1$ categories, the expert assesses the lower and upper quartiles for the i^{th} category ($i = 2, \ldots, k - 1$). These assessments are the marginal unconditional lower and upper quartiles of θ_i.

The bar charts through which the expert makes her assessments are similar to those in Figs. 2.3 and 2.4. As in the elicitation method for the Connor–Mosimann distribution, the method of Sect. 2.2.2 is used to estimate the parameters of the beta(a_i, b_i) distribution from the assessed median and quartiles of θ_i.

Assessing the covariance matrix $\mathbf{R}$ is trickier. Clemen and Reilly (1999) consider the task of assessing correlations for copulas and propose three methods of eliciting $\mathbf{R}$, but none of the methods can guarantee that $\mathbf{R}$ will be a positive definite matrix, which is a requirement of the multivariate normal distribution. Instead, Elfadaly and Garthwaite (2017) modify an elicitation method of Kadane et al. (1980) to form a method of assessing $\mathbf{R}$ that ensures positive definiteness. The following are the assessment tasks that the expert performs.

(iv) The expert is asked to give conditional quartile assessments for specified categories under the assumption that her earlier median assessments are the correct values for other categories. First, the expert is asked to assume that p_1 equals the initial assessment of its median and assesses a lower quartile L_2 and an upper quartile U_2 for p_2. Then, for each remaining p_i, $i = 3, \ldots, k - 1$, the expert assesses the two quartiles L_i and U_i conditional on $p_1, \ldots, p_{i-1}$ each equalling its assessed median.
(v) The last task that the expert performs is to assess conditional medians. Most of the conditional values are equal to assessments of medians that the expert has given, but for one category the conditional value does not match an assessed median, so it should drive the expert to alter her opinion. Figure 2.5 is an example of the bar chart used for the assessment of conditional medians. The grey bars show the expert's initial median assessments but the expert is asked to assume

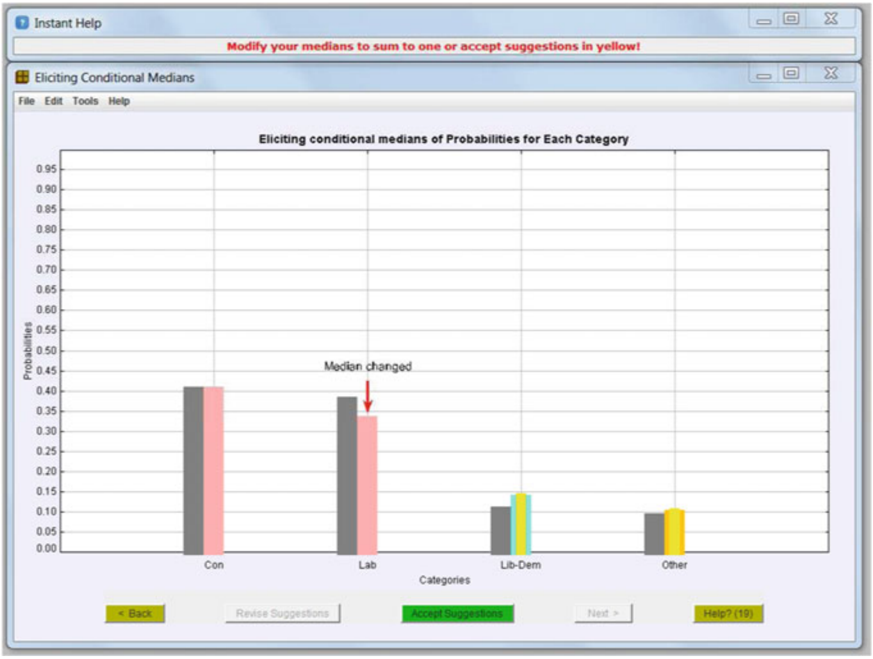

Modify your medians to sum to one or accept suggestions in yellow!

Fig. 2.5 Software suggestions for conditional medians

> that the pink bars are the correct values for the proportions in the leftmost two categories. For the second of these categories, the suggested value does not equal the expert's original median (the pink and grey bars are of different heights) so the expert needs to change her median assessments for the remaining categories. The blue and orange boxes are the expert's new assessed medians—they did not satisfy the unit sum constraint exactly, and the yellow bars are values suggested by the computer that meet this constraint and are close to the expert's new assessments. The expert is invited to accept the computer's suggestions as a reasonable representation of her opinions, or to revise them.

To obtain **R**, assessments about the p_i must be translated into information about $\mathbf{w} = (w_1, \ldots, w_{k-1})'$, where $w_i = \phi^{-1}[G_i(\theta_i)]$. This is straightforward for the required assessments, which are medians and quartiles of p_i *that were conditional assessments for specified values of* $p_1, \ldots, p_{i-1}$. Let m_j^* denote the value specified for p_j $(j = 1, \ldots, i-1)$ and suppose that L_i, M_i and U_i are the conditional lower quartile, median and upper quartile of $p_i | p_1 = m_1^*, \ldots, p_{i-1} = m_{i-1}^*$. Then, from (2.16),

$$L_i^{\#} = \frac{L_i}{1 - \sum_{j=1}^{i-1} m_j^*}, \quad M_i^{\#} = \frac{M_i}{1 - \sum_{j=1}^{i-1} m_j^*} \quad \text{and} \quad U_i^{\#} = \frac{U_i}{1 - \sum_{j=1}^{i-1} m_j^*} \tag{2.18}$$

are the corresponding conditional lower quartile, median and upper quartile of $\theta_i|\theta_1, \ldots, \theta_{i-1}$. Also, once the parameters of the beta distribution $g_i(.)$ have been chosen, the transformation from θ_i to w_i is a known, monotonically increasing transformation, as $w_i = \phi^{-1}[G_i(\theta_i)]$. Consequently, quantiles of θ_i yield the corresponding quantiles of w_i. From conditional medians and quartiles of the w_i, the covariance matrix of $\mathbf{w}$ is obtained using the method of Kadane et al. (1980), and this yields the correlation matrix $\mathbf{R}$.

Full details of the method are given in Elfadaly and Garthwaite (2017), and software implementing it may be downloaded from http://statistics.open.ac.uk/elicitation. The software includes substantial *help notes* to facilitate use of the elicitation method.

2.3.3 Assessing a Vine Prior Distribution

2.3.3.1 Details of the Approach

As with the Gaussian copula, it is possible to specify the prior distribution for $\theta_1, \ldots, \theta_{k-1}$, given in (2.16), which defines the prior distribution for $p_1, \ldots, p_k$. The parameters $\theta_1, \ldots, \theta_{k-1}$ can be interpreted as the conditional probability that an observation falls into a particular category given that it did not fall into any of the previous categories. In the Connor–Mosimann distribution $\theta_1, \ldots, \theta_k$ are independent, whereas in the Gaussian copula and vine distributions they are dependent. A D-vine can be used to represent the prior distribution $f(\theta_1, \ldots, \theta_{k-1})$. The general structure of a D-vine in this context is given in Fig. 2.6. This type of vine is fully defined by the ordering of the variables in the first tree of the vine. Without loss of generality, suppose this ordering is $(\theta_1, \ldots, \theta_{k-1})$.

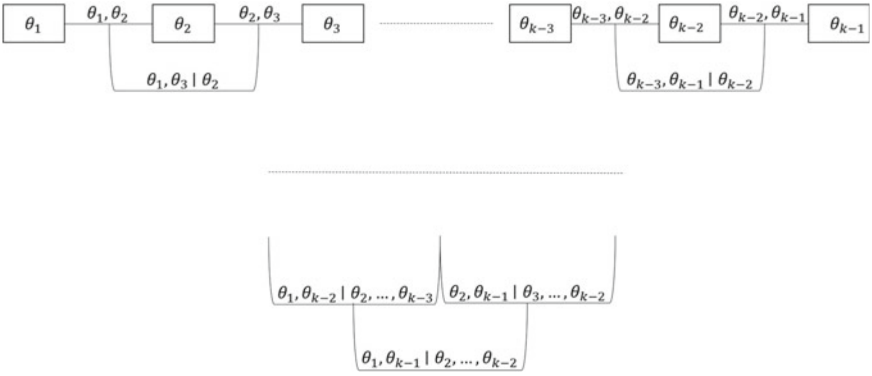

θ_1
θ_1, θ_2
θ_2
θ_2, θ_3
θ_3
$\theta_1, \theta_3 | \theta_2$
θ_{k-3}
$\theta_{k-3}, \theta_{k-2}$
θ_{k-2}
$\theta_{k-2}, \theta_{k-1}$
θ_{k-1}
$\theta_{k-3}, \theta_{k-1} | \theta_{k-2}$
$\theta_1, \theta_{k-2} | \theta_2, \ldots, \theta_{k-3}$
$\theta_2, \theta_{k-1} | \theta_3, \ldots, \theta_{k-2}$
$\theta_1, \theta_{k-1} | \theta_2, \ldots, \theta_{k-2}$

Fig. 2.6 The structure of a general D-vine in $k - 1$ dimensions

The prior distribution is given by the vine distribution representing the D-vine in Fig. 2.6. If a simplified vine structure is assumed, in which the conditional bivariate copulas do not depend on the variables that are conditioned on, the prior distribution is of the form

$$\begin{aligned} f(\theta_1, \ldots, \theta_{k-1}) = & \prod_{i=1}^{k-1} [f_i(\theta_i)] \times \prod_{i=1}^{k-2} \left[c_{i,i+1}\left(F_i(\theta_i), F_{i+1}(\theta_{i+1})\right)\right] \\ & \times \prod_{i=1}^{k-3} \prod_{j=i+2}^{k-1} \left[c_{i,j}\left(F_{i|i+1,\ldots,j-1}(\theta_i|\theta_{i+1}), F_{j|i+1,\ldots,j-1}(\theta_j|\theta_{i+1}, \ldots, \theta_{j-1})\right)\right], \end{aligned} \tag{2.19}$$

where $f_i(\theta_i)$ is the marginal prior PDF of θ_i, $c_{i,j}(\cdot, \cdot)$ is a bivariate copula PDF, $F_i(\theta_i)$ is the marginal prior CDF of θ_i and $F_{i|i+1,\ldots,j-1}(\theta_i|\theta_{i+1} \ldots, \theta_{j-1})$ is the conditional prior CDF of $\theta_i|\theta_{i+1}, \ldots, \theta_{j-1}$.

In order to fully specify the D-vine requires specifying the marginal distributions of each of the probabilities, $f_i(\theta_i)$, the unconditional copulas in the first tree of the vine, $c_{i,i+1}(\cdot, \cdot)$, which represent the unconditional relationships between the multinomial probabilities, and the conditional copulas in trees 2 to $k-2$ of the vine, $c_{i,j}(\cdot, \cdot)$, which represent the conditional relationships between the multinomial probabilities. It is not necessary to specify the prior conditional distributions $F_{i|i+1,\ldots,j-1}(\cdot|\cdot)$, as they can be calculated from the other specifications. For further details, see Wilson (2018).

There are two main stages to the elicitation: the marginal distributions and the bivariate copulas. As in the Gaussian copula approach, it is assumed that each θ_i has a marginal beta distribution with parameters a_i and b_i. The elicitation tasks in this stage are the same as tasks (i)−(iii) in the Gaussian copula approach. The method of Sect. 2.2.2 can be used, as suggested there, to determine the parameters of the beta distributions based on the medians and quartiles of θ_i. Another approach, from Wilson (2018), is to calculate the parameter values which exactly match each of the three pairs of the three quartiles (lower quartile and median, upper quartile and median, lower quartile and upper quartile) and use these to find the means and variances of these beta distributions, $(\mu_{i,j}, \sigma^2_{i,j})$, for $j = 1, 2, 3$. The mean and variance of θ_i can then be calculated as a weighted average of these values, i.e.

$$\mu_i = w_{i,1}\mu_{i,1} + w_{i,2}\mu_{i,2} + w_{i,3}\mu_{i,3}, \tag{2.20}$$

$$\sigma_i^2 = \frac{1}{w_{i,1}^2 + w_{i,2}^2 + w_{i,3}^2}\left(w_{i,1}^2\sigma_{i,1}^2 + w_{i,2}^2\sigma_{i,2}^2 + w_{i,3}^2\sigma_{i,3}^2\right), \tag{2.21}$$

for weights $w_{i,j}$, where $w_{i,1} + w_{i,2} + w_{i,3} = 1$. The weights could be chosen to be equal by default or a higher weight could be given to pairs of quantiles in which the expert is more confident. The parameter values (a_i, b_i) can then be found directly.

To assess the unconditional and conditional bivariate copulas, consider the unconditional probabilities, in the chosen ordering, $(p_1, \ldots, p_{k-1})$. For the copulas involving θ_1, the expert performs the following tasks.

(iv) She is asked to suppose that the values of $p_1, \ldots, p_{i-1}$ have been observed at their specified medians in the marginal elicitation, $p_1 = m_1^*, \ldots, p_{i-1} = m_{i-1}^*$. It would be expected in this case that her median of p_i would remain as m_i^*. She is asked to confirm this is the case.

(v) She is then asked for her upper and lower quartiles of p_i again conditioning on $p_1 = m_1^*, \ldots, p_{i-1} = m_{i-1}^*$. It would be expected that learning that the values of the proportions in the previous categories were equal to her prior medians would reduce her uncertainty about the value of p_i given that there is dependence between p_i and $p_1, \ldots, p_{i-1}$ in her prior beliefs.

These two steps are performed for each of the copulas which contain θ_1, with the resulting quartiles being L_i, M_i and U_i, which are from the conditional distribution of $p_i | p_1, \ldots, p_{i-1}$. To convert to quantiles of $\theta_i | \theta_1, \ldots, \theta_{i-1}$ (Elfadaly and Garthwaite 2017), we use (2.18).

A number of candidate parametric bivariate copulas are fitted to these three quantiles using least squares. Possible choices here are the Gaussian copula, t-copula, Gumbel copula, Clayton copula and Frank copula. The best fitting of these copulas using least squares is chosen to represent the bivariate relationship between $\theta_1, \theta_i | \theta_2, \ldots, \theta_{i-1}$. Full details are provided in Wilson (2018). To specify the copulas not involving θ_1, the expert performs the following task.

(vi) She is asked to suppose that an observation hasn't fallen into the first $j - 1$ categories. This leaves $p_j, \ldots, p_{k-1}$, where the remaining unconditional probabilities are $p_j = \theta_j$ and $p_{j+k} = \theta_{j+k} \prod_{r=j}^{j+k-1} (1 - \theta_r)$. Steps (iv) and (v) are then repeated to obtain the copulas between $\theta_j, \theta_i \mid \theta_{j+1}, \ldots, \theta_{i-1}$, where $i > j$.

Step (vi) is repeated for each $j \geq 2$ until all of the copulas in the vine have been specified.

2.3.3.2 Elicitation Exercise

In this section, the results of an elicitation exercise are reported, conducted to assess the ability of experts to specify a prior distribution for multinomial probabilities using vines and to assess the impact of this prior structure. The expert chosen is a mathematics teacher from a comprehensive school (for students between 11 and 16) in England with no specialist training in statistics. He was first given a tutorial in probability and statistical concepts necessary for the elicitation. He was then asked to assess the proportions of pupils from his school who would achieve specific grades in their GCSE exams in August 2016.

This part of the elicitation took place in two stages: first the marginal distributions for the probabilities were elicited and then the dependencies were elicited, both using the methods in Sect. 2.3.3.1. The expert was then asked to consider a scenario in which the pupils had been given extra lessons in mathematics prior to their exams. The elicitation procedure was repeated for this new scenario. The expert felt comfortable giving all of the required specifications.

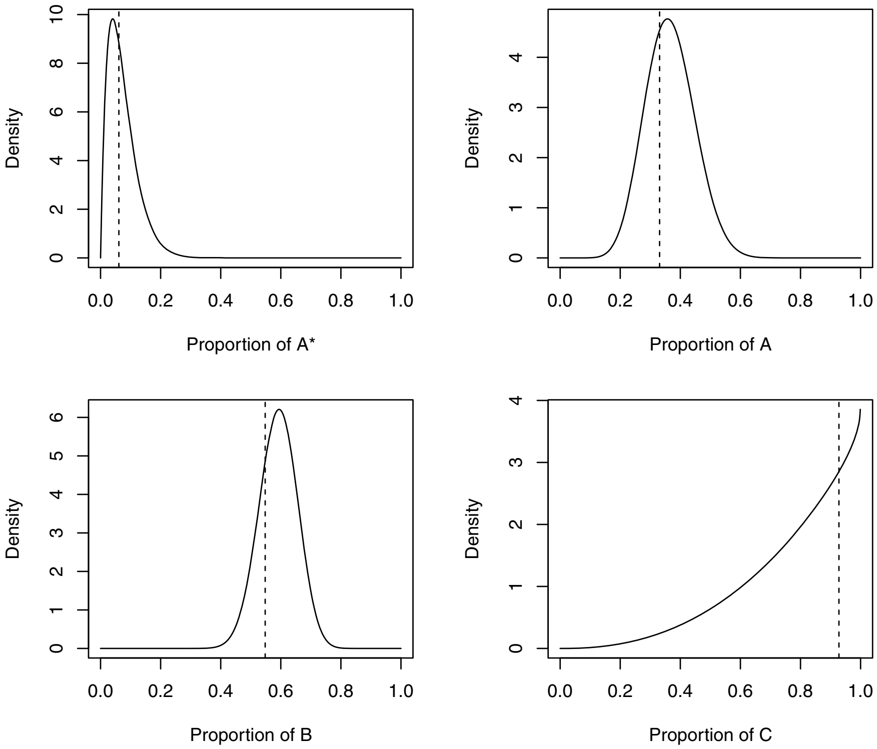


Fig. 2.7 Prior beta distributions for the proportion of students achieving A*, A, B and C given that they did not achieve a higher grade, and the actual proportions

In this section, the elicitation session with the expert is discussed in order to show how the elicitation of the quantities detailed in the previous section can be realistically achieved in practice. In particular, the accuracy of the prior distributions assessed by the expert is considered. To do this, the actual proportions of students achieving specific GCSE grades in the teacher's school are compared to their prior distribution. The effect of conditioning on some proportions in the prior will be investigated to assess the suitability of the expert's assessments of dependencies.

Initially, the marginal specifications made by the expert are considered. Given the three quantile specifications made for the conditional proportion of students achieving each grade given that they did not achieve a higher grade, a beta distribution is fitted to each proportion using the methods described above. The resulting marginal prior distributions are given in Fig. 2.7, alongside the actual conditional proportion of students achieving that grade in 2016, given that they didn't achieve a higher grade, which for A*, A, B, C, and below C were, respectively, (0.06, 0.33, 0.55, 0.93, 1). The first proportion is the unconditional proportion of students with grade A*.

The true conditional proportion of students achieving each grade was given a non-zero probability in each prior distribution. In the case of the proportions of students

Table 2.2 The interquartile ranges as specified by the expert for each of the marginal conditional proportions and the actual value of the conditional proportion

Grade	Lower quartile	Upper quartile	Actual proportion
θ_1	0.05	0.10	0.06
θ_2	0.34	0.42	0.33
θ_3	0.56	0.62	0.55
θ_4	0.48	0.61	0.93

achieving each of the four grades, the actual proportions are reasonably close to the mode of the relevant prior distribution. In the case of the proportions of students achieving an A conditional on not achieving an A* and achieving a B conditional on not achieving an A* or A, the observed proportions are slightly below the prior mode and in the case of the proportions of students achieving grades A* and C conditional on not achieving any of the higher grades the observed proportions are slightly above the prior mode. It seems that more students either did very well in their exam or quite poorly in their exam than this teacher expected.

Consider the raw specifications made by the expert. Table 2.2 provides the interquartile ranges as specified by the expert for each of the marginal conditional proportions and the actual value of the proportion. In the table, θ_1 is the proportion of students achieving an A*, θ_2 is the proportion achieving an A of those who did not achieve an A*, θ_3 is the proportion achieving a B of those who did not achieve an A* or A and θ_4 is the proportion achieving a C of those who did not achieve an A*, A or B.

One of the proportions, that for θ_1, fell within the interquartile range of the expert. The proportions of students achieving grades A and B conditional on higher grades, θ_2 and θ_3, were very close to the lower limits of the interquartile ranges of the expert. Thus, the expert's upper and lower quartiles seem reasonable, and may be consistent with 50% of observations falling within the assessed interquartile range over a larger number of assessments.

The effect of the dependencies expressed by the expert can be assessed. This could be done by comparing observed conditional proportions of students achieving specific grades to their posterior distributions having observed the proportions of students achieving other grades. However, as interest is in the dependencies in the prior distribution, instead the observed proportions are compared to the prior conditional distributions for the grades of interest conditioned on specific values of the other proportions. This provides a better understanding of the dependencies expressed in the prior distribution.

Specifically, the prior distributions for the proportions of students achieving grades A and B given that they didn't achieve higher grades are compared with their prior

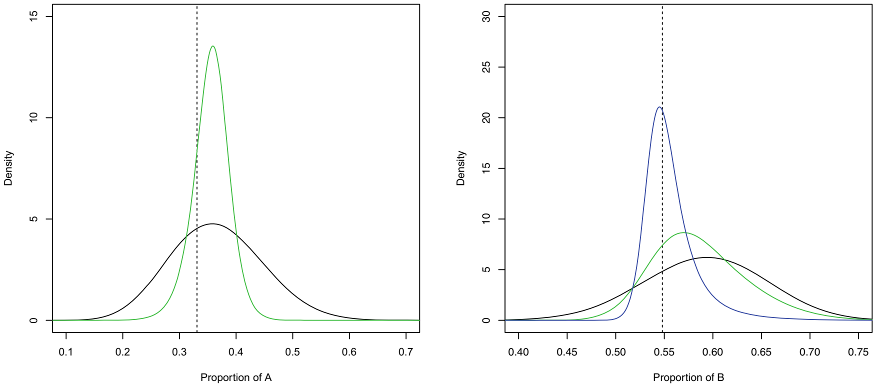


Fig. 2.8 Prior beta distributions for θ_2 and θ_3 (black), the prior conditional distributions of $\theta_2|\theta_1 = 0.06$, and $\theta_3|\theta_2 = 0.33$ (green) and $\theta_3|\theta_1 = 0.06, \theta_2 = 0.33$ (blue) and the conditional proportions of students achieving these grades (vertical dashed lines)

conditional distributions conditioning on the observed proportions above. That is, the prior conditional distributions for $\theta_2|\theta_1 = 0.06, \theta_3|\theta_2 = 0.33$ and $\theta_3|\theta_1 = 0.06, \theta_2 = 0.33$. For example, $\theta_2|\theta_1 = 0.06$ is the proportion of students achieving an A who did not achieve an A* given that 6% of students achieved an A*. The results are given in Fig. 2.8.

In each case, the dependency between the proportions has led to a reduction in uncertainty when conditioning on the proportion of students achieving a previous grade. That is, conditioning on one of the proportions reduces the uncertainty about the proportions of students achieving other grades. In the case of θ_2, θ_1 is conditioned on being very close to the expert's median value. The expert has specified positive dependence between θ_1 and θ_2, with the fitted copula giving Kendall's Tau value of 0.75 based on the values that the expert gave. Thus, the conditional distribution for $\theta_2|\theta_1 = 0.06$ has a mode which is very close to that of the prior distribution for θ_2.

In the case of the prior conditional distribution of $\theta_3|\theta_2 = 0.33$, then θ_2 is conditioned at a value slightly lower than the expert's prior mode. The expert assessed the dependence between θ_2 and θ_3 to be positive, with Kendall's Tau value of 0.43. Thus the conditional distribution of $\theta_3|\theta_2 = 0.33$ has a mode which is lower than that of the unconditional prior distribution. The expert assessed the dependence between $\theta_1, \theta_3|\theta_2 = 0.33$ to be negative, with Kendall's Tau value of -0.41. Thus, conditioning on θ_1 to be above its prior mode has led to the prior conditional distribution of $\theta_3|\theta_1 = 0.06, \theta_2 = 0.33$ having a mode lower than that of $\theta_3|\theta_2 = 0.33$. In each case, there is a reduction in uncertainty when conditioning on variables.

Based on this analysis, it seems that the dependencies specified by the expert in the elicitation exercise are reasonable.

2.4 Eliciting Prior Distributions for Multinomial Models that Contain Covariates

Multinomial sampling models often contain covariates that influence membership probabilities for the different categories. Methods of eliciting a prior distribution for this sampling model have been developed for the special case of logistic regression, in which there are only two categories (c.f. Bedrick et al. 1996; Garthwaite et al. 2013) but, until recently, the general case with more than two categories has attracted little attention. The focus is the task of eliciting a prior distribution for a multinomial logistic regression model, in which there are more than two categories and a generalised linear regression model that links covariates to the membership probabilities.

Elfadaly and Garthwaite (2019) develop a new elicitation method for this task that quantifies opinion as a multivariate normal prior distribution on the regression coefficients. They first address the simpler case where there are no covariates and represent opinion as a logistic normal distribution. This distribution was introduced as a flexible sampling distribution to model compositional data, i.e. proportions that sum to one over the simplex (c.f. Aitchison 1986). The distribution has a more flexible dependence structure and more parameters than the Dirichlet distribution, making it an attractive choice as a prior distribution for a multinomial sampling model without covariates. O'Hagan and Forster (2004), Sect. 12.14–12.19 gives some theoretical results on the prior–posterior analysis for this prior. The benefit of the logistic normal distribution is that it can be extended in a natural way to form a prior distribution that is appropriate for a multinomial logistic regression model. Moreover, the elicitation method that yields the parameters of a logistic normal distribution can be expanded to elicit the parameters of the corresponding multivariate normal prior distribution (Elfadaly and Garthwaite 2019). The elicitation method is implemented in interactive software that is freely available on the web at http://statistics.open.ac.uk/elicitation.

In this section, the elicitation method for a multinomial model without covariates is reviewed, and then its extension of multinomial models that contain covariates is discussed. In Sect. 2.4.1, the logistic normal prior distribution and its assumptions are briefly reviewed. The assessment tasks and their use to encode the hyperparameters of the prior and obtain feedback are given in Sect. 2.4.2. In Sect. 2.4.3, the scope of the logistic normal prior is extended so that it is a suitable prior distribution for a multinomial logistic regression model.

2.4.1 The Logistic Normal Prior Distribution

Aitchison (1986) introduced different forms of multivariate logistic transformations from normally distributed variates to $\mathbf{p} = (p_1, \ldots, p_k)'$. The one most widely used is the additive transformation that defines additive logistic normal distributions. Following Aitchison, put $\boldsymbol{Y}_{k/1} = (Y_2, \ldots, Y_k)'$, where the first category is suppressed. The additive logistic transformation from $\boldsymbol{Y}_{k/1}$ to $\mathbf{p}$ is defined by

$$p_1 = 1/[1 + \sum_{j=2}^{k} \exp(Y_j)], \quad p_i = \exp(Y_i)/[1 + \sum_{j=2}^{k} \exp(Y_j)], \quad i = 2, \ldots, k, \tag{2.22}$$

with the inverse transformation

$$Y_i = \log(p_i/p_1) \equiv \log(r_i), \quad i = 2, \ldots, k. \tag{2.23}$$

We refer to p_1 as the fill-up variable and the first category as the fill-up category. Its choice is arbitrary, in principle, but we believe that choosing it as the most common category makes performing the assessment tasks easier for the expert. The vector $\mathbf{p}$ has a logistic normal distribution if

$$\boldsymbol{Y}_{k/1} \sim \text{MVN}(\boldsymbol{\mu}_{k/1}, \boldsymbol{\Sigma}_{k/1}). \tag{2.24}$$

Assuming the prior distribution takes this form, an expert's opinion needs to be quantified to encode the hyperparameters $\boldsymbol{\mu}_{k/1} = (\mu_2, \ldots, \mu_k)$ and $\boldsymbol{\Sigma}_{k/1}$.

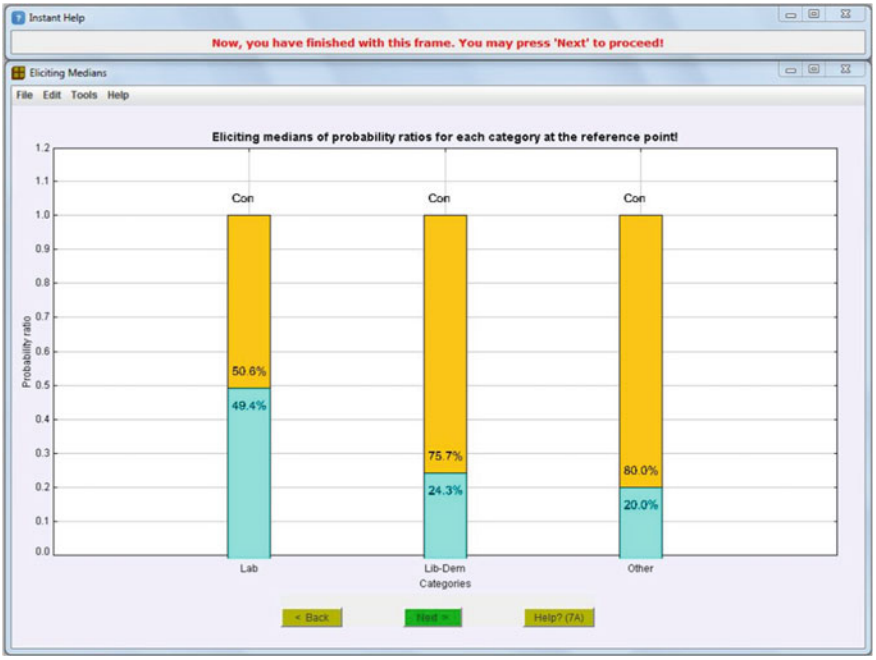
2.4.2 Eliciting Priors for Multinomial Models Without Covariates

The expert gives her assessments for the probability ratios $r_i = p_i/p_1$ $(i = 2, \ldots, k)$. Assessing ratios of proportions is a reasonably easy task that is frequently used in elicitation methods (c.f. Elfadaly and Garthwaite 2019, and the references therein). Using a bar chart within the software (c.f. Fig. 2.9), the expert is asked to focus on the proportion r_i of those items that fall in the ith category (represented as the volume in the lower (blue) part of a box) relative to the proportion that falls in the first (fill-up) category, represented as the volume in the upper (orange) part of the box. For example, to obtain the median assessment at the first column of Fig. 2.9 in the context of voting behaviour in a local election, the expert was asked "Of the people who vote for either the Conservative party or the Labour party, what proportion will vote for Labour?" She assessed the median of this proportion as 49.4%.

On a successive set of bar charts similar to that in Fig. 2.9, the expert assesses medians and quartiles of r_i to encode $\boldsymbol{\mu}_{k/1}$ and $\boldsymbol{\Sigma}_{k/1}$. The main approach is to transform the assessed quantiles of r_i into the corresponding quantiles of Y_i using the monotonicity of the log transformation. This is detailed as follows.

2.4.2.1 Encoding $\boldsymbol{\mu}_{k/1}$ and $\boldsymbol{\Sigma}_{k/1}$

(i) To obtain $\boldsymbol{\mu}_{k/1}$, the expert assesses her medians, m_i^*, of each probability ratio r_i. Then, using (2.23)–(2.24), the monotonicity of the log transformation and the



Instant Help

Now, you have finished with this frame. You may press 'Next' to proceed!

Eliciting Medians

Eliciting medians of probability ratios for each category at the reference point!

Fig. 2.9 Assessing medians of probability ratios

symmetry of the distribution of Y_i $(i = 2, \ldots, k)$, the components of $\boldsymbol{\mu}_{k/1}$ are simply encoded as

$$\mu_i = E(Y_i) = \log(m_i^*), \quad \text{for } i = 2, \ldots, k. \tag{2.25}$$

To determine $\boldsymbol{\Sigma}_{k/1} = \mathrm{Var}(\boldsymbol{Y}_{k/1})$, two extra sets of assessments are required.

(ii) On a bar chart similar to that in Fig. 2.9, the expert assesses the lower (upper) quartiles, L_2^* (U_2^*) for the probability ratio r_2. This gives the variance of Y_2 as (Elfadaly and Garthwaite 2019)

$$\mathrm{Var}(Y_2) = \left\{[\log(U_2^*) - \log(L_2^*)]/1.349\right\}^2, \tag{2.26}$$

where 1.349 is the interquartile range of a standard normal distribution. Then for the remaining categories, the expert assesses conditional lower (upper) quartiles L_{i+1}^* (U_{i+1}^*) of r_{i+1} given that $r_j = m_j^*$ for $j = 2, \ldots, i$ and $i = 2, \ldots, k$. These conditional quartile assessments are used to encode the conditional variances (Elfadaly and Garthwaite 2019)

$$\begin{aligned}\mathrm{Var}[Y_{i+1} \mid Y_2 = \log(m_2^*), \ldots, Y_i = \log(m_i^*)]&\\ = \left\{[\log(U_{i+1}^*) - \log(L_{i+1}^*)]/1.349\right\}^2, &\quad \text{for } i = 2, \ldots, k-1.\end{aligned} \tag{2.27}$$

(iii) For $i = 2, \ldots, k-1$, the expert is asked to assess conditional medians of r_i given values for $r_{i+1}, \ldots, r_k$ that are chosen from the previously assessed medians and lower quartiles. Specifically, the expert assesses m_{2j}^* as her conditional median of $(r_j \mid r_2 = L_2^*)$. This gives the conditional mean of Y_j

$$E[Y_j \mid Y_2 = \log(L_2^*)] = \log(m_{2j}^*), \quad \text{for } j = 3, \ldots, k. \tag{2.28}$$

Then using a series of bar chart graphs, the expert assesses the values m_{ij}^* as the conditional medians of $(r_j \mid r_2 = m_2^*, \ldots, r_{i-1} = m_{i-1}^*, r_i = L_i^*)$ for $i = 3, \ldots, k-1$, $j = i+1, \ldots, k$. The following set of conditional means are obtained:

$$\begin{aligned}E[Y_j \mid Y_2 = \log(m_2^*), \ldots, Y_{i-1} = \log(m_{i-1}^*), Y_i = \log(L_i^*)]&\\ = \log(m_{ij}^*), \quad \text{for } i = 3, \ldots, k-1;\ j = i+1, \ldots, k.&\end{aligned} \tag{2.29}$$

Based on a method developed by Kadane et al. (1980), Elfadaly and Garthwaite (2019) use the quantities encoded in (2.25)–(2.29) to obtain $\boldsymbol{\Sigma}_{k/1} = \mathrm{Var}(\mathbf{Y}_{k/1})$ as a matrix that is certain to be positive definite.

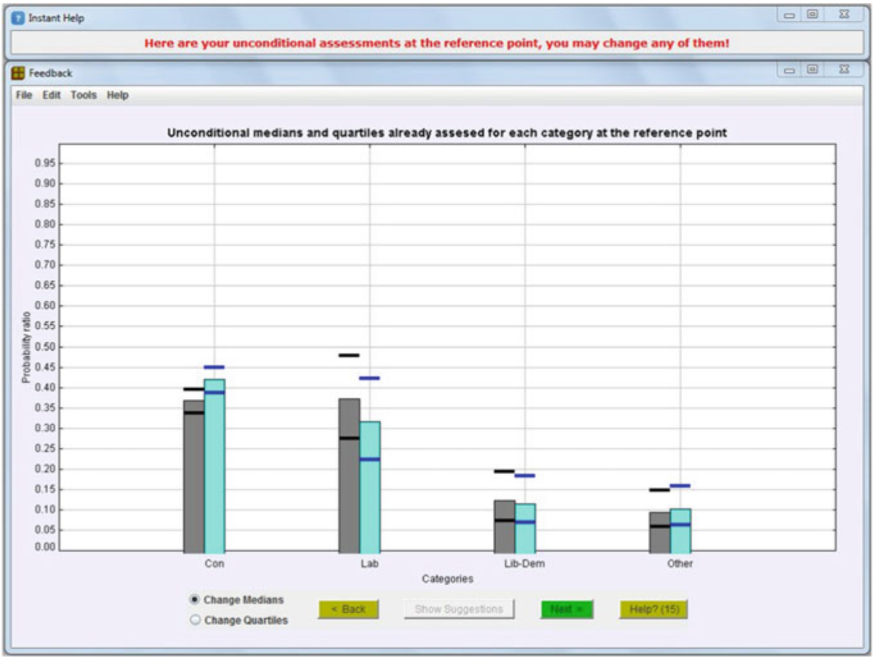
2.4.2.2 Feedback on the Elicited Prior Distribution

The expert is given feedback through a bar chart displaying the unconditional median and quartiles of each p_j $(j = 1, \ldots, k)$ that are implied by her prior distribution (c.f. Fig. 2.10). She is then invited to keep revising those quantiles until they form an acceptable representation of her opinion.

In giving her feedback, the expert is encouraged to consider her opinion from a different perspective. Here, she gives and revises assessments on the probabilities p_j instead of the probability ratios r_j. Also, the required assessments here are all unconditional assessments, while most of the initial assessments were conditional.

There are no closed-form equations for the unconditional moments or quartiles of the logistic normal distribution, so Elfadaly and Garthwaite (2019) developed a method for estimating the quartiles of p_j from the encoded $\boldsymbol{\mu}_{k/1}$ and $\boldsymbol{\Sigma}_{k/1}$. They also determined these hyperparameters from the expert's revised quantiles of the p_j.

These feedback screens can also be a useful means of reducing the number of assessments that are required when the model contains covariates, as detailed in the next section.


Instant Help

Here are your unconditional assessments at the reference point, you may change any of them!

Feedback

Fig. 2.10 Feedback showing marginal medians and quartiles of each p_j: Initial median assessments are given as the fixed grey bars. The blue bars represent the expert's interactively revised medians, and the short horizontal (blue) lines are the revised quartiles

2.4.3 Eliciting Priors for Multinomial Models with Covariates

Elfadaly and Garthwaite (2019) also address the common situation where the membership probabilities of the multinomial sampling model are influenced by explanatory covariates. The problem is expressed as a multinomial logistic regression model for which a multivariate normal prior distribution on the regression coefficients is assumed. The hyperparameters of this prior distribution are encoded by extending the method discussed in Sect. 2.4.2. The method is repeatedly applied to quantify the expert's opinion at different specific values of the covariates. But first the sampling and prior models are described, and their main assumptions are provided.

2.4.3.1 Sampling and Prior Models

Let $p_i(\boldsymbol{\xi})$ denote the probability that an observation with m covariate values $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_m)'$ falls in the ith category of a multinomial model ($i = 1, \ldots, k$). The multinomial logistic (logit) model is obtained by putting Y_i equal to $\alpha_i + \boldsymbol{\xi}'\boldsymbol{\beta}_i$ in

(2.22), where α_i and $\boldsymbol{\beta}_i = (\beta_{1,i}, \ldots, \beta_{m,i})'$ are the constant and vector of regression coefficients for the i^{th} category ($i = 2, \ldots, k$). Hence,

$$p_i(\boldsymbol{\xi}) = \begin{cases} 1/[1 + \sum_{j=2}^{k} \exp(\alpha_j + \boldsymbol{\xi}'\boldsymbol{\beta}_j)], & i = 1 \\ \exp(\alpha_i + \boldsymbol{\xi}'\boldsymbol{\beta}_i)/[1 + \sum_{j=2}^{k} \exp(\alpha_j + \boldsymbol{\xi}'\boldsymbol{\beta}_j)], & i = 2, \ldots, k. \end{cases} \tag{2.30}$$

Rearranging the regression coefficients into a matrix, say $\mathbf{B}$, of the form

$$\mathbf{B} = \left[\begin{pmatrix} \alpha_2 \\ \boldsymbol{\beta}_2 \end{pmatrix}, \ldots, \begin{pmatrix} \alpha_k \\ \boldsymbol{\beta}_k \end{pmatrix} \right], \tag{2.31}$$

the rows of $\mathbf{B}$ are defined as $\boldsymbol{\alpha}$, $\boldsymbol{\beta}_{(r)}$, for $r = 1, 2, \ldots, m$, so that

$$\boldsymbol{\alpha} = (\alpha_2, \ldots, \alpha_k)' \quad \text{and} \quad \boldsymbol{\beta}_{(r)} = (\beta_{r,2}, \ldots, \beta_{r,k})'. \tag{2.32}$$

Under the assumption that $(\boldsymbol{\alpha}', \boldsymbol{\beta}'_{(1)}, \ldots, \boldsymbol{\beta}'_{(m)})'$ has a multivariate normal prior distribution, MVN($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$), the hyperparameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are to be encoded.

It is assumed that, given $\boldsymbol{\alpha}$, the vectors $\boldsymbol{\beta}_{(r)}$ and $\boldsymbol{\beta}_{(s)}$ are *a priori* independent for all r and s ($r \neq s$); hence, $\boldsymbol{\Sigma}_{|\alpha} = \text{Var}(\boldsymbol{\beta}'_{(1)}, \ldots, \boldsymbol{\beta}'_{(m)} \,|\, \boldsymbol{\alpha})$ is a block-diagonal matrix:

$$\boldsymbol{\Sigma}_{|\alpha} = \begin{pmatrix} \boldsymbol{\Sigma}_{\beta,1|\alpha} & \text{O} & \text{O} \\ \text{O} & \ddots & \text{O} \\ \text{O} & \text{O} & \boldsymbol{\Sigma}_{\beta,m|\alpha} \end{pmatrix}, \tag{2.33}$$

where $\boldsymbol{\Sigma}_{\beta,r|\alpha} = \text{Var}(\boldsymbol{\beta}_{(r)}|\boldsymbol{\alpha})$. The elements of $\boldsymbol{\alpha}$ must be correlated to reflect the unit sum constraint of $\mathbf{p}$, and so must be the elements of each $\boldsymbol{\beta}_{(r)}$.

Each continuous (categorical) covariate is given a reference value (level) and, in the elicitation process, they are assumed to vary one at a time, while other covariates are assumed to be at their reference values/levels. Each covariate is centred so that its reference value/level is zero.

Let $\boldsymbol{\xi}_0$ be the $m \times 1$ vector of 0s. When $\boldsymbol{\xi} = \boldsymbol{\xi}_0$, all variables take their reference value (0) and $\boldsymbol{\xi}_0$ is referred to as the *reference point*. Also, for $r = 1, \ldots, m$, let $\boldsymbol{\xi}_r^*$ denote an $m \times 1$ vector whose elements are 0 apart from its rth element, which equals some specific value, ξ_r^* say. For categorical covariates, ξ_r^* is set to 1.

Having listed all the assumptions of the model, in the rest of this section, the main assessment tasks that are required to obtain $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in the elicitation method of Elfadaly and Garthwaite (2019) are reviewed.

2.4.3.2 Encoding $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

Let $\boldsymbol{\mu} = (\boldsymbol{\mu}'_\alpha, \boldsymbol{\mu}'_{\beta,1}, \ldots, \boldsymbol{\mu}'_{\beta,m})'$, where $\boldsymbol{\mu}_\alpha = E(\boldsymbol{\alpha})$ and $\boldsymbol{\mu}_{\beta,r} = E(\boldsymbol{\beta}_{(r)})$ for $r = 1, \ldots, m$. To elicit $\boldsymbol{\mu}_\alpha$, the expert is asked to consider just the subpopulation of items whose covariates are all at the reference point ($\boldsymbol{\xi} = \boldsymbol{\xi}_0$). The same assessment tasks as in Sect. 2.4.2.1(i) will then give $\boldsymbol{\mu}_\alpha = E(\boldsymbol{Y}_{k/1} \,|\, \boldsymbol{\xi}_0)$.

To elicit $\boldsymbol{\mu}_{\beta,r}$, the expert repeats the assessment tasks of Sect. 2.4.2.1(i) for the subpopulation for which $\boldsymbol{\xi} = \boldsymbol{\xi}_r^*$. These yield $E(\boldsymbol{Y}_{k/1} \mid \boldsymbol{\xi}_r^*) = \boldsymbol{\mu}_\alpha + \xi_r^* \boldsymbol{\mu}_{\beta,r}$, as ξ_r^* is the only non-zero element of $\boldsymbol{\xi}_r^*$. This gives $\boldsymbol{\mu}_{\beta,r} = [E(\boldsymbol{Y}_{k/1} \mid \boldsymbol{\xi}_r^*) - \boldsymbol{\mu}_\alpha]/\xi_r^*$.

The submatrices of $\boldsymbol{\Sigma}$ are encoded as follows. For $r = 1, \ldots, m$, define $\boldsymbol{\Sigma}_\alpha$, $\boldsymbol{\Sigma}_{\alpha,\beta,r}$ and $\boldsymbol{\Sigma}_{\beta,r}$ from the conformal partitioning

$$\text{Var}\begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta}_{(r)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_\alpha & \boldsymbol{\Sigma}'_{\alpha,\beta,r} \\ \boldsymbol{\Sigma}_{\alpha,\beta,r} & \boldsymbol{\Sigma}_{\beta,r} \end{pmatrix}. \tag{2.34}$$

The assessment tasks in Sect. 2.4.2.1(ii)–(iii) at the reference point (i.e. assuming $\boldsymbol{\xi} = \boldsymbol{\xi}_0$) gives $\boldsymbol{\Sigma}_\alpha = \text{Var}(\boldsymbol{Y}_{k/1} \mid \boldsymbol{\xi}_0)$. The expert is then asked to repeat the same assessment tasks under the two assumptions that $\boldsymbol{\xi} = \boldsymbol{\xi}_r^*$ and that the earlier median assessments of $r_2(\boldsymbol{\xi}_0), \ldots, r_k(\boldsymbol{\xi}_0)$ are the true values of these ratios (hence fixing the value of $\boldsymbol{\alpha}$). This gives $\text{Var}(\boldsymbol{Y}_{k/1} \mid \boldsymbol{\xi} = \boldsymbol{\xi}_r^*, \boldsymbol{\alpha}) = \text{Var}(\xi_r^* \boldsymbol{\beta}_{(r)} \mid \boldsymbol{\alpha})$, which determines $\boldsymbol{\Sigma}_{\beta,r|\alpha}$ from

$$\boldsymbol{\Sigma}_{\beta,r|\alpha} = \text{Var}(\boldsymbol{Y}_{k/1} \mid \boldsymbol{\xi} = \boldsymbol{\xi}_r^*, \boldsymbol{\alpha}) \,/\, (\xi_r^*)^2. \tag{2.35}$$

Repeating the process for $r = 1, \ldots, m$ gives $\boldsymbol{\Sigma}_{\beta,1|\alpha}, \ldots, \boldsymbol{\Sigma}_{\beta,m|\alpha}$, and $\boldsymbol{\Sigma}_{|\alpha}$ from (2.33).

To complete the encoding of $\boldsymbol{\Sigma}$, it remains to elicit $\boldsymbol{\Sigma}_{\alpha,\beta,r}$. To do this a number of $k - 1$ different values are chosen for $\boldsymbol{\alpha}^*$, say $\boldsymbol{\alpha}_2^*, \ldots, \boldsymbol{\alpha}_k^*$ (for details on their choice, see Elfadaly and Garthwaite 2019) and, for each $\boldsymbol{\alpha}_i^*$, assessments are obtained from the expert that determine $E(\boldsymbol{\beta}_{(r)}|\boldsymbol{\alpha} = \boldsymbol{\alpha}_i^*)$. Specifically, she is given chosen values for $r_2(\boldsymbol{\xi}_0), \ldots, r_k(\boldsymbol{\xi}_0)$ that yield $\boldsymbol{\alpha} = \boldsymbol{\alpha}_i^*$. Under the assumption that these are the true probability ratios at the reference point and that the population is restricted to items for which $\boldsymbol{\xi} = \boldsymbol{\xi}_r^*$, the expert then repeats the assessment tasks of Sect. 2.4.2.1(iii). These yield $E(\boldsymbol{Y}_{k/1}|\boldsymbol{\xi}_r^*, \boldsymbol{\alpha} = \boldsymbol{\alpha}_i^*)$, which gives

$$E(\boldsymbol{\beta}_{(r)}|\, \boldsymbol{\alpha} = \boldsymbol{\alpha}_i^*) = \{E(\boldsymbol{Y}_{k/1}|\, \boldsymbol{\xi}_r^*, \boldsymbol{\alpha} = \boldsymbol{\alpha}_i^*) - \boldsymbol{\alpha}_i^*\}/\xi_r. \tag{2.36}$$

Finally, put $\boldsymbol{\Sigma}_{\alpha,\beta} = (\boldsymbol{\Sigma}'_{\alpha,\beta,1}, \ldots, \boldsymbol{\Sigma}'_{\alpha,\beta,m})'$ and $\boldsymbol{\Sigma}_\beta = \boldsymbol{\Sigma}_{|\alpha} + \boldsymbol{\Sigma}_{\alpha,\beta} \boldsymbol{\Sigma}_\alpha^{-1} \boldsymbol{\Sigma}'_{\alpha,\beta}$. Then a positive definite matrix $\boldsymbol{\Sigma}$ is obtained from

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_\alpha & \boldsymbol{\Sigma}'_{\alpha,\beta} \\ \boldsymbol{\Sigma}_{\alpha,\beta} & \boldsymbol{\Sigma}_\beta \end{pmatrix}. \tag{2.37}$$

The software provides a short-cut option to reduce the number of required assessments, which is a very beneficial tool if there are several covariates in the model. Instead of repeating the assessment tasks at $\boldsymbol{\xi}_r^*$, the expert may choose to just modify the marginal quartiles of the probabilities that were given as feedback (see Sect. 2.4.2.2) at the reference point, i.e. those at $\boldsymbol{\xi} = \boldsymbol{\xi}_0$. The expert revises those assessments (as the blue bars on a bar chart similar to that in Fig. 2.10) to reflect her opinions when the values of the covariates are $\boldsymbol{\xi}_r^*$ instead of $\boldsymbol{\xi}_0$ (with assessments fixed as the grey bars).

This option greatly shortens the elicitation process, but the price is imposing the correlation structure of the reference point into all situations of $\boldsymbol{\xi} = \boldsymbol{\xi}_r^*$. However, in practice this will not often seem an unrealistic assumption.

2.5 Summary

The Dirichlet distribution has been utilised extensively, as a result of its conjugacy with the multinomial distribution, as the natural prior distribution for multinomial probabilities. In the first part of this chapter, an approach for the careful elicitation and diagnostic assessment of the Dirichlet distribution was presented, within the SHELF framework. This tool allows the user to assess whether the Dirichlet distribution is a suitable representation of an expert's beliefs.

The relative inflexibility of the Dirichlet distribution, particularly in the specification of dependencies between multinomial probabilities, has resulted in various more flexible alternatives to the Dirichlet distribution being proposed. This raises the question of how to decide which of the models should be used to represent an expert's opinion. When one model is a simplification of another model, then it can be relatively straightforward to see whether the additional complexity is beneficial. For example, when the parameters of the Connor–Mosimann distribution meet certain conditions ($b_i = a_{i+1} + b_{i+1}$ for $i = 1, \ldots, k-2$; see Sect. 2.3.1), then it becomes a Dirichlet distribution. Hence, after quantifying opinion as a Connor–Mosimann distribution, examining whether these conditions are approximately satisfied can indicate whether a simpler Dirichlet distribution should be fitted to the assessments. More generally, the assessments given by the expert (medians, quartiles, conditional medians, etc.) can be compared with the corresponding quantities given by a model; large differences would suggest that the model is not representing the expert's opinions satisfactorily, and a more flexible model is needed. However, further research is needed before practical guidance can be given about the approximate levels of flexibility and complexity for particular types of problem.

This chapter has provided details of some of the most promising approaches for quantifying opinion about a multinomial distribution, with a particular emphasis on the elicitation tasks necessary to elicit the prior distribution in each case. It is therefore a handy resource for those confronted with the task of eliciting multinomial probabilities.

References

Aitchison, J. (1986). *The statistical analysis of compositional data*. London: Chapman and Hall.

Bedford, T., & Cooke, R. M. (2002). Vines-a new graphical model for dependent random variables. *Annals of Statistics*, *30*, 1031–1068.

Bedford, T., Daneshkhah, A., & Wilson, K. J. (2016). Approximate uncertainty modeling in risk analysis with vine copulas. *Risk Analysis*, *36*, 792–815.

Bedrick, E. J., Christensen, R., & Johnson, W. (1996). A new perspective on priors for generalized linear models. *Journal of the American Statistical Association*, *91*, 1450–1460.

Bunn, D. W. (1978). Estimation of a Dirichlet prior distribution. *Omega*, *6*, 371–373.

Chaloner, K., & Duncan, G. T. (1987). Some properties of the Dirichlet-multinomial distribution and its use in prior elicitation. *Communications in Statistics-Theory and Methods*, *16*, 511–523.

Clemen, R. C., & Reilly, T. (1999). Correlations and copulas for decision and risk analysis. *Management Science*, *45*, 208–224.

Connor, R. J., & Mosimann, J. E. (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association*, *64*, 194–206.

Dickey, J.M., Jiang, J.M., & Kadane, J.B. (1983). Bayesian methods for multinomial sampling with noninformatively missing data. Technical Report 6/83-#15, State University of New Yourk at Albany, Department of Mathematics and Statistics.

EFSA. (2014). Guidance on expert knowledge elicitation in food and feed safety risk assessment. *EFSA Journal*, 12(6). https://doi.org/10.2903/j.efsa.2014.3734.

Elfadaly, F. G., & Garthwaite, P. H. (2013). Eliciting Dirichlet and Connor-Mosimann prior distributions for multinomial models. *Test*, *22*, 628–646.

Elfadaly, F. G., & Garthwaite, P. H. (2017). Eliciting Dirichlet and Gaussian copula prior distributions for multinomial models. *Statistics and Computing*, *27*, 449–467.

Elfadaly, F.G., & Garthwaite, P.H. (2019). https://doi.org/10.1111/rssa.12546. Submitted for publication.

Evans, M., Guttman, I., & Li, P. (2017). Prior elicitation, assessment and inference with a Dirichlet prior. *Entropy*, *19*, 564. https://doi.org/10.3390/e19100564.

Fan, D. Y. (1991). The distribution of the product of independent beta variables. *Communications in Statistics-Theory and Methods*, *20*, 4043–4052.

Frees, E. W., & Valdez, E. A. (1998). Understanding relations using copulas. *North American Actuarial Journal*, *2*, 1–25.

Garthwaite, P. H., Al-Awadhi, S. A., Elfadaly, F. G., & Jenkinson, D. J. (2013). Prior distribution elicitation for generalized linear and piecewise-linear models. *Journal of Applied Statistics*, *40*, 59–75.

Garthwaite, P. H., Kadane, J. B., & O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, *100*, 680–701.

Gosling, J. P. (2018). SHELF: The sheffield elicitation framework. In L. C. Dias, A. Morton, & J. Quigley (Eds.), *Elicitation: The science and art of structuring judgement*. New York: Springer.

Joe, H. (1997). *Multivariate models and dependence concepts*. London: Chapman & Hall.

Jouini, M. N., & Clemen, R. T. (1996). Copula models for aggregating expert opinions. *Operations Research*, *44*, 444–457.

Kadane, J. B., Dickey, J. M., Winkler, R., Smith, W., & Peters, S. (1980). Interactive elicitation of opinion for a normal linear model. *Journal of the American Statistical Association*, *75*, 845–854.

Nelsen, R. B. (1999). Lecture notes in statistics. *An introduction to copulas* (Vol. 139). New York: Springer-Verlag.

Oakley, J.E. (2017). *SHELF: Tools to support the sheffield elicitation framework*. R package version 1.3.0. https://github.com/OakleyJ/SHELF.

Oakley, J.E., O'Hagan, A. (2010). *SHELF: The sheffield elicitation framework (version 3.0)*. School of Mathematics and Statistics, University of Sheffield. http://tonyohagan.co.uk/shelf.

Oakley. J.E. (2010). Eliciting univariate probability distributions. In K. Böcker (Ed.), *Rethinking risk measurement and reporting: Volume I*. London: Risk Books

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., et al. (2006). *Uncertain judgements: Eliciting expert probabilities*. Chichester: John Wiley.

O'Hagan, A., & Forster. J. (2004). B*ayesian Inference, volume 2B of Kendall's Advanced Theory of Statistics* (2nd ed.). London: Arnold.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

van Dorp, J. R., & Mazzuchi, T. A. (2004). Parameter specification of the beta distribution and its Dirichlet extensions utilizing quantiles. In A. K. Gupta & S. Nadarajah (Eds.), *Handbook of beta distribution and its applications*. New York: Marcel Dekker Inc.

Werner, Christoph, Bedford, Tim, Cooke, Roger M., Hanea, Anca M., & Morales-Napoles, Oswaldo. (2017). Expert judgement for dependence in probabilistic modelling: A systematic literature review and future research directions. *European Journal of Operational Research*, *258*(3), 801–819.

Wilson, K. J. (2018). Specification of informative prior distributions for multinomial models using vine copulas. *Bayesian Analysis*, *13*, 749–766.

Zapata-Vázquez, R. E., O'Hagan, A., & Bastos, L. S. (2014). Eliciting expert judgements about a set of proportions. *Journal of Applied Statistics*, *41*(9), 1919–1933.

Chapter 3
Are Performance Weights Beneficial? Investigating the Random Expert Hypothesis

Deniz Marti, Thomas A. Mazzuchi, and Roger M. Cooke

Abstract Expert elicitation plays a prominent role in fields where the data are scarce. As consulting multiple experts is critical in expert elicitation practices, combining various expert opinions is an important topic. In the Classical Model, uncertainty distributions for the variables of interest are based on an aggregation of elicited expert percentiles. Aggregation of these expert distributions is accomplished using linear opinion pooling relying on performance-based weights that are assigned to each expert. According to the Classical Model, each expert receives a weight that is a combination of the expert's statistical accuracy and informativeness for a set of questions, the values of which are unknown at the time the elicitation was conducted. The former measures "correspondence with reality," a measure of discrepancy between the observed relative frequencies of seed variables' values falling within the elicited percentile values and the expected probability based on the percentiles specified in the elicitation. The later gauges an expert's ability to concentrate high probability mass in small interquartile intervals. Some critics argue that this performance-based model fails to outperform the models that assign experts equal weights. Their argument implies that any observed difference in expert performance is just due to random fluctuations and is not a persistent property of an expert. Experts should therefore be treated equally and equally weighted. However, if differences in experts' performances are due to random fluctuations, then hypothetical experts created by randomly recombining the experts' assessments should perform statistically as well as the actual experts. This hypothesis is called the *random expert hypothesis*. This hypothesis is investigated using 44 post-2006 professional expert elicitation studies obtained through the TU Delft database. For each study, 1000 hypothetical expert panels are simulated whose elicitations are a random mix of all expert elicitations within that study. Results indicate that actual expert statistical accuracy performance is significantly better than that of randomly created experts. The study does not consider

D. Marti (✉) · T. A. Mazzuchi
The George Washington University, Washington, USA
e-mail: hazanmarti@gwu.edu

R. M. Cooke
Resources for the Future, Washington, USA

A. M. Hanea et al. (eds.), *Expert Judgement in Risk and Decision Analysis*,
International Series in Operations Research & Management Science 293,
https://doi.org/10.1007/978-3-030-46474-5_3

experts' informativeness but still provides strong support for performance-based weighting as in the Classical Model.

3.1 Introduction

Expert elicitation can play a prominent role in the decision-making process in risk assessment, system safety, reliability, and many other fields, particularly in fields where it is difficult to obtain data input (e.g., Einhorn 1974; Cooke and Goossens 2008; Mosleh et al. 1988; Singpurwalla 1988; Spetzler and von Holstein 1975; Wallsten and Budescu 1983; Otway and von Winterfeldt 1992; Aspinall 2010; Chap. 10, this volume). Disciplines that involve high levels of uncertainty combined with insufficient data include, but not are limited to, disaster management, epidemiology, intelligence, public and global health, environment, and security, all of which require robust probabilistic assessments (e.g., Ryan et al. 2012; Keeney and Von Winterfeldt 1989; Hald et al. 2016). In such fields, there might be cost and time considerations, as well as technical impracticalities to data collection, which result in limited scientific data. Sometimes, it is not practical to collect data due to the nature of events. Ultimately, absent or insufficient data lead to poor risk assessments and judgment, resulting in failure either to make informed decisions or to design reliable decision-making processes. Thus, in order to properly characterize the uncertainty in such fields, experts' inputs play a vital role (Cooke and Goossens 2008; Otway and von Winterfeldt 1992). Experts, in the absence of empirical data, are requested to provide information, which could be elicited in various forms such as probability elicitation, parameter estimation, and quantity estimation (Clemen and Winkler 1999). These forms of expert elicitations are essential for uncertainty characterization and risk and policy models.

The standard expert elicitation practice is to consult with multiple experts. Clemen and Winkler (1999) note that the reason for consulting multiple experts is to collect as much data as possible, which could be considered the same as a motivation to increase the sample size in an experiment. This raises the concern on how to fully encapsulate diverse expert judgments in a single input for the analysis. Morgan et al. (1992) noted that factors that lead to combining expert opinions must be chosen so that experts' knowledge can be optimally reflected in the ultimate outcome. Thus, the natural question that arises is "How should one combine multiple opinions?" While a significant body of literature has addressed this issue (see for review Ouchi 2004; Clemen 1989; Morgan et al. 1992), perhaps among the proposed methods, opinion pooling has been the most commonly used approach. Stone (1961) initially coined a strategy for combining opinions: opinion pooling, which was later substantially reviewed by other scholars (see for example, French 1981; Genest and Zidek 1986).

The linear opinion pool is a very practical and straightforward axiomatic method. It is, in fact, a weighted average of multiple probability distributions

$$f(\theta) = \sum_{i=1}^{n} w_i f_i(\theta) \tag{3.1}$$

Here, θ is the unknown quantity of interest,$f_i(\theta)$ is the density function of expert$_i$, w_i represents the weight assigned to expert$_i$, and n is the number of experts. The combined distribution is represented by $f(\theta)$, referred to as the decision maker's probability distribution. Each weight can be interpreted as the expert's relative contribution. If the decision maker has little evidence to judge experts' weights, then each weight is simply distributed equally to the experts, that is $w_i = 1/n$. This approach is called Equal Weighting (EW), and treats each expert equally. However, this approach does not give the decision maker the power to optimize the use of experts' opinions. The underlying assumption of using equal weights is that experts contribute equally. A pre-commitment to EW usually implies that experts' performance will not be measured at all. Consequently, the EW decision maker's performance cannot be validated. This potentially compromises the impact of expert judgment in science-based decision making.

The most prevalent approach addressing this concern is the Classical Model (Cooke 1991), which suggests a weighting mechanism that is based on experts' performances, rather than weighting experts equally. Some scholars argue that performance-based weighting does not outperform equal weighting in terms of the proposed performance criteria. Clemen (2008) provided the most thorough critique of the Classical Model. His results were based on a small sample of expert studies and thus were inconclusive. However, his work advanced the debate and motivated subsequent studies (e.g., Eggstaff et al. 2014; Colson and Cooke 2017), which eventually demonstrated the out-of-sample superiority of the Classical Model's performance-based approach, relative to equal weighting. Following on this work, this chapter seeks to evaluate the appropriateness of the *random expert hypothesis.* This work is novel in the sense of testing the fundamental premises of two aggregation approaches, EW and PW. Simply stated, this hypothesis investigates the claim that any expert's performance in performance-based weighting is due to chance. In this chapter, the random expert hypothesis will be evaluated with respect to the statistical accuracy measures for 44 most up-to-date datasets from the TU Delft database (Cooke and Goossens 2008).

3.1.1 Classical Model

The Classical Model is grounded in the argument that experts differ in terms of their performances—that is, in their ability to assess uncertainty and communicate it properly. Therefore, their performances should be quantified and then reflected in the weighting framework. The model addresses the naturally arising question of how experts' performance can be measured. The model proposes that the Decision Maker's distribution, (1), is obtained via performance-based weights whose values are determined by the aforementioned measures of experts' statistical accuracy and informativeness. The performances on these two criteria are assessed via an elicitation procedure using predetermined seed variables whose exact values are known by the analyst.

The elicitation procedure involves requesting experts to provide their inputs for a predetermined number, say N, of seed variables, the values of which are usually known post hoc (Cooke 1991). The common practice is to ask experts for their estimates of 5th, 50th, and 95th quantiles for seed variables though other percentile could be used as well. The two performance criteria are measured using these elicitations and the true realizations of the seed variables. The specified percentiles reflect the experts' judgments about this unknown quantity in terms of specified statistical bins. For example, by specifying an elicitation for the 5th percentile, $q_{5,}$ the expert considers that the probability that the true realization of the seed variable is smaller than q_5 is 0.05. Similarly, the 50th percentile, $q_{50,}$ suggests that the expert believes that there is 50% probability of observing the true value to be less than q_{50}, etc.

In addition to these assessed percentiles, the analyst specifies an overshoot percentage (commonly 10%, see Cooke 1991 for more details) in order to determine the complete support for the experts' distributions. Once elicitations are compiled, the analyst assesses the experts' performances using the true realizations of the seed variables. Specifically, the analyst determines how informative the expert distributions are relative to a minimally informative distribution on the support and how well the expert's uncertainty assessments via the specified percentile values match with the realization of the seed variables (i.e., statistical accuracy).

(1) Informativeness

Informativeness score gauges the additional contribution of the expert's elicitation relative to a background measure. That is, it answers the question of "does the expert provide any additional information than a minimally informative distribution?" To measure experts' performance with respect to this criterion, the analyst first combines the expert opinions for each seed variable into a single range, the lower and upper bounds of which are determined by, respectively, the minimum and the maximum of elicited values for each seed variable and the realization of these variables. Then, by using a 10% overshoot percentage, the entire cumulative distributions are computed for each expert. These elicited distributions for each expert are compared with a minimally informative background measure, usually the uniform distribution, which expresses complete uncertainty over the range. The more additional information an expert's distribution gives relative to the base knowledge, the higher the *information* scores he or she would receive.

(2) Statistical Accuracy

Statistical accuracy (a.k.a., calibration score) is a measure of the extent to which the expert's quantile assessment matches with reality. Cooke (1991) incorporated this idea into the model by using a hypothesis test. The null hypothesis is that the experts' percentile assessments correspond to reality. The p value associated with this hypothesis constitutes the statistical accuracy score. That is, lower p value indicates less evidence about the experts' statistical accuracy performance. Following the computations below, the analyst determines the frequency of true realizations' occurrence in specified inter-quantile intervals, bounded by the specified quantiles.

3.1.2 The Debate on Aggregating Expert Elicitations Mechanisms: Performance-Based Weights (PW) Versus Equal Weights (EW)

The debate around how to aggregate expert elicitations revolve around two fundamental approaches: combining expert elicitations based on equal weights (EW) or based on their performance-based weights (PW). The Classical Model uses a performance-based approach. The model's main premise suggests that performance-based weighting mechanism ensures higher quality and improves the task for which the expert elicitation is done. There is a substantial body of knowledge that supports the use of performance weights (e.g., Aspinall et al. 2016; Bamber and Aspinall 2013; Colson and Cooke 2017; Wilson 2017). However, others have advocated the use of equal weights to combine expert elicitations (Clemen and Winkler 1999; Clemen 1989). They argued that equal weights perform as well as performance-based weights; therefore, there is no need to undertake an intensive expert elicitation procedure (e.g., Clemen 2008). Some of these critics failed to provide substantial evidence and details of their research procedure (e.g., replicable codes), so their findings are not considered to be conclusive.

Perhaps, among the EW advocates, the most productive contribution was Clemen (2008) who critiqued the Classical Model implementations for solely depending on in-sample validation. He argued that the concern about this validation technique was that it uses the dataset to determine the performances and also to validate the model. He suggested using out-of-sample validation and compared EW and PW. Specifically, Clemen (2008) performed a remove-one-at-a-time (ROAT) method, whereby seed variables are removed one at a time. Performance weights are computed based on the remaining seed variables, and these weights are used to predict the removed item. He found that PW failed to statistically outperform EW (PW outperformed EW in 9 out of 14 studies). Two concerns were raised about these findings: One, Clemen (2008) used a nonrandom sample and failed to justify his data choices. Second, the ROAT approach leads to systematic biases, whereby each removed item can penalize an expert who did poorly on that particular item (Cooke and Goossens 2008; Colson and Cooke 2017). This bias was addressed (Colson and Cooke 2017) by a more substantial approach, the cross-validation technique that uses a certain percentage of dataset, instead of a single seed variable. The dataset is split into a training set to determine the performance weights and a test set to predict the removed items. Eggstaff et al. (2014) performed an extensive cross-validation analysis on all possible sets of training and test variables and found that PW statistically outperforms EW.

These examples of previous studies confirmed the validity of the Classical Model; nonetheless, the debate continues. The model has been validated in studies that include different number of seed variables and experts (e.g., Tyshenko et al. 2011; Jaiswal et al. 2012; Bamber et al. 2016; Aspinall 2010; Aspinall et al. 2016). The debate so far focused on the validity of the Classical Model, in different validation approaches (i.e., in-sample, ROAT, and cross validation). However, it is also necessary to analyze the fundamental assumptions of the two competing approaches. No previous studies have tested the core distinction between the two camps of the debate: do the differences in performance reflect persistent differences in the experts, or are they an artifact caused by random influences introduced by the elicitation itself? For example, if the difference is due to one expert having a good day, or being influenced by domestic or professional stressors, or having more information about particular seed variable, etc., then the equal weighting scheme may be warranted. The EW approach assumes that any apparent differences in expert performance are due to such random influences and would not persist beyond the particular elicitation context. On the other hand, the PW approach suggests that performance differences reflect "properties of the experts," which persist beyond particular elicitation context. Focusing on the fundamental assumption that performance differences are persistent enables the formulation of this assumption as a testable statistical hypothesis termed the *Random Expert Hypothesis (REH)*: apparent differences in expert performance are due to random stressors affecting the elicitation.

3.2 Random Expert Hypothesis (REH)

The REH states that apparent differences in expert performance are due to random stressors of the elicitation. If this hypothesis were true, then randomly reallocating the assessments among the experts should have no effect on the performance of the expert panel. This "random scrambling" is precisely defined below. Under the REH, the scores of the best and worst performing experts in the original panel should be statistically indistinguishable from those of the best and worst experts after scrambling the assessments. The variation in expert scores in the original panel should be statistically indistinguishable from the variation in the scrambled panels. There are many ways of scrambling the experts' assessments and this allows a determination of the distributions of scores that result from randomly redistributing the stressors over the experts.

Note that random scrambling will have no effect on the EW combination. This underscores the fact that EW *implies* the REH. In consequence (modus tollens), if REH is (statistically) rejected, then so is EW. In this sense, REH provides a more powerful test of the assumption underlying the use of EW. Note also that if all experts in a panel are "equally good" or "equally bad," then the REH may actually be true for that panel. Indeed, this sometimes happens. The use of PW depends on the fact that such panels are in the minority. Testing the REH on a set of cases allows for gauging the size of that minority.

The REH was tested by a process of creating random panels of experts whose elicitations are derived from the experts within the original expert panel. For example, suppose an expert judgment panel includes ten experts, each of whom assessed 5th, 50th, and 95th percentiles for each seed variable. A hypothetical expert judgment panel would have ten randomly created experts, each of whose elicitations are randomly drawn without replacement from the original assessments for each variable. This process is repeated 1000 times. If there is not a systematic difference between randomly created experts and the original experts, as the REH implies, then one would expect that in approximately half of those 1000 runs, the original experts would outperform the random experts. The performance measure used in this study is statistical accuracy; informativeness and full performance weights will be considered in a future study.

Figure 3.1 displays the process of random expert creation for three experts and three seed variables. For example, Random Expert 1 takes the assessment of Original Expert 2 for Seed Variable #1, the assessment of Original Expert 1 for Seed Variable #2, and finally the assessment of Original Expert 3 for Seed Variable #3. Random Expert 2 chooses randomly from the remaining experts, and Random Expert 3 gets the remaining elicitations. Ultimately, a hypothetical expert judgment panel is composed by creating as many scrambled random experts as in the original experts.

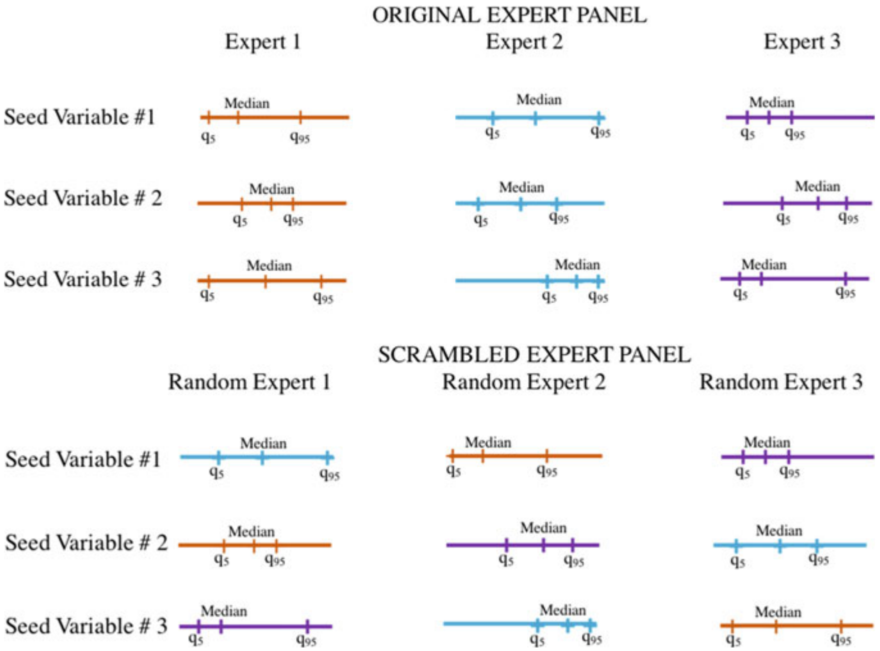


Fig. 3.1 An illustration of random expert creation process. q_5 corresponds to the 5th percentile elicitation, the median corresponds to the 50th percentile elicitation, and q_{95} corresponds to the 95th percentile elicitation

The Classical Model assumes that expert performances in terms of statistical accuracy may vary systematically with respect to the persistent differences. Specifically, such persistent differences are the reasons why the “best performing expert” performs the best and the “poorest performing expert” performs the poorest. However, when the elicitations are scrambled for a very large number of runs, then scrambled experts should perform the same, statistically. That is, the “best performing expert” does no longer perform as much better than the other experts; the “poorest performing expert” does no longer perform as much poorly than the rest. In other words, the scrambling process eliminates the systematic variation, which implies a smaller standard deviation.

If REH is false, then the original expert panel should look statistically different from the population of scrambled panels. The systematic differences among experts, as posited by the Classical Model, lead to a larger average score and smaller standard deviation of the score. The maximum score (i.e., the best performing expert’s score) of the original panel is expected to be higher than that of the scrambled panel. Similarly, the minimum score (i.e., the poorest performing expert’s score) of the original panel is expected to be lower than that of the scrambled panel.

There are a number of ways in which a test could be constructed to examine whether the original expert panel comes from the same distribution as the scrambled panels. In this study, four tests of REH are identified. Specifically, if REH were true, then

(1) The probability is 50% that the average of the experts’ statistical accuracies in the original panel is higher than that of a scrambled panel
(2) The probability is 50% that the standard deviation of the experts’ statistical accuracies in the original panel is higher than that of a scrambled panel
(3) The probability is 50% that the maximum of the experts’ statistical accuracies in the original panel is higher than that of a scrambled panel
(4) The probability is 50% that the minimum of the experts’ statistical accuracies in the original panel is lower than that of a scrambled panel.

These predictions of REH were tested based on experts’ statistical accuracy performances measured by the statistical accuracy score. The statistical accuracy score is the focus since it is the main characteristics of the performance-based weights (see the Cooke 1991 for discussion), while the information score has a role of modulating the statistical accuracy score.

3.3 Expert Judgment Data

TU Delft database provides extensive datasets of expert elicitations that were conducted based on Classical Model framework (Cooke and Goossens 2008). This database has been recently updated with new studies that were performed starting from 2006 to 2015 (Colson and Cooke 2017). As summarized by Colson and Cooke (2017), these studies were done by organizations such as Bristol University, the British Government, United States Department of Homeland Security, World Health Organization, and the US Environmental Protection Agency, etc. Studies were performed in two formats of structured expert judgment, in three percentiles and five percentiles. Experts are asked to elicit the seed variables for 5th, 50th, and 95th percentiles in the former format, and 5th, 25th, 50th, 75th, and 95th percentiles in the latter format. These elicitations were compiled by Cooke and Goosens (2008) and made available to the researchers and recently updated (available at http://rogermcooke.net/). This study focuses on all 44 datasets that are available in the new post-2006 expert judgment database. 27 of these 44 datasets came from studies, which were performed in three-percentile format, and 17 of the 44 datasets were performed in five-percentile format: experts were asked to provide five percentiles for the elicited variables.

Table 3.1 summarizes the names, the percentile format, number of experts, number of seed variables, and associated references for each expert judgment panel. The studies are across wide range of domains such as environmental risk, bioterrorism, air traffic control, and volcano eruptions. The number of experts in the panels of these studies ranged from 4 to 21, and the number of seed variables ranged from 8 to 48. The three-percentile format data has 298 experts who elicited 386 seed variables in total, which yielded 4597 elicitations in total. The five-percentile format data has 111 experts who elicited 170 seed variables in total, which yielded a total of 1117 elicitations.

3.4 Hypothesis Testing

44 studies presented in Table 3.1 are used to test the random expert hypothesis. For each study, hypothetical expert judgment panels consisting of randomly scrambled experts are simulated in 1000 runs. The extent to which this data support the REH can be statistically examined by a Binomial test for each of the four statistical metrics, namely, average, standard deviation, maximum, and the minimum scores of expert panels for each study.

$$\mathrm{H_0 : r = 0.5}$$

$$\mathrm{H_a : r > (<)0.5}$$

where *r* is the percentage of the studies in which the original experts outperform the random experts.

r is the success probability in which the success, “outperformance,” is defined as follows:

1. The average statistical accuracy score of the original expert panels is higher than that of a scrambled expert panel

Table 3.1 Expert judgment studies are illustrated with the number of seed variables and experts, and percentile formats

Study	Percentile format	# of experts	# of seed variables	Subject
UMD	3	9	11	Nitrogen removal in Chesapeake Bay
USGS	3	18	32	Volcanos
arsenic	3	9	10	Air quality levels for arsenic
Biol Agents	3	9	10	Human dose–response curves for bioterror agents
Geopolit	3	9	16	Geopolitics
ATCEP	3	5	10	Air traffic controllers human error
Daniela	3	4	10	Fire prevention and control
eBBP	3	14	15	XMRV blood/tissue infection transmission risks
create	3	7	10	Terrorism
effErupt	3	14	8	Icelandic fissure eruptions: source characterization
erie	3	10	15	Establishment of Asian Carp in Lake Erie
FCEP	3	5	8	Flight crew human error
Sheep	3	14	15	Risk management policy for sheep scab control
Hemophilia	3	18	8	Hemophilia
Liander	3	11	10	Underground cast iron gas-lines
PHAC	3	10	12	Additional CWD factors
TOPAZ	3	21	16	Tectonic hazards for radwaste siting in Japan
SPEED	3	14	16	Volcano hazards (Vesuvius and Campi Flegrei, Italy)
TDC	3	18	17	Volcano hazards (Tristan da Cunha)

(continued)

Table 3.1 (continued)

Study	Percentile format	# of experts	# of seed variables	Subject
GL	3	9	13	Costs of invasive species in Great Lakes
Goodheart	3	5	10	Airport safety
Ice	3	10	11	Sea level rise from ice sheets melting due to global warming
puig-gdp	3	9	13	Emission forecasts from Mexico
puig-oil	3	6	19	Oil emissions and prices
YTBID (CDC)	3	14	48	Return on investment for CDC warnings
Gerestenberger	3	12	13	Probabilistic seismic-hazard model for canterbury
CWD	3	14	10	Infection transmission risks: Chronic wasting disease from deer to humans
Nebraska	5	4	10	Grant effectiveness, child health insurance enrollment
San Diego	5	7	10	Effectiveness of surgical procedures
BFIQ	5	7	11	Breastfeeding and IQ
France	5	5	10	Future antimicrobial resistance in France
Italy	5	4	8	Future antimicrobial resistance in Italy
Spain	5	5	10	Future antimicrobial resistance in Spain
UK	5	6	10	Future antimicrobial resistance in UK
Arkansas	5	4	10	Grant effectiveness, child health insurance enrollment
CoveringKids	5	5	10	Grant effectiveness, child health insurance enrollment
dcpn_Fistula	5	8	10	Effectiveness of obstetric fistula repair

(continued)

Table 3.1 (continued)

Study	Percentile format	# of experts	# of seed variables	Subject
Florida	5	7	10	Grant effectiveness, child health insurance enrollment
Illinois	5	5	10	Grant effectiveness, child health insurance enrollment
Obesity	5	4	10	Grant effectiveness, childhood obesity
Tobacco	5	7	10	Grant effectiveness, childhood obesity
Washington	5	5	10	Grant effectiveness, child health insurance enrollment
cdc-roi	5	20	10	Return on investment for CDC warnings
IQ-earn	5	8	11	Effects of increases in IQ in India on the present value of Lifetime earnings

Note The references to the data can be found in the Appendix

2. The standard deviation of statistical accuracy scores of the original expert panels is higher than that of a scrambled expert panel
3. The maximum statistical accuracy scores of the original expert panels is higher than that of a scrambled expert panel
4. The minimum statistical accuracy scores of the original expert panels is lower than that of a scrambled expert panel.

3.5 Results

The data were analyzed in two different formats: (1) in three-percentile format data, including all 44 available datasets (thus five-percentile datasets were converted to three-percentile datasets), (2) in five-percentile format data, including only five-percentile elicitations.

3.5.1 *The Analysis of the Three-Percentile Format Data*

The average, standard deviation, the maximum, and the minimum scores of the original experts are compared with those of the random experts in each randomly created 1000 expert panels. The Binomial tests are performed for all 44 datasets available in three-percentile format.

The statistical accuracy scores were computed for the three-percentile format data, consisting of 27 studies that were originally performed as a three-percentile format, and 17 five-percentile studies that were converted to three-percentile format. Table 3.2 provides the statistical summaries of the original experts' statistical accuracy scores: summaries, average, standard deviation, maximum, and minimum scores of expert panels.

Then, four statistical metrics—average, standard deviation, maximum, and minimum of the statistical accuracy scores—were computed for the original expert panels and for each of the 1000 scrambled expert panels. Then, for each expert panel, the percentage that the original experts' corresponding statistics ranked higher than (lower for the minimum) those of the 1000 scrambled expert panels was determined. Under the REH, the original expert panels' metrics should be ranked above (below) those of the scrambled expert panels 50% of the time. Table 3.3 illustrates the actual percentages determined for each dataset.

For example, in Table 3.3, the corresponding percentage for the average score in the study UMD is shown as 99.7%. This indicates that in 99.7% of the scrambled panels (997 out of 1000 simulation runs), the average scores of the original experts are greater than those of the randomly scrambled experts. Similarly, in the UMD study, in 96.4% of the scrambled panels (964 out of 1000 simulation runs), the standard deviation of the experts in the original panel are greater than those of scrambled experts, indicating a larger variation in the original expert score in most cases. The best performing expert in UMD study outperforms the best performing expert of the scrambled panels in 95.4% of the time. This means that, in 954 out of 1000 simulation runs randomly created expert panels, the best performing experts are outperformed by the original best performing expert. Finally, 100% for the minimum score displayed in Table 3.3 shows that the minimum score of the original expert panel was lower than those of all scrambled panels, indicating that the score of the poorest performing expert of the original panel performed the poorest compared to all random experts.

Figure 3.2 shows that in 16 out of 44 studies, the original experts outperformed more than 95% (i.e., 950 out of 1000 simulation runs) of the scrambled expert panels. Similarly, in 5 studies, the original experts outperformed the scrambled experts in 85–95% of the time. In total, in 33 out of 44 studies, the original experts' average scores ranked higher than those of the scrambled experts from 1000 expert panels at least 80% of the time.

Figure 3.3 shows that, in 10 studies, the standard deviation of the experts' statistical accuracy scores in the original panel is larger than those in more than 95% of the 1000 randomly created expert panels. In 28 out of 44 studies, the variation in the original expert data is larger than the variation in the scrambled expert panels at least 80% of the time.

Figure 3.4 shows that, in 10 studies, the best performing expert in the original expert panel outperforms the best performing expert in the random expert panels in more than 95% of the time. In 26 out of 44 studies, the best performing original

Table 3.2 Statistical accuracy scores of the original experts for the three-percentile format elicitation data

Study No.	Study name	Average	Standard deviation	Max	Min
1	UMD	1.33E-01	2.69E-01	7.06E-01	3.21E-14
2	USGS	3.15E-03	1.17E-02	5.55E-02	7.12E-13
3	arsenic	4.84E-03	1.18E-02	3.57E-02	9.86E-07
4	Biol Agents	5.70E-02	1.16E-01	3.11E-01	1.42E-06
5	Geopolit	5.10E-02	1.01E-01	2.30E-01	1.42E-06
6	ATCEP	2.99E-02	4.47E-02	1.01E-01	1.42E-06
7	Daniela	1.88E-01	2.57E-01	5.54E-01	4.35E-07
8	eBBP	2.00E-01	2.63E-01	8.33E-01	8.91E-06
9	create	3.57E-03	6.37E-03	1.71E-02	8.91E-06
10	effErupt	2.91E-02	5.46E-02	1.85E-01	8.91E-06
11	erie	2.27E-01	2.46E-01	6.61E-01	1.08E-08
12	FCEP	1.75E-01	2.84E-01	6.64E-01	5.12E-05
13	Sheep	5.64E-02	1.70E-01	6.43E-01	1.62E-11
14	hemophilia	1.88E-01	2.28E-01	6.64E-01	2.66E-04
15	Liander	3.18E-04	8.37E-04	2.81E-03	3.50E-08
16	PHAC	9.71E-03	2.46E-05	7.50E-05	2.43E-10
17	TOPAZ	3.08E-02	1.00E-01	2.43E-10	4.42E-12
18	SPEED	1.83E-02	6.03E-02	2.27E-01	2.88E-12
19	TDC	1.03E-01	2.72E-01	9.89E-01	1.02E-12
20	GL	6.13E-02	1.51E-01	4.54E-01	1.91E-09
21	Goodheart	1.47E-01	2.76E-01	7.07E-01	7.99E-04
22	Ice	8.53E-02	1.50E-01	3.99E-01	5.84E-06
23	puig-gdp	3.68E-02	9.16E-02	2.77E-01	5.04E-12
24	puig-oil	1.72E-03	4.17E-03	1.02E-02	3.27E-12
25	YTBID (CDC)	1.43E-01	2.23E-01	9.68E-01	5.80E-07
26	Gerestenberger	6.35E-02	6.29E-02	1.52E-01	1.88E-05
27	CWD	7.62E-02	1.47E-01	4.93E-01	1.07E-06
28	Nebraska	1.89E-03	3.71E-03	7.46E-03	4.54E-10
29	San Diego	3.45E-04	5.91E-04	1.31E-03	8.36E-11
30	BFIQ	1.24E-01	2.33E-01	6.38E-01	2.28E-04
31	France	1.56E-01	3.09E-01	7.07E-01	1.54E-07
32	Italy	1.70E-01	3.14E-01	6.40E-01	5.86E-07
33	Spain	4.70E-06	9.07E-06	2.08E-05	1.29E-10
34	UK	1.49E-01	2.72E-01	6.83E-01	6.17E-09
35	Arkansas	8.00E-02	1.56E-01	3.14E-01	1.07E-06
36	CoveringKids	2.76E-01	3.02E-01	6.83E-01	9.86E-07

(continued)

Table 3.2 (continued)

Study No.	Study name	Average	Standard deviation	Max	Min
37	dcpn_Fistula	6.54E-04	1.13E-03	2.81E-03	9.86E-07
38	Florida	2.24E-02	2.36E-02	4.70E-02	5.21E-06
39	Illinois	1.75E-02	3.23E-02	7.50E-02	5.45E-08
40	Obesity	6.67E-02	9.06E-02	1.92E-01	2.47E-10
41	Tobacco	2.06E-01	2.39E-01	6.83E-01	5.99E-03
42	Washington	6.29E-02	1.04E-01	2.44E-01	5.99E-04
43	cdc-roi	1.08E-01	1.46E-01	4.93E-01	3.50E-08
44	IQ-earn	6.88E-02	1.26E-01	3.70E-01	1.70E-07

Note First 27 datasets were expert elicitations based on three-percentile format (5th, 50th, and 95th percentiles) and last 17 studies were converted into three-percentile format by truncating the 25th and the 75th percentiles)

Table 3.3 The percentage of original experts' corresponding statistics ranked higher than (lower for the minimum) those of the 1000 randomly created expert panels (the entire available data in three-percentile format)

Study No.	Study name	Average (%)	Standard deviation (%)	Max (%)	Min (%)
1	UMD	99.70	96.40	95.40	100.00
2	USGS	86.60	84.50	79.40	80.10
3	arsenic	57.80	60.80	56.50	43.40
4	Biol Agents	84.20	73.10	60.30	69.80
5	Geopolit	87.20	82.70	76.30	54.80
6	ATCEP	95.80	94.70	93.90	99.50
7	Daniela	91.90	64.70	63.60	99.70
8	eBBP	99.10	91.40	83.30	88.30
9	create	23.00	34.70	20.80	13.10
10	effErupt	85.90	80.50	54.00	88.80
11	erie	87.10	71.10	75.00	100.00
12	FCEP	93.30	84.50	85.00	92.00
13	Sheep	98.80	97.70	97.80	99.20
14	hemophilia	90.40	77.30	24.20	34.70
15	Liander	21.20	26.50	25.30	36.40
16	PHAC	56.60	48.70	22.50	99.70
17	TOPAZ	98.00	98.00	98.00	8.00
18	SPEED	97.90	97.50	97.50	97.60
19	TDC	100.00	100.00	97.50	99.10
20	GL	100.00	99.40	99.10	98.80
21	Goodheart	82.70	83.90	83.10	34.70
22	Ice	95.00	91.50	82.10	55.00

(continued)

Table 3.3 (continued)

Study No.	Study name	Average (%)	Standard deviation (%)	Max (%)	Min (%)
23	puig-gdp	96.70	96.40	96.30	99.30
24	puig-oil	97.20	97.20	97.20	73.60
25	YTBID (CDC)	97.90	94.20	80.30	88.70
26	Gerestenberger	6.35	6.29	15.19	73.80
27	CWD	82.90	79.80	78.40	71.50
28	Nebraska	76.80	78.70	78.70	97.80
29	San Diego	91.10	90.40	85.10	74.10
30	BFIQ	80.10	90.40	80.60	48.30
31	France	99.80	98.90	98.90	97.20
32	Italy	80.50	85.80	81.60	99.70
33	Spain	59.40	40.80	35.70	88.30
34	UK	96.30	89.10	88.50	99.90
35	Arkansas	97.90	96.20	95.20	81.90
36	CoveringKids	96.00	85.00	62.90	98.70
37	dcpn_Fistula	11.60	14.90	9.30	17.10
38	Florida	54.20	33.70	14.70	60.30
39	Illinois	79.40	72.80	72.70	76.40
40	Obesity	93.10	90.50	90.50	99.90
41	Tobacco	40.40	49.30	36.30	58.10
42	Washington	26.70	40.80	37.70	50.90
43	cdc-roi	94.80	61.10	58.10	91.60
44	IQ-earn	2.30	16.60	26.40	99.60

expert outperformed the best performing random expert in at least 80% of the 1000 scrambled expert panels.

Figure 3.5 shows that, in 19 studies, the poorest performing expert in the original expert panel performed poorer than the poorest performing expert in the randomly created expert panels. In 26 out of 44 studies, the original experts' minimum score is lower than the random expert panel's minimum score in at least 80% of the 1000 expert panels.

In above results, the percentage score may be a function of the number of experts and the corresponding spread in the calibration scores of the experts. The exact determination is a subject of future research. However, the empirical results from above suggest that the REH may not be appropriate. To more formally test the REH, a statistical test is needed. The test selected was the Binomial test due to its appropriateness for dichotomous outcomes and its nonparametric nature.

The Binomial test results show that the average of the statistical accuracy results of the original experts outperformed the randomly created experts more than 50% of the time, in three statistical metrics: on average ($p = 2.65\text{E-}06$), on standard deviation

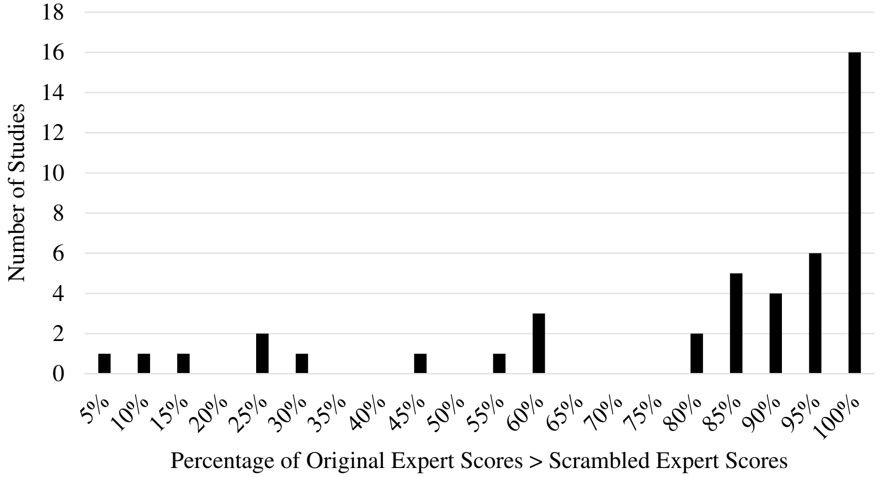


Fig. 3.2 Distribution of percentage of original experts' average statistical accuracy scores ranked higher among those of scrambled experts in 1000 hypothetical expert panels based on 44 studies

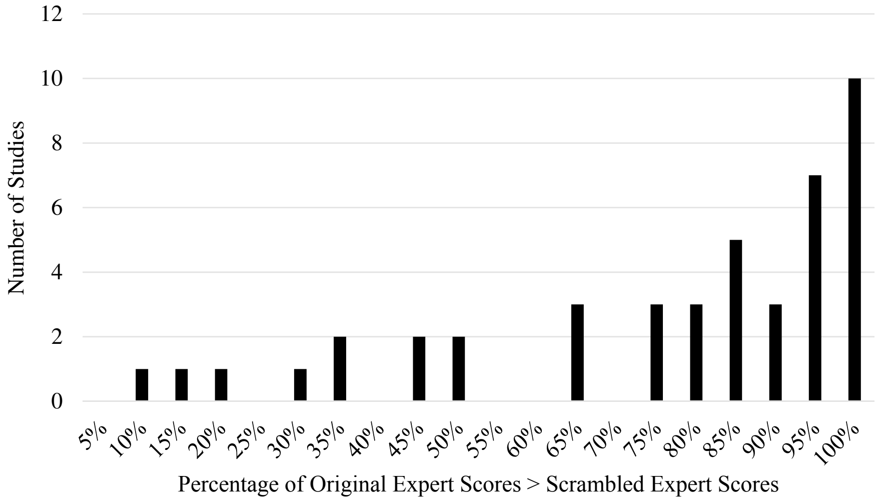


Fig. 3.3 Distribution of percentage of the standard deviation of the original experts' statistical accuracy scores ranked higher among those of scrambled experts in 1000 hypothetical expert panels based on 44 studies

(p = 1.94E-04), and on maximum scores (p = 6.3E-04). Also, the minimum of the original experts performed significantly poorer than the poorest performing randomly created experts more than 50% of the time (p = 1.27E-05).

Overall, the results of the random expert hypothesis testing show that, in a significant number of studies, the scrambled experts fail to perform as well as the original

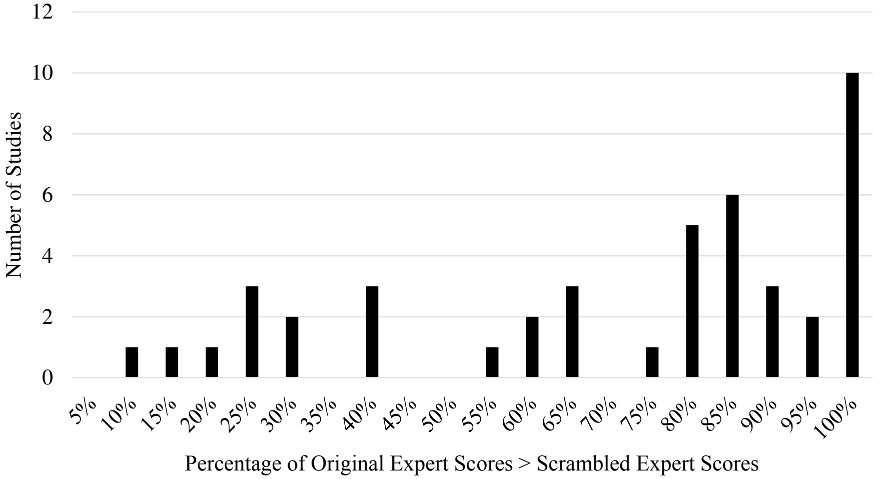


Fig. 3.4 Distribution of percentage of the maximum of the original experts' statistical accuracy scores ranked higher among those of scrambled experts in 1000 hypothetical expert panels based on 44 studies

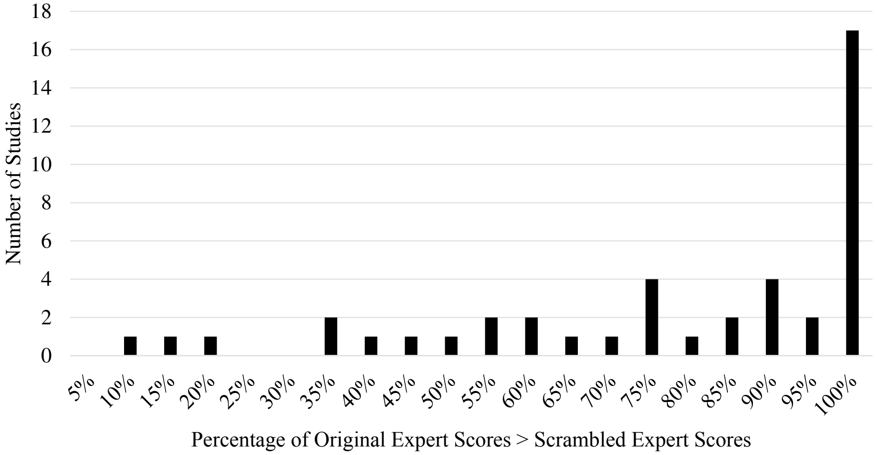


Fig. 3.5 Distribution of percentage of the minimum of the original experts' statistical accuracy scores ranked lower among those of scrambled experts in 1000 hypothetical expert panels based on 44 studies

experts. Specifically, in most studies, the original experts outperformed the scrambled experts in an overwhelmingly large percentage of the hypothetical expert panels. Binomial test results suggest that the original experts ranked higher than the scrambled experts in three statistical summaries, the average, standard deviation, and the maximum of statistical accuracy score, and ranked lower than in terms of the minimum score. This indicates that the hypothesis that expert performances occur due to randomness is extremely unlikely.

3.5.2 Analysis of the Five-Percentile Format Data

The conventional elicitation format in Structured Expert Judgment practices is three-percentile format; however, in some cases, analysts would prefer five-percentile format where they ask experts their elicitations in five percentiles such as 5th, 25th, 50th, 75th, and 95th percentiles. Therefore, it is deemed important to test the REH on alternative elicitation formats. In this section, the same analyses performed in the previous section to the entire dataset available were computed for 17 studies that were originally performed as five-percentile format.

Table 3.4 shows statistical accuracy scores of five-percentile elicited data. The summary statistics shown in the table were incorporated into the next analyses where the corresponding statistics of the scrambled panels were compared with the original experts. Table 3.5 shows the percentages that the average, standard deviation, and the maximum of the original experts outperformed the random expert panels, and the

Table 3.4 Average, standard deviation, max, and min of original experts' statistical accuracy scores for the five-percentile format elicitation data

Study No.	Study name	Average	Standard deviation	Max	Min
28	Nebraska	8.35E-03	1.64E-02	3.30E-02	7.34E-09
29	San Diego	6.97E-04	1.43E-03	3.82E-03	1.02E-09
30	BFIQ	1.45E-01	2.56E-01	6.92E-01	3.02E-04
31	France	1.37E-01	2.88E-01	6.52E-01	1.99E-07
32	Italy	1.37E-01	2.88E-01	6.52E-01	1.99E-07
33	Spain	7.02E-06	1.00E-05	2.24E-05	1.02E-09
34	UK	6.42E-02	9.21E-02	1.85E-01	1.96E-08
35	Arkansas	1.93E-02	3.39E-02	6.98E-02	1.15E-05
36	CoveringKids	3.28E-01	3.40E-01	7.56E-01	6.23E-06
37	dcpn_Fistula	1.81E-03	3.10E-03	7.62E-03	6.23E-06
38	Florida	3.81E-02	4.63E-02	1.25E-01	1.18E-05
39	Illinois	3.68E-02	5.48E-02	1.32E-01	3.32E-07
40	Obesity	1.66E-01	2.11E-01	4.40E-01	4.09E-09
41	Tobacco	2.06E-01	2.61E-01	6.88E-01	1.05E-03
42	Washington	3.14E-02	3.09E-02	6.98E-02	3.82E-03
43	cdc-roi	1.30E-01	2.25E-01	7.20E-01	2.18E-07
44	IQ-earn	7.96E-02	1.56E-01	4.54E-01	6.97E-07

Table 3.5 The percentage of original experts' corresponding statistics ranked higher (lower for minimum) than those of the 1000 scrambled expert panels for five-percentile format elicitation data

Study name	Average (%)	Standard deviation (%)	Max (%)	Min (%)
Nebraska	85.30	85.60	85.60	97.70
San Diego	88.60	88.60	88.70	79.50
BFIQ	82.50	87.70	86.60	59.20
France	99.70	99.70	99.70	99.30
Italy	96.50	97.60	98.20	100.00
Spain	62.30%	50.30	47.90	90.40
UK	48.30	51.30	45.00	99.90
Arkansas	72.70	74.90	74.80	76.00
CoveringKids	96.40	82.00	75.50	98.60
dcpn_Fistula	18.80	20.90	15.60	10.70
Florida	50.00%	32.10	29.30	71.30
Illinois	80.10	70.90	72.20	82.60
Obesity	99.20	94.90	94.90	99.80
Tobacco	32.20	50.80	45.70	91.10
Washington	3.50	4.00	3.00	18.80
cdc-roi	96.50	92.10	77.00	94.80
IQ-earn	2.60	17.10	17.00	99.80

minimum of the original expert score is lower than the minimum of the scrambled experts.

Figure 3.6 shows that, in 7 out of 17 studies, the original experts outperformed more than 95% (i.e., 950 out of 1000 simulation runs) of the scrambled expert panels. In total, in 11 out of 17 studies, the original experts' average scores ranked higher than those of the scrambled experts from 1000 expert panels at least 80% of the time.

Figure 3.7 shows that, in 6 studies, the standard deviation of the experts' statistical accuracy scores in the original panel is larger than those in more than 95% of the 1000 randomly created expert panels. In 10 out of 17 studies, the variation in the original expert data is larger than the variation in the scrambled expert panels at least 80% of the time.

Figure 3.8 shows that, in 3 studies, the best performing expert in the original expert panel outperforms the best performing expert in the random expert panels in more than 95% of the time. In 9 out of 17 studies, the best performing original expert outperformed the best performing random expert in at least 80% of the 1000 scrambled expert panels.

Figure 3.9 shows that, in 8 studies, the poorest performing expert in the original expert panel performed poorer than the poorest performing expert in the randomly created expert panels. In 11 out of 17 studies, the original experts' minimum score is lower than the random expert panel's minimum score in at least 80% of the 1000 expert panels.

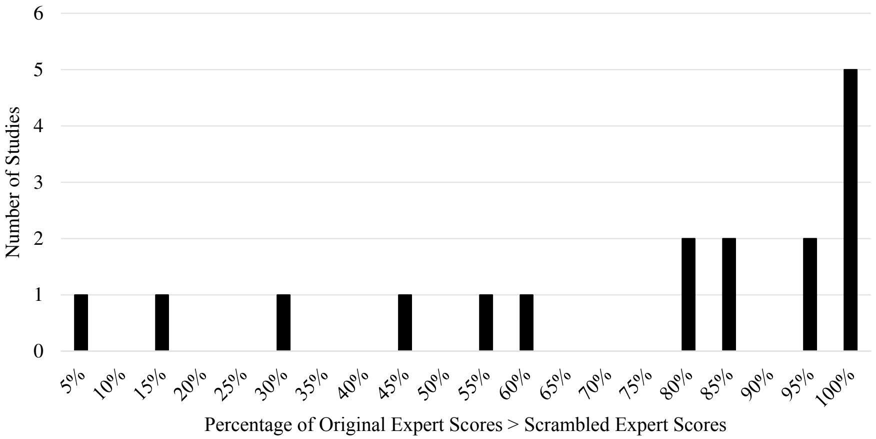


Fig. 3.6 Distribution of percentage of the average statistical accuracy of the original experts' statistical accuracy scores ranked higher among those of scrambled experts in 1000 hypothetical expert panels based on 17 studies that were originally elicited in five-percentile formats

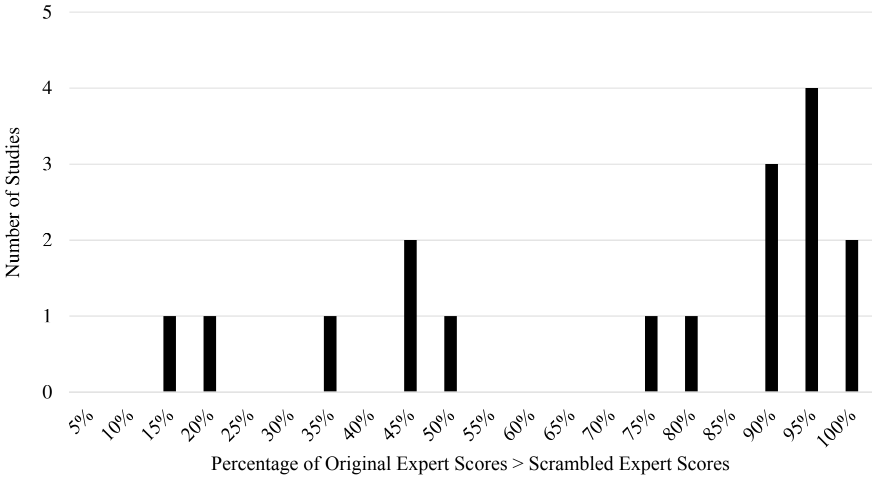


Fig. 3.7 Distribution of percentage of the standard deviation of the original experts' statistical accuracy scores ranked higher among those of scrambled experts in 1000 hypothetical expert panels based on 17 studies that were originally elicited in five-percentile formats

The Binomial results show that the average statistical accuracy scores of the original experts outperformed the randomly created experts in more than 50% of the 17 studies ($p = 0.024$). However, the Binomial test results show that the proportions of the studies in which standard deviation and maximum scores of the original experts outperform those of random experts were not statistically significant ($p = 0.167$ and $p = 0.17$, respectively). Finally, the Binomial test results indicate a significant

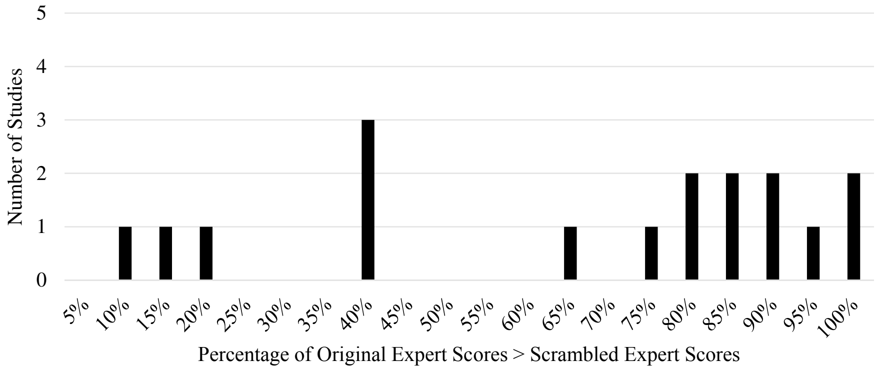


Fig. 3.8 Distribution of percentage of the maximum of the original experts' statistical accuracy scores ranked higher among those of scrambled experts in 1000 hypothetical expert panels based on 17 studies that were originally elicited in five-percentile formats

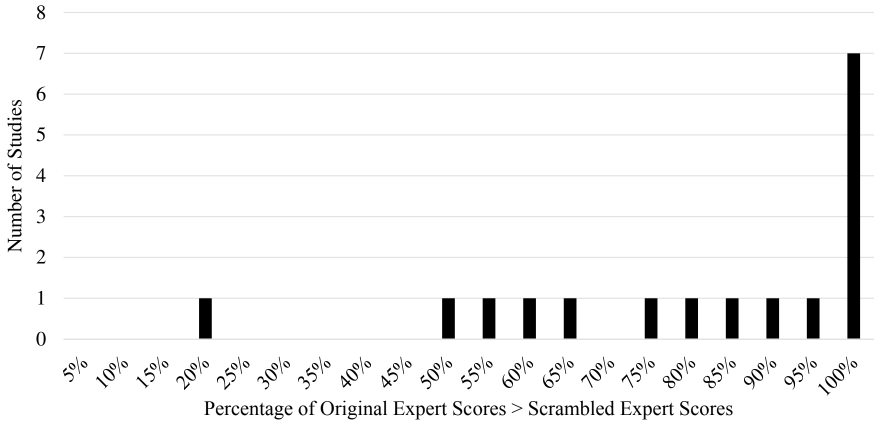


Fig. 3.9 Distribution of percentage of the minimum of the original experts' statistical accuracy scores ranked lower among those of scrambled experts in 1000 hypothetical expert panels based on 17 studies that were originally elicited in five-percentile formats

proportion of the studies in which the minimum of the original expert statistical accuracy scores was outperformed by the minimum of the random experts ($p = 0.00117$). As expected, the statistical tests with only 17 studies have much lower power.

3.5.3 A Sign Test Between the Three-Percentile Format and Five-Percentile Format Elicitation Data

Studies that were originally conducted to collect as in five percentiles (i.e., 5th, 25th, 50th, 75th, and 95th percentiles) format were converted into the three-percentile format (5th, 50th, and 95th percentiles). The average statistical accuracy scores of original experts in both formats were computed and compared by a two-sided sign test. This test was done in R, using "Wilcoxon Rank Sum and Signed Rank Test" function. The test results show that the difference between three and five-percentile formats were not statistically different ($W = 144$, $p = 1$). The corresponding test results for the standard deviation ($W = 145, p = 1$), maximum ($W= 137, p = 0.81$), and minimum ($W = 118, p = 0.37$) were also not significant, indicating that when experts are asked their elicitations in either three or five-percentile formats, their statistical accuracy did not significantly change. This implies that the number of probability bins and in turn bin range (e.g., whether covers 25% or 45%) do not significantly influence experts' statistical accuracy.

Furthermore, the sign test was performed to test whether the original experts' outperformance percentages differed in three-percentile format than in five-percentile format. Sign test results show that there was not a statistical difference between the three-percentile format analysis and five-percentile format analysis in terms of percentages that the original experts outperform the scrambled experts in 1000 simulations ($W = 146.5$, $p = 0.96$). Similar analysis was done for standard deviation ($W = 146, p = 0.97$; i.e., the percentage that the original experts outperform the random experts in their standard deviation), for the maximum ($W = 141$, $p = 0.92$; i.e., the percentage that the original experts outperform the scrambled experts in their maximum scores), and for the minimum ($W = 131\ p = 0.65$; i.e., the percentage that the minimum score of the original experts is less than the minimum score of the random expert panels).

3.6 Concluding Remarks

This book chapter addresses the fundamental limitation of the equal weighting approach, namely that experts are expected to be interchangeable. This assumption has severe implications because it treats the best performing experts equally with the poor performing experts. Specifically, it leads to a depreciation of the maximum value of the expert input by undervaluing useful expert elicitations and overvaluing redundant or misleading elicitations of poorly performing experts. In order to address the aforementioned limitation of the equal weighting approach, the random expert hypothesis was used to test if experts should be treated equally. The results provide strong evidence that the original expert panels outperform randomly created experts. Specifically, the performances of the original experts with those of randomly scrambled experts were compared in terms of their statistical accuracy. Results show that the original experts perform better than the randomly created experts; their statistical accuracy scores spread more since there are good and poor performing experts, which illustrates the potential problem of the equal weight approach. It may not be reasonable to assign all experts equal weights.

The present study also tested whether the results are replicated in the different elicitation format, specifically three versus five-percentile format. This analysis has significant practical implications. Showing the differences in statistical accuracy in different elicitation formats offers valuable insights to analysts so that they can decide the number of bins that they would ask experts to elicit. If there are performance differences between the three and five-quantile formats, they are too small to be detected with the current dataset. This question could be revisited in the future as more data become available.

This study focused on comparing performances in terms of statistical accuracy scores. As proposed by the Classical Model (e.g., Cooke 1991; Cooke et al. 2008), the statistical accuracy score is the dominant component in expert decision weight computations. Specifically, the Classical Model gives the power to the analyst to exclude the assessment of an expert whose statistical accuracy performance is less than a given threshold. In other words, it is the statistical accuracy that determines whether an experts' input is included into the analysis. As aforementioned, the information score functions serve as a modulating factor for evaluating expert performances. There may be cases where experts can provide large intervals indicating greater uncertainty in their estimates, which would still guarantee a high statistical accuracy score yet may not be as informative. Information score is an effective way to penalize those experts. Therefore, it is encouraged to investigate the random expert hypothesis based on decision weights that encompasses both statistical accuracy and information score. In future studies, thorough analyses including large dataset will be analyzed.

Finally, it is useful to compare this study with previous cross-validation studies (Eggstaff et al. 2014; Colson and Cooke 2017). Those studies considered all non-trivial splits of the statistical accuracy variables into training and test sets. The Classical Model performance weight was initialized on each training set and compared to equal weighting on the respective test sets. Although these studies showed significant out-of-sample superiority for performance weighting, the results were tempered by the fact that the performance weighting based on each training set is not the same as the performance weighting based on all variables. There was an out-of-sample penalty for statistical accuracy which decreased with training set size, but which obviously could not be eliminated. Hence, the superiority of performance weighting was largely driven by the higher informativeness of the performance weighted decision maker. The present results utilize the full set of statistical accuracy variables and do not consider informativeness. This suggests that performance weighting is also superior with respect to statistical accuracy in addition to informativeness. Working this out is a task for future research.

Appendix

Data references table

Study name	References
UMD	Koch, Benjamin J., Filoso, S., Cooke, R. M. Hosen, J. D., Colson, A.R. Febria, Catherine M., Palmer, M. A., (2015) Nitrogen in stormwater runoff from Coastal Plain watersheds: The need for empirical data, reply to Walsh, Elementa DOI https://doi.org/10.12952/journal.elementa.000079. https://www.elementascience.org/articles/79 Koch, Benjamin J., Febria, Catherine M., Cooke, Roger M. Hosen, Jacob D., Baker, Matthew E., Colson, Abigail R. Filoso, Solange, Hayhoe, Katharine, Loperfido, J.V., Stoner, Anne M.K., Palmer, Margaret A., (2015) Suburban watershed nitrogen retention: Estimating the effectiveness of storm water management structures, Elementa, https://doi.org/10.12952/journal.elementa.000063 https://www.elementascience.org/articles/63
USGS	Newhall, C. G., & Pallister, J. S. (2015). Using multiple data sets to populate probabilistic volcanic event trees. In Volcanic Hazards, Risks and Disasters (pp. 203–232)
arsenic	Hanzich, J.M. (2007) Achieving Consensus: An Analysis Of Methods To Synthesize Epidemiological Data For Use In Law And Policy. Department of Public Health & Primary Care, Institute Of Public Health, University of Cambridge; unpublished MPhil thesis, 66 pp + appendices
Biol Agents	Aspinall & Associates (2006). REBA Elicitation. Commercial-in-confidence report, pp. 26
Geopolit	Ismail and Reid (2006). "Ask the Experts" presentation
ATCEP	Morales-Nápoles, O., Kurowicka, D., & Cooke, R. (2008). EEMCS final report for the causal modeling for air transport safety (CATS) project
Daniela	Forys, M.B., Kurowicka, D., Peppelman, B.(2013) "A probabilistic model for a gas explosion due to leakages in the grey cast iron gas mains" Reliability Engineering & System Safety volume 119, issue, year 2013, pp. 270–279
eBBP	Tyshenko, M.G., S. ElSaadany, T. Oraby, M. Laderoute, J. Wu, W. Aspinall and D. Krewski (2011) Risk Assessment and Management of Emerging Blood-Borne Pathogens in Canada: Xenotropic Murine Leukaemia Virus-Related Virus as a Case Study for the Use of a Precautionary Approach. Chapter in: *Risk Assessment* (ISBN 979-953-307-765-8) Cashman, N.R., Cheung, R., Aspinall, W., Wong, M. and Krewski, D. (2014) Expert Elicitation for the Judgment of Prion Disease Risk Uncertainties associated with Urine-derived and Recombinant Fertility Drugs. Submitted to: Journal of Toxicology and Environmental Health
create	Bier V.M, Kosanoglu, F, Shin J, unpublished data, nd
effErupt	Aspinall, W.P. (2012) Comment on "Social studies of volcanology: knowledge generation and expert advice on active volcanoes" by Amy Donovan, Clive Oppenheimer and Michael Bravo [*Bull Volcanol* (2012) 74:677–689] Bulletin of Volcanology, 74, 1569–1570. https://doi.org/10.1007/s00445-012-0625-x

(continued)

(continued)

Study name	References
erie	Colson, Abigail R., Sweta Adhikari, Ambereen Sleemi, and Ramanan Laxminarayan. (2015) "Quantifying Uncertainty in Intervention Effectiveness with Structured Expert Judgment: An Application to Obstetric Fistula." BMJ Open, 1–8. https://doi.org/10.1136/bmjopen-2014-007233 Cooke, R.M., Wittmann, M.E., Lodge, D.M., Rothlisberger, J.D., Rutherford E.S., Zhang, H. and Mason, D.M. (2014) "Out-of-Sample Validation for Structured Expert Judgment of Asian Carp Establishment in Lake Erie", Integrated Environmental Assessment and Management, open access. DOI: https://doi.org/10.1002/ieam.1559 Zhang, H, Rutherford E.S., Mason, D.M., Breck, J,T,, Wittmann M.E., Cooke R.M., Lodge D.M., Rothlisberger J.D., Zhu X., and Johnson, T B., (2015) Forecasting the Impacts of Silver and Bighead Carp on the Lake Erie Food Web, Transactions of the American Fisheries Society, Volume 145, Issue 1, pp 136–162, https://doi.org/10.1080/00028487.2015.1069211
FCEP	Leontaris, G., & Morales-Nápoles, O. (2018). ANDURIL—A MATLAB toolbox for ANalysis and Decisions with UnceRtaInty: Learning from expert judgments. SoftwareX, 7, 313–317
Sheep	Hincks, T., Aspinall, W. and Stone, J. (2015) Expert judgement elicitation exercise to evaluate Sheep Scab control measures: Results of the Bayesian Belief Network analysis. University of Bristol PURE Repository Working Paper (forthcoming)
hemophilia	Fischer K, Lewandowski D, Janssen MP. Estimating unknown parameters in haemophilia using expert judgement elicitation. Haemophilia. 2013 Sep;19(5):e282–e288
Liander	Forys, M.B., Kurowicka, D., Peppelman, B.(2013) "A probabilistic model for a gas explosion due to leakages in the grey cast iron gas mains" Reliability Engineering & System Safety volume 119, issue, year 2013, pp. 270–279
PHAC	Oraby,T., Tyshenko, M.G., Westphal, M., Darshan, S., Croteau, M., Aspinall, W., Elsaadany, S., Cashman, N. and Krewski, D. (2011) Using Expert Judgments to Improve Chronic Wasting Disease Risk Management in Canada. Journal of Toxicology and Environmental Health, in press. Volume 74, Issue 2-4, 2011 Special Issue: Prion Research in Perspective 2010
TOPAZ	Scourse, E., Aspinall, W.P. and Chapman, N. (2014) Using expert elicitation to characterise long-term tectonic risks to radioactive waste repositories in Japan. Journal of Risk Research, https://doi.org/10.1080/13669877.2014.971334
SPEED	Hicks, A., Barclay, J., Simmons, P. and Loughlin, S. (2014). "An interdisciplinary approach to volcanic risk reduction under conditions of uncertainty: a case study of Tristan da Cunha." Nat. Hazards Earth Syst. Sci. 14(7): 1871-1887. https://doi.org/10.5194/nhess-14-1871-2014. www.nat-hazards-earth-syst-sci-discuss.net/1/7779/2013/ Bevilacqua, A., Isaia, R., Neri, A., Vitale, S., Aspinall, W.P. and eight others (2015) Quantifying volcanic hazard at Campi Flegrei caldera (Italy) with uncertainty assessment: I. Vent opening maps. Journal of Geophysical Research—Solid Earth; AGU. https://doi.org/10.1002/2014jb011775
TDC	Scourse, E., Aspinall, W.P. and Chapman, N. (2014) Using expert elicitation to characterise long-term tectonic risks to radioactive waste repositories in Japan. Journal of Risk Research, https://doi.org/10.1080/13669877.2014.971334

(continued)

(continued)

Study name	References
GL	Rothlisberger,J.D. Finnoff, D.C. Cooke,R.M. and Lodge, D.M. (2012) "Ship-borne nonindigenous species diminish Great Lakes ecosystem services" Ecosystems (2012) 15: 462–476 https://doi.org/10.1007/s10021-012-9522-6 Rothlisberger, J.D., Lodge, D.M. Cooke, R.M. and Finnoff, D.C. (2009) "Future declines of the binational Laurentian Great Lakes fisheries: recognizing the importance of environmental and cultural change" *Frontiers in Ecology and the Environment;* https://doi.org/10.1890/090002
Goodheart	Goodheart, B. (2013). Identification of causal paths and prediction of runway incursion risk by means of Bayesian belief networks. Transportation Research Record: Journal of the Transportation Research Board, (2400), 9–20.
Ice	Bamber, J.L., and Aspinall, W.P., (2012) An expert judgement assessment of future sea 1evel rise from the ice sheets, Nature Climate Change, PUBLISHED ONLINE: January 6, 2012| https://doi.org/10.1038/nclimate1778. http://www.nature.com/nclimate/journal/vaop/ncurrent/full/nclimate1778.html
puig-gdp	Puig, D., Morales-Nápoles, O., Bakhtiari, F., & Landa, G. (2018). The accountability imperative for quantifying the uncertainty of emission forecasts: evidence from Mexico. Climate Policy, 18(6), 742–751
puig-oil	Puig, D., Morales-Nápoles, O., Bakhtiari, F., & Landa, G. (2018). The accountability imperative for quantifying the uncertainty of emission forecasts: evidence from Mexico. Climate Policy, 18(6), 742–751
YTBID (CDC)	Colson, Abigail R., M.A. Cohen, S. Regmi, A. Nandi, R. Laxminarayan (2015) "Structured Expert Judgment for Informing the Return on Investment in Surveillance: The Case of Environmental Public Health Tracking." Working Paper. Center for Disease Dynamics, Economics & Policy
Gerestenberger	Gerstenberger, M. C., et al. (2016). "A Hybrid Time-Dependent Probabilistic Seismic-Hazard Model for Canterbury, New Zealand." Seismological Research Letters. Vol. 87 Doi: https://doi.org/10.1785/0220160084 Gerstenberger, M.C.; McVerry, G.H.; Rhoades, D.A.; Stirling, M.W. (2014) Seismic hazard modeling for the recovery of Christchurch, New Zealand.*Earthquake Spectra, 30(1):* 17–29; https://doi.org/10.1193/021913eqs037m Gerstenberger, M.C.; Christophersen, A.; Buxton, R.; Allinson, G.; Hou, W.; Leamon, G.; Nicol, A. (2013) Integrated risk assessment for CCS. p. 2775–2782; https://doi.org/10.1016/j.egypro.2013.06.162 IN: Dixon, T.; Yamaji, K. (eds) *11th International Conference on Greenhouse Gas Control Technologies, 18th-22nd November 2012, Kyoto International Conference Center, Japan.* Elsevier. *Energy procedia 37*

(continued)

(continued)

Study name	References
CWD	Tyshenko, M.G., ElSaadany, S., Oraby, T., Darshan, S., Catford, A., Aspinall, W., Cooke, R. and Krewski, D. (2012) Expert judgement and re-elicitation for prion disease risk uncertainties. *International Journal of Risk Assessment and Management*, 16(1–3), 48–77. https://doi.org/10.1504/ijram.2012.047552 Tyshenko, M.G., S. ElSaadany, T. Oraby, S. Darshan, W. Aspinall, R. Cooke, A. Catford, and D. Krewski (2011) Expert elicitation for the judgment of prion disease risk uncertainties. *J Toxicol Environ Health* A.; 74(2–4):261–285 Oraby,T., Tyshenko, M.G., Westphal, M., Darshan, S., Croteau, M., Aspinall, W., Elsaadany, S., Cashman, N. and Krewski, D. (2011) Using Expert Judgments to Improve Chronic Wasting Disease Risk Management in Canada. Journal of Toxicology and Environmental Health, in press. Volume 74, Issue 2–4, 2011 Special Issue: Prion Research in Perspective 2010
Nebraska	Attribution study for Robert Wood Johnson Covering Kids & Families in Pennsylvania, Washington, Nebraska, Illinois, Arkansas, and Florida, conducted by Center for Disease Dynamics, Economics & Policy, 2012
San Diego	Attribution study for Robert Wood Johnson Covering Kids & Families in Pennsylvania, Washington, Nebraska, Illinois, Arkansas, and Florida, conducted by Center for Disease Dynamics, Economics & Policy, 2012
BFIQ	Colson, A. Cooke, R.M., Lutter, Randall, (2016) How Does Breastfeeding Affect IQ? Applying the Classical Model of Structured Expert Judgment, Resources for the Future, RFF DP16–28 http://www.rff.org/research/publications/how-does-breastfeeding-affect-iq-applying-classical-model-structured-expert
France	Abigail R. Colson, Itamar Megiddo, Gerardo Alvarez-Uria, Sumanth Gandra, Tim Bedford, Alec Morton, Roger M. Cooke, Ramanan Laxminarayan (ns). “Quantifying Uncertainty about Future Antimicrobial Resistance: Comparing Structured Expert Judgment and Statistical Forecasting Methods.”
Italy	Abigail R. Colson, Itamar Megiddo, Gerardo Alvarez-Uria, Sumanth Gandra, Tim Bedford, Alec Morton, Roger M. Cooke, Ramanan Laxminarayan (ns). “Quantifying Uncertainty about Future Antimicrobial Resistance: Comparing Structured Expert Judgment and Statistical Forecasting Methods.”
Spain	Abigail R. Colson, Itamar Megiddo, Gerardo Alvarez-Uria, Sumanth Gandra, Tim Bedford, Alec Morton, Roger M. Cooke, Ramanan Laxminarayan (ns). “Quantifying Uncertainty about Future Antimicrobial Resistance: Comparing Structured Expert Judgment and Statistical Forecasting Methods.”
UK	Abigail R. Colson, Itamar Megiddo, Gerardo Alvarez-Uria, Sumanth Gandra, Tim Bedford, Alec Morton, Roger M. Cooke, Ramanan Laxminarayan (ns). “Quantifying Uncertainty about Future Antimicrobial Resistance: Comparing Structured Expert Judgment and Statistical Forecasting Methods.”
Arkansas	Attribution study for Robert Wood Johnson Covering Kids & Families in Pennsylvania, Washington, Nebraska, Illinois, Arkansas, and Florida, conducted by Center for Disease Dynamics, Economics & Policy, 2012.
CoveringKids	Attribution study for Robert Wood Johnson Covering Kids & Families in Pennsylvania, Washington, Nebraska, Illinois, Arkansas, and Florida, conducted by Center for Disease Dynamics, Economics & Policy, 2012

(continued)

(continued)

Study name	References
dcpn_Fistula	Aspinall,W. Devleesschauwer, B. Cooke, R.M., Corrigan,T., Havelaar, A.H., Gibb, H., Torgerson, P., Kirk, M., Angulo, F., Lake, R., Speybroeck, N., and Hoffmann, S. (2015) World Health Organization estimates of the relative contributions of food to the burden of disease due to selected foodborne hazards: a structured expert elicitation. PLOS ONE,: January 19, 2016 https://doi.org/10.1371/journal.pone.0145839
Florida	Attribution study for Robert Wood Johnson Covering Kids & Families in Pennsylvania, Washington, Nebraska, Illinois, Arkansas, and Florida, conducted by Center for Disease Dynamics, Economics & Policy, 2012
Illinois	Attribution study for Robert Wood Johnson Covering Kids & Families in Pennsylvania, Washington, Nebraska, Illinois, Arkansas, and Florida, conducted by Center for Disease Dynamics, Economics & Policy, 2012
Obesity	Colson, Abigail R., R.M. Cooke, R. Laxminarayan. (2015) "Attributing Impact to a Charitable Foundation's Programs with Structured Expert Judgment." Working Paper. Center for Disease Dynamics, Economics & Policy
Tobacco	Colson, Abigail R., R.M. Cooke, R. Laxminarayan. (2015) "Attributing Impact to a Charitable Foundation's Programs with Structured Expert Judgment." Working Paper. Center for Disease Dynamics, Economics & Policy
Washington	Attribution study for Robert Wood Johnson Covering Kids & Families in Pennsylvania, Washington, Nebraska, Illinois, Arkansas, and Florida, conducted by Center for Disease Dynamics, Economics & Policy, 2012
cdc-roi	Colson, Abigail R., M.A. Cohen, S. Regmi, A. Nandi, R. Laxminarayan (2015) "Structured Expert Judgment for Informing the Return on Investment in Surveillance: The Case of Environmental Public Health Tracking." Working Paper. Center for Disease Dynamics, Economics & Policy
IQ-earn	Randall Lutter, Abigail Colson, and Roger Cooke (ns), (ns), "Effects of Increases in IQ in India on the Present Value of Lifetime Earnings

References

Aspinall, W. (2010). A route to more tractable expert advice. *Nature, 463*(7279), 294.

Aspinall, W. P., Cooke, R. M., Havelaar, A. H., Hoffmann, S., & Hald, T. (2016). Evaluation of a performance-based expert elicitation: WHO global attribution of foodborne diseases. *PLoS ONE, 11*(3), e0149817.

Bamber, J. L., & Aspinall, W. P. (2013). An expert judgement assessment of future sea level rise from the ice sheets. *Nature Climate Change, 3*(4), 424.

Bamber, J. L., Aspinall, W. P., & Cooke, R. M. (2016). A commentary on "how to interpret expert judgment assessments of twenty-first century sea-level rise" by Hylke de Vries and Roderik SW van de Wal. *Climatic Change, 137*(3–4), 321–328.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5*(4), 559–583.

Clemen, R. T. (2008). Comment on Cooke's classical method. *Reliability Engineering & System Safety, 93*(5), 760–765.

Clemen, R. T., & Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis, 19*(2), 187–203.

Colson, A. R., & Cooke, R. M. (2017). Cross validation for the classical model of structured expert judgment. *Reliability Engineering & System Safety, 163,* 109–120.

Cooke, R. (1991). *Experts in uncertainty: Opinion and subjective probability in science*. Oxford University Press on Demand.

Cooke, R. M., & Goossens, L. L. (2008). TU Delft expert judgment data base. *Reliability Engineering & System Safety, 93*(5), 657–674.

Eggstaff, J. W., Mazzuchi, T. A., & Sarkani, S. (2014). The effect of the number of seed variables on the performance of Cooke′ s classical model. *Reliability Engineering & System Safety, 121,* 72–82.

Einhorn, H. J. (1974). Expert judgment: Some necessary conditions and an example. *Journal of Applied Psychology, 59*(5), 562.

French, S. (1981). Consensus of opinion. *European Journal of Operational Research, 7*(4), 332–340.

Genest, C., & Zidek, J. V. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science, 1*(1), 114–135.

Hald, T., Aspinall, W., Devleesschauwer, B., Cooke, R., Corrigan, T., Havelaar, A. H., et al. (2016). World Health Organization estimates of the relative contributions of food to the burden of disease due to selected foodborne hazards: A structured expert elicitation. *PLoS ONE, 11*(1), e0145839.

Jaiswal, K. S., Aspinall, W., Perkins, D., Wald, D., & Porter, K. A. (2012). Use of expert judgment elicitation to estimate seismic vulnerability of selected building types. In *15th World Conference on Earthquake Engineering (WCEE)*. Lisbon, Portugal, Sept (pp. 24–28).

Keeney, R. L., & Von Winterfeldt, D. (1989). On the uses of expert judgment on complex technical problems. *IEEE Transactions on Engineering Management, 36*(2), 83–86.

Morgan, M. G., Henrion, M., & Small, M. (1992). *Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis*. Cambridge university press.

Mosleh, A., Bier, V. M., & Apostolakis, G. (1988). A critique of current practice for the use of expert opinions in probabilistic risk assessment. *Reliability Engineering & System Safety, 20*(1), 63–85.

Otway, H., & von Winterfeldt, D. (1992). Expert judgment in risk analysis and management: process, context, and pitfalls. *Risk Analysis, 12*(1), 83–93.

Ouchi, F. (2004). A literature review on the use of expert opinion in probabilistic risk analysis.

Ryan, J. J., Mazzuchi, T. A., Ryan, D. J., De la Cruz, J. L., & Cooke, R. (2012). Quantifying information security risks using expert judgment elicitation. *Computers & Operations Research, 39*(4), 774–784.

Singpurwalla, N. D. (1988). Foundational issues in reliability and risk analysis. *SIAM Review, 30*(2), 264–282.

Spetzler, C. S., & Stael von Holstein, C. A. S. (1975). Exceptional paper—probability encoding in decision analysis. *Management Science, 22*(3), 340–358.

Stone, M. (1961). The opinion pool. *The Annals of Mathematical Statistics*, 1339–1342.

Tyshenko, M. G., ElSaadany, S., Oraby, T., Darshan, S., Aspinall, W., Cooke, R., et al. (2011). Expert elicitation for the judgment of prion disease risk uncertainties. *Journal of Toxicology and Environmental Health, Part A, 74*(2–4), 261–285.

Wallsten, T. S., & Budescu, D. V. (1983). State of the art—encoding subjective probabilities: A psychological and psychometric review. *Management Science, 29*(2), 151–173.

Wilson, K. J. (2017). An investigation of dependence in expert judgement studies with multiple experts. *International Journal of Forecasting, 33*(1), 325–336.

Chapter 4
Customized Structural Elicitation

Rachel L. Wilkerson and Jim Q. Smith

4.1 Background

Expert elicitation is a powerful tool when modelling complex problems especially in the common scenario when current probabilities are unknown and data is unavailable for certain regions of the probability space. Such methods are now widely developed, well understood, and have been used to model systems in a variety of domains including climate change, food insecurity, and nuclear risk assessment (Barons et al. 2018; Rougier and Crucifix 2018; Hanea et al. 2006.) However, eliciting expert probabilities faithfully has proved to be a sensitive task, particularly in multivariate settings. We argue that first eliciting structure is critical to the accuracy of the model, particularly as conducting a probability elicitation is time and resource-intensive.

An appropriate model structure fulfils two criteria. Firstly, it should be compatible with how experts naturally describe a process. Ideally, modellers should agree on a structure using natural language. Secondly, any structure should ideally have the potential to eventually be embellished through probabilistic elicitation into a full probability model. It is often essential to determine that the structure of a problem as desired by a domain expert is actually consistent with the class of structural models considered. The logic and dynamics of Bayesian networks (BN) often do not match with an experts' description of a problem. When this happens, a customizing approach as we illustrate below generates flexible models that are a more accurate representation of the process described by the domain expert. We show that these

R. L. Wilkerson (✉)
University of Warwick, Warwick, UK
e-mail: R.L.Wilkerson@warwick.ac.uk

J. Q. Smith
Alan Turing Institute, University of Warwick, Warwick, UK
e-mail: J.Q.Smith@warwick.ac.uk

A. M. Hanea et al. (eds.), *Expert Judgement in Risk and Decision Analysis*,
International Series in Operations Research & Management Science 293,
https://doi.org/10.1007/978-3-030-46474-5_4

alternative graphical models often admit a supporting formal framework and subsequent probabilistic model similar to a BN while more faithfully representing the beliefs of the experts.

While there are several protocols for eliciting probability distributions such as the Cooke method, SHELF, and IDEA protocols (Cooke 1991; Hagan and Oakley 2014; Hanea et al. 2018), the process of determining the appropriate underlying structure has not received the same attention. Protocols for eliciting structural relationships between variables in the continuous range have been developed (Tim and Roger 2001), and basic guidelines for eliciting a discrete Bayesian network structure are available and well documented (Korb and Nicholson 2009; Smith 2010). These methods are widely applicable but are rarely customized to structural elicitation of models other than the BN. However, it is possible to develop customizing protocols to elicit structure. We illustrate this through the case studies in this chapter.

Towards this end, this chapter explores examples of real case studies that are better suited to eliciting bespoke structure. We illustrate how experts' natural language description of a problem can determine the structure of a model. Programs to alleviate food insecurity in the United States serve as a running example. Even within this domain we are able to show that different problem dynamics are naturally more suited to particular structures, and eliciting these custom structures creates more compelling models. We show that these bespoke structures can subsequently be embellished into customized probabilistic graphical models that support a full probabilistic description.

4.2 Eliciting Model Structure

Structured expert elicitation begins with a natural language description of the problem from domain experts. An expert describes the components of a system and how they are related, and a structure often emerges organically. This process may be aided by the use of informal graphs, a widespread practice. However, the methods used by the facilitators for systematically translating these diagrams into their logical consequences and finally embellishing these into a full probability model are often not supported. Nevertheless, there are certain well-developed classes of graphical models that do support this translation. The most popular and best supported by available software is the Bayesian network. However, other graphical frameworks have emerged, each with its own representative advantages. These include event trees, chain event graphs, and dynamic analogues of these (Collazo et al. 2017; Barclay et al. 2015). We describe some of the competing frameworks and suggest how one can be selected over another.

4.2.1 Choosing an Appropriate Structure

Choosing between candidate structures may not be straightforward. Some domain problems may be compatible with existing structures, while others might require creating new classes of probabilistic graphical models. The task of developing a bespoke graphical framework that supports a translation into a choice of probability models is usually a labour-intensive one requiring some mathematical skills. While some domain problems will require the modeller to undertake developing a customized model class, there are also several such frameworks already built, forming a tool kit of different frameworks (Collazo et al. 2017; Smith 1993; Smith and Figueroa 2007; Liverani and Smith 2015). We give guidelines below to help the modeller decide which of these methods most closely match the problem explanation given by the domain experts.

As a running example, we consider the drivers of food insecurity. The illustrations we use throughout the chapter are based on meetings with actual domain experts. We have simplified these case studies so that we can illustrate the elicitation process as clearly as possible. A meeting of advocates discusses the effect of food insecurity on long-term health outcomes. One advocate voices that food insecurity stems from insufficient resources to purchase food. The experts collectively attest that the two main sources of food are personal funds like disposable income or government benefit programs. The government benefit programs available to eligible citizens include child nutrition programs that provide free school breakfast, lunch, and after school snacks, the Supplemental Nutrition Assistance Program (SNAP), and Temporary Assistance for Needy Families (TANF). From this discussion among experts, modellers need to resolve the discussion into several key elements of the system. One potential set of elements drawn from the expert discussion is shown below:

- Government benefits, B: the rate at which a particular neighbourhood is participating in all available government programs.
- Disposable Income, I: the average amount of income available for purchasing food in the neighbourhood.
- Food insecurity, F: the rate at which families and individuals in a neighbourhood experience insufficient access to food.
- Long-term health outcomes, H: measured by an overall health index defined at the neighbourhood level.

There are several guiding principles to help modellers create a structure that is faithful to the experts' description.

4.2.1.1 Scope

One common difficulty that appears in many structural elicitation exercises is the tendency of expert groups to think only in terms of measured quantities, rather than underlying drivers. Food insecurity and poverty researchers often consider elements of the system as documented for policy-makers, whereas those with a firsthand knowledge of food insecurity may consider a different set of drivers, like personal trauma (Dowler and O'Connor 2012; Chilton and Rose 2009). Anecdotes of food insecurity may often draw out key, overlooked features of the system, but a well-defined problem scope is critical to prevent a drifting purpose. The responsibility of guiding the conversation continually towards general representations rather than particular instances falls to the modeller.

4.2.1.2 Granularity

Elicitations typically begin with a coarse description before refining the system. Considering refinements and aggregations can help the experts' opinions of the key elements of the system to coalesce. For instance, rather than modelling all the government benefits together in B, we could have defined a variable for each benefit program, child nutrition programs, C, and financial support for individuals S. Because the experts are interested in the well-being of the neighbourhood as a whole, it is sensible to model the problem with aggregate rather than individual benefits. The granularity of key elements depends on the modeller's focus. Thinking of the problem at different spatial levels may help to choose the appropriate granularity.

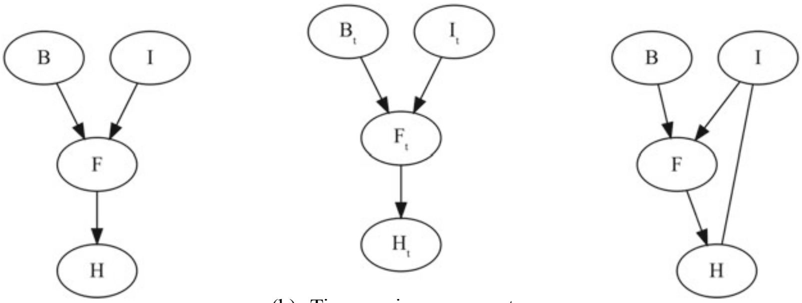
4.2.1.3 Potential Interventions

Another guiding principle during the structural elicitation is ensuring that possible interventions are represented by the system components. For instance, if the policy experts wanted to know what would happen after increasing all benefit programs simultaneously, modelling benefits collectively as B would be appropriate. But if they want to study what happens by intervening on child nutrition programs, then separating this node into C, child nutrition programs, and allowing B to represent additional benefit programs would compose a more suitable model.

4.2.1.4 Context Dependence

As the key elements of the system emerge, testing the structure by imagining these key elements in a different structure may either restrict or elucidate additional model features. The drivers that cause food insecurity at the neighbourhood level may vary greatly from those that provoke food insecurity at the individual household level.

For this running example, the experts focus on the neighbourhood level. They speak about each of the variables as the particular incidence rates for a neighbourhood. The modeller could then draw a dependence structure for random variables from their discussion about the dependence between these measurements. This structure would be most conducive to a Bayesian network. An example of one tentative BN structure has tried to accommodate these points in Fig. 4.1a.



(a) BN of food insecurity at the neighbourhood level

(b) Time series representation of food insecurity drivers over time

(c) Hybrid representation of food insecurity drivers

Fig. 4.1 Three different representations with nodes and edges customized to the experts' beliefs

4.2.1.5 Importance of Temporal Processes

Another key modelling decision is whether or not to use a dynamic network model. Are the experts speaking about potential interventions that are time-dependent or not? Do the key elements of the process change drastically over time? Few elements of a system are ever truly static, but dynamic models should only be chosen when the temporal element is crucial to the experts' description of the system as they are often more computationally intensive.

In contrast to the static example of measurements given above, suppose that the experts believe that yearly fluctuations in disposable income I directly affect the rates of food insecurity F. This is a dynamic process. Another expert might draw on literature that shows the linear relationship between I and F. Using a standard Bayesian network for this problem description would not capture the temporal information or the strength between each of the pairs of nodes. The quantities of the graph here are not static random variables, but rather its nodes appear to be representing processes. In this case, a more appropriate choice for the graphical elements would be to represent them as time series B_t, I_t, F_t, H_t. This graph is shown at a single time point in Fig. 4.1b. The probabilistic model can be embellished into a number of different stochastic descriptions as will be discussed later in this chapter.

The meaning of the graph begins to change as the modeller learns more about the structure of a problem. We suggest ways in which we could begin to frame different models for a desired context in terms of nodes and edges. Nodes for general graphical models can be any mathematical objects suitable to the given domain, provided that the system can be actually represented in terms of a probabilistic distribution which is consistent with the meaning we can ascribe to the model edges.

Once we have established the nodes, the relationships between variables must be represented. These are usually expressed in terms of oriented edges or colourings in the vertices. Continuing with our toy example, the advocates promptly recognize that government benefits and disposable income directly impact the state of food insecurity. It also appears natural, as another expert attests, to associate the long-term health as dependent on food insecurity. These three relationships give us the graph in Fig. 4.1a.

The experts comment that the available money for food purchasing directly affects how much food a family can buy, making directed edges a natural fit for B to F and I to F. However, the relationship between long-term health outcomes and disposable income is less clear. One advocate mentions that individuals and families who are battling chronic illness or faced with an outstanding medical bill are less likely to have disposable income, and thus more likely to be food insecure. However, using the typical BN machinery, adding an edge between long-term health outcomes and disposable income would induce a cycle in the graph and thus render the BN inadmissible.

One common solution would be to simply ignore this information and proceed only with the BN given previously. A second solution would be to embellish the model into a dynamic representation that could formally associate this aspect of the process by expressing instantaneous relationships in a single time slice of effects between nodes on different time slices. A time slice simply denotes the observations of the variables at a given time point. Another method might be to incorporate an undirected edge that could be used to represent the ambiguous relationship between I and H. The result is a hybrid graph with undirected and directed edges with its own logic shown in Fig. 4.1c.

Whatever semantic we choose, edges should represent the experts' natural language description of the relationships. Returning to the instance in which the experts speak about food insecurity as a time series, the edges represent regression coefficients as the system unfolds. As we will show below, directed acyclic graphs (DAGs) are particularly convenient for modelling. However, there are graphical representations that permit cycles, should the modeller wish to focus on the cyclic nature of F and H. The choice between the type and orientation of edge affects the semantics of the model as shown below.

4.2.2 Stating Irrelevancies and Checking Conditional Independence Statements

Suppose we choose to represent a domain expert's problem with a BN. Often, it is more natural for experts to impart meaning to the edges present in a graphical model. Unfortunately, it is the absence of edges that represent the conditional independences. To facilitate a transparent elicitation process, these conditional independence relationships can be expressed in a more accessible way as questions about which variables are irrelevant to the other.

Domain experts who are not statistically trained do not naturally read irrelevance statements from a BN. So it is often important to explicitly unpick each compact irrelevance statement written in the graph and check its plausibility with the domain expert.

Generally, suppose the domain expert believes that X is irrelevant for predicting Y given the measurement Z. That is, knowing the value of X provides no additional information about Y given information about Z. These beliefs can be written as $X \perp\!\!\!\perp Y \mid Z$, read as X is independent of Y conditional on Z.

For our example, the missing edges indicate three conditional independence relationships $H \perp\!\!\!\perp B \mid F$, $H \perp\!\!\!\perp I \mid F$ and $B \perp\!\!\!\perp I$. To check these, the modeller would ask the following questions to the domain expert:

- If we know what the food insecurity status is, does knowing what the disposable income is provide any additional information about long-term health?
- Assuming we know the food insecurity level, does the government benefit level offer any more insight into the long-term health of a neighbourhood?
- Does knowing disposable income levels of a neighbourhood provide further information about the government benefit level?

This last question might prompt the expert to realize that indeed, disposable income influences eligibility for government benefits, so an edge would be added between B and I.

These questions can also be rephrased according to the semigraphoid axioms, a simplified set of rules that hold for a given set of conditional independence statements. It is helpful to include these as they provide a template for different rule-based styles for other frameworks that capture types of natural language. More details can be found in Smith (2010). The first such axiom is given below.

Definition 4.1 The **symmetry property** requires that for three disjoint measurements X, Y, and Z:

$$X \perp\!\!\!\perp Y \mid Z \Leftrightarrow Y \perp\!\!\!\perp X \mid Z$$

This axiom asserts that assuming Z is known, if X tells us nothing new about Y, then knowing Y also provides no information about X.

The second, stronger semigraphoid axiom is called perfect composition (Pearl 2014). Thus, for any four measurements X, Y, Z, and W:

Definition 4.2 **Perfect composition** requires that for any four measurements X, Y, Z, and W:

$$X \perp\!\!\!\perp (Y, Z) \mid W \Leftrightarrow X \perp\!\!\!\perp Y \mid (W, Z) \text{ and } X \perp\!\!\!\perp Z \mid W$$

Colloquially, this tells us that assuming W is known, then if neither Y nor Z provides additional information about X, then two statements are equivalent. Firstly, if two pieces of information Y and Z do not help us know X, then each one on its own also does not help model X. Secondly, if one of the two is given initially alongside W, the remaining piece of information still does not provide any additional information about X. Further axioms are recorded and proved in Pearl (2009). For the purposes of elicitation, these axioms prompt common language questions which can be posed to a domain expert to validate a graphical structure. Given the values of the vector of variables in Z, learning the values of Y would not help the prediction of X. Note that when we translate this statement into a predictive model, then this would mean that we know $p(x \mid y, z) = p(x \mid y)$.

BNs encode collections of irrelevance statements that translate into a collection of conditional independence relationships. This can be thought of as what variable measurements are irrelevant to another. Relationships of the form $X \perp\!\!\!\perp Y \mid Z$ can be read straight off the graph as missing edges indicate conditional independence relationships. BNs obey the global Markov property that each node is independent of its non-descendants given its parents (Pearl 2009). By identifying the non-descendants and parents of each node, the entire collection of independence relationships is readily apparent. To see this in our example, consider the node representing long-term health, H. In Figs. 4.1a, b, $\{B, I\}$ are its non-descendants, and F is its parent, so we know that $H \perp\!\!\!\perp B \mid F$ and $H \perp\!\!\!\perp I \mid F$.

The independences can be read from the graph using the d-separation criteria. We can determine the conditional independence between three sets of variables A, B, and S using d-separation. Investigating d-separation from the graph requires inspecting the moralized ancestral graph of all variables of interest, denoted as $(\mathcal{G}_{\mathrm{An}(A \cup B \cup S)})^m$ (Pearl 2009; Smith 2010). This includes the nodes and edges of the variables of interest and all their ancestors. Then, we moralize the graph, drawing an undirected edge between all pairs of variables with common children in the ancestral graph. After disorienting the graph (replacing directed edges on the graph with undirected ones) and deleting the given node and its edges, we can check conditional independence between variables of interest. If there is a path between the variables, then they are dependent on the BN; otherwise, they are independent.

Pearl and Verma (1995) proved the d-separation theorem for BNs, definitively stating the conditional independence queries that can be answered from the topology of the BN in Fig. 4.1a. The d-separation criteria and associated theorems formalize this process of reading off conditional independence relationships from a graph.

Theorem 4.1 *Let A, B, and S be disjoint subsets of a DAG $\mathcal{G}$. Then S d-separates A from B if and only if S separates A from B in $(\mathcal{G}_{An(A \cup B \cup S)})^m$, the moralized ancestral graph for the set.*

The proof is given in Steffen (1996).

As an example, consider the BN of the drivers of food insecurity shown in Fig. 4.1a. The d-Separation theorem tells us that H is d-separated from B and I given the separating set F. We see that in the moralized graph, F d-separates every path from the node H to a node in the set $\{B, I\}$. Thus, d-separation allows us to consider the relationships between any three subsets of variables in the DAG.

Separation theorems have been found for more general classes of graphs including chain graphs, ancestral graphs, and chain event graphs (Bouckaert and Studeny 1995; Andersson 2001; Richardson and Spirtes 2002). Another class of graphical model, vines, weakens the notion of conditional independence to allow for additional forms of dependence structure (Bedford and Cooke 2002). The results of the separation theorem for BNs can also be used to explore independence relationships in classes of graphs that are BNs with additional restrictions such as those imposed by the multi-regression dynamic models (Smith 1993) and flow graphs (Smith and Figueroa 2007).

When the structure is verified, it can then be embellished to a full probability model, provided it meets the original assumptions of our model. Understanding the relationship between the elicited conditional independence statements implied by the graph ensures that we do not elicit equivalent statements, thereby reducing the number of elicitation tasks. Even more importantly, the probabilities will respect the expert's structural hypotheses–hypotheses that are typically much more securely held than their numerical probability assessment.

In a discrete BN, this process involves populating the conditional probability tables with probabilities either elicited from experts or estimated from data. Alternatively, our food insecurity drivers' example could be embellished to a full probability representation of a continuous BN. Discrete BNs will be populated by conditional independence tables that assign probabilities to all possible combinations of the values of each term in the factorized joint probability density. New computational approaches for continuous BNs allow for scalable inference and updating of the BN in a high-dimensional, multivariate setting (Hanea et al. 2006). The probabilities underpinning this model can be elicited using additional protocols and procedures from other chapters of this book.

4.3 Examples from Food Insecurity Policy

4.3.1 Bayesian Network

Structural elicitation for a Bayesian network is well studied (Smith 2010; Korb and Nicholson 2009). To see this process in action, consider a food insecurity example. The United States Department of Agriculture (USDA) administers the national School Breakfast Program (SBP), serving free or reduced price meals to eligible students.

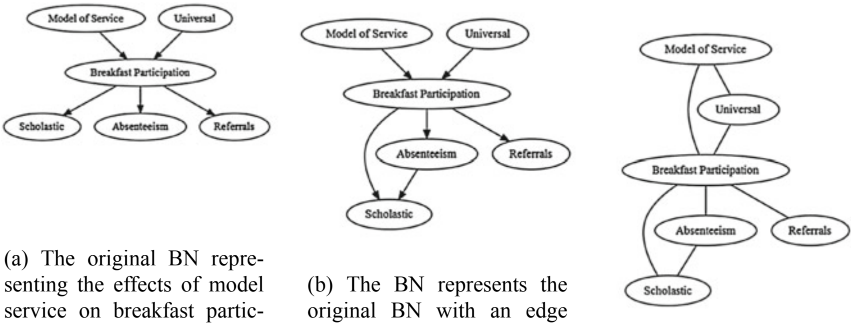
A key element of the system is understanding the programmatic operations. Participation in SBP is not as high as it is for the school lunch program (Nolen and Krey 2015). The traditional model of breakfast service involves students eating in the cafeteria before the beginning of school. Advocates began promoting alternative models of service to increase school breakfast participation. These include Grab n Go, in which carts are placed through the school hallways and students select a breakfast item en route to class, or Breakfast in the Classroom, where all students eat together during the first period of the day. Only schools which have 80% of students eligible for free or reduced lunch are eligible for universal school breakfast. This means that breakfast is offered to every child in the school, regardless of their free or reduced status. This policy was implemented to reduce stigma of receiving a free meal.

The experts would also like to understand the effects of not eating breakfast. Advocates, principals, and teachers have hypothesized that eating school breakfast impacts scholastic achievement. Food-insecure children struggle to focus on their studies. Schools also show a reduced rate in absenteeism, as children and parents have the added incentive of breakfast to arrive at school. Some evidence suggests eating breakfast may also reduce disciplinary referrals, as hungry children are more likely to misbehave.

The data for this problem comes from a set of schools who are all eligible for universal breakfast, but some have chosen not to implement the program while others have. As universal breakfast status can be used as a proxy for socio-economic background of students attending a school, we have narrowed the population to schools with low socio-economic status. The group of experts does not describe a temporal process here. They do not mention changes in breakfast participation throughout the school year, yearly fluctuations, or a time series of participation rates. Thus, it is natural for the modeller to begin with a BN approach. Given this information about breakfast, led by a facilitator, the modeller could consolidate the discussion into the following nodes:

- X_1 Model of Service (Yes, No): indicates whether or not an alternative model of service as been implemented.
- X_2 Universal (Yes, No): indicates whether or not an eligible school has opted into universal service, as opposed to checking the economic status of the student at each meal.
- X_3 Breakfast Participation (High, Medium, Low): the binned participation rates at each school.
- X_4 Scholastic Achievement (High, Medium, Low): the standardized test score for each school.
- X_5 Absenteeism (High, Low): the binned absenteeism rate for the year.
- X_6 Disciplinary Referrals (High, Low): absolute number of disciplinary referrals.

We note that this list of nodes is focused on understanding the effects of school breakfast participation and specific type of breakfast service model. Certainly, there are other reasons for absenteeism and disciplinary referrals besides whether or not a student had a good breakfast, but these are beyond the scope of this model. How do we determine the structure of this model from these measurable random variables? From this set of nodes, the decision-maker is queried about the possible relationship between all possible sets of edges. For instance, we could ask, does knowing whether or not the school has opted into universal breakfast give any other information about whether or not the school has implemented an alternative breakfast model? In this case, the decision-makers believe X_1 does not give any additional information about X_2, because the program model is subject to approval from the cafeteria managers and teachers, whereas the decision to implement universal breakfast is primarily the decision of the principal. Thus no edge is placed between X_1 and X_2. Both X_1 and X_2 are helpful in predicting X_3, so an arrow is drawn between each of these pairs. X_4 is affected by X_3. These relationships can be seen in Fig. 4.2a.



(a) The original BN representing the effects of model service on breakfast participation and academic outcomes

(b) The BN represents the original BN with an edge added through the described verification process.

(c) The ancestral, moralized DAG of the central BN.

Fig. 4.2 Exploring the conditional independence relationships expressed by the directed BN and its moralized analogue

It is important to note that if we had taken the population to be all schools rather than those with a low socio-economic status, then X_2 would affect X_4, X_5, and X_6 because universal school lunch would then be a proxy for low socio-economic status.

Suppose we know a school has a low breakfast rate, and we want information about their absenteeism. Will knowing anything about scholastic achievement provide any additional information about absenteeism? In order to check this with d-separation, we examine the ancestral graph $\mathscr{G}_{\mathrm{An}(X_4,X_5,X_6)}$, the moralized graph $(\mathscr{G}_{\mathrm{An}(X_4,X_5,X_6)})^m$ shown in Fig. 4.2c. If there is not a path between X_4 and X_5, then we can say that X_4 is irrelevant to X_5. However, if there is a path between X_4 and X_5 that does not pass through our given X_3, then the two variables are likely to be dependent. Thus, we can use the d-separation theorem to check the validity of the BN. We may also ask equivalent questions by symmetry. For instance, suppose we know a school has a low breakfast rate and we want to know information about their scholastic achievement. Will additional information about absenteeism be relevant to scholastic achievement? Asking such a question may prompt our group of decision-makers to consider that students who miss classes often perform worse on exams. Revising the BN is in order, so we add an additional edge from X_5 to X_4. The BN in Fig. 4.2a represents the beliefs of the domain experts. This encodes the following irrelevance statements:

- Knowing the model of service provides no additional information about whether or not the school district has implemented universal breakfast.
- The model of service provides no additional information about scholastic achievement, absenteeism, or referrals given that we know what the percentage of students who eat breakfast is.
- Knowing absenteeism rates provides no additional information about disciplinary referrals given that we know what the breakfast participation rate is.
- Knowing scholastic achievement rates provides no additional information about disciplinary referrals given that we know what the breakfast participation and absentee rates are.

When these irrelevance statements are checked, the domain experts realize that there is an additional link in that absenteeism affects scholastic achievements. Thus we draw an additional arrow between X_4 and X_5 as shown in Fig. 4.2b. The relationship between referrals and absenteeism is disputed in the literature and among experts, so, at least in this first instance, we omit this edge.

Once the experts agree on the structure and verify it using the irrelevance statements, then the modeller may elicit the conditional distributions. Taken together, the BN represents a series of local judgements. The joint probability mass function of a BN is represented by

$$p(\boldsymbol{x}) = \prod_{i=1}^{n} p(x_i|\mathrm{pa}(x_i))$$

where $\mathrm{pa}(x_i)$ indicates the parent set of x_i. For our example,

$$p(\boldsymbol{x}) = p(x_1)p(x_2)p(x_3|x_1, x_2)p(x_4|x_3, x_5)p(x_5|x_3)p(x_6|x_3)$$

Many of these distributions may be estimated by data, and unknown quantities may be supplied through structured expert elicitation. For instance, consider the sample question: what is the probability that scholastic achievement is high given that breakfast participation rate is medium and the absentee rate is low? When the conditional probability tables are completed, the BN can be used to estimate effects of intervention in the system according to Pearl (2009).

4.3.2 Chain Event Graph

To illustrate an instance when a bespoke representation is more appropriate than the BN, consider the example of obtaining public benefits to address food insecurity. The USDA's Supplemental Nutrition Assistance Program (SNAP) provides funds for food to qualifying families and individuals through Electronic Benefit Transfer (EBT). Although 10.3% of Americans qualify for the program, Loveless (2010) estimates that many more citizens are eligible for benefits than actually receive them. Policy-makers and advocates want to understand what systemic barriers might prevent eligible people from accessing SNAP. The application process requires deciding to apply, having sufficient documentation to apply (proof of citizenship, a permanent address), a face-to-face interview, and correct processing of the application to receive funds.

The structural elicitation phase includes speaking with domain experts to gather a reasonably comprehensive list of steps in the process. Domain experts include caseworkers, advocates, and individuals applying through the system. For our example, Lara et al. (2013) collected this information through interviews at 73 community-based organizations in New York State and categorized it according to access, eligibility, and benefit barriers. This qualitative information collection is crucial to developing an accurate model. From the qualitative studies, the key barriers were identified as

- Face-to-face interviews not waived
- Same-day application not accepted
- Excessive documentation required
- Expedited benefit (available to households in emergency situations) not issued
- Failed to receive assistance with application documents
- Barriers experienced by special population: elderly and immigrant
- Ongoing food stamp not issued within 30 days
- EBT card functionality issues.

The events selected should be granular enough to encompass the key points at which an applicant would drop out of the process, but coarse enough to minimize model complexity. An important part of the qualitative analysis process includes combining anecdotal evidence into similar groupings. For instance, the benefits office refused to waive the in-office interview for an applicant who did not have transportation to the application centre. In a separate instance, an interview was not waived for a working single mother with four children who could not attend because she was at work. While there are different contexts to each example, the central problem is the failure to waive the face-to-face interview. This type of node consolidation aids in reducing model complexity.

Discretising events can be a convenient way to clarify the model structure. Checking that the discretization covers all possible outcomes from that event ensures that the model is an accurate representation of the problem. For our example, one possible discretization with four variables of the problem is as follows:

- X_r: At-risk population? (Regular, Elderly, Immigrant)
  - Regular: Households not part of an at-risk population
  - Elderly: Household head is over 65
  - Immigrant: Household head is a citizen, but immigration status of members of the household is uncertain
- X_a: Decision to apply (Expedited, Regular application, Decides not to apply)
  - Expedited: Same-day applications, used in cases of emergency food insecurity
  - Regular application: The standard procedure
  - Decides not to apply: Eligible households who elect not to apply for a variety of reasons
- X_v: Application Verdict (Rejected, Accepted, Revision Required)
  - Rejected: Failed application, no possibility of resubmission
  - Accepted: Successful application
  - Revision required: Application must be resubmitted because of missing documentation, missed interview, or other reasons

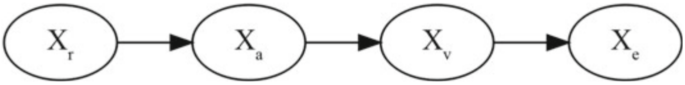


Fig. 4.3 An inadmissible BN for the public benefits application process example

- X_e: Utilizing an EBT card (Card successfully used for transactions, transaction errors)
  - Card used for transactions: EBT arrives within the 30-day deadline and is successfully used at a grocery store
  - Transaction errors: Card either does not arrive or returns an error at the grocery store

Figure 4.3 shows a simple BN approach to the natural language problem. Assume that the conditional independence relationships have been checked and that we can now supply the conditional probabilities. As we begin this process, note that some of the probabilities are nonsensical. For example, we must supply a probability for quantities like the probability of having an accepted application given that the eligible citizen decided not to apply, and the probability of successfully utilizing EBT given that the application was rejected. This probability setting sounds absurd to elicit structurally and will be distracting during the probability elicitation.

The application process is difficult to coerce into a BN because the problem is highly asymmetrical. For instance, applicants with insufficient documentation will not have the chance to interview and will not progress through the system. Now, if we consider again the natural language of the experts, we notice that this process is described as a series of events that have a natural ordering. Applicants must first decide to apply, then receive a verdict, and finally use their EBT card. The notion of being a member of an at-risk population does not have an explicit ordering, but we can reasonably order it before the other events as it may affect how downstream events unfold.

Collazo et al. (2017) shows that ordering demographic information at the beginning often coincides with higher scoring models during model selection for this class of graphs. Shafer (1996) has argued that event trees are a more natural way to express probabilistic quantities, so we will instead express this problem as an event tree in Fig. 4.4. We next show how, in this instance, there is an alternative graphical frame-

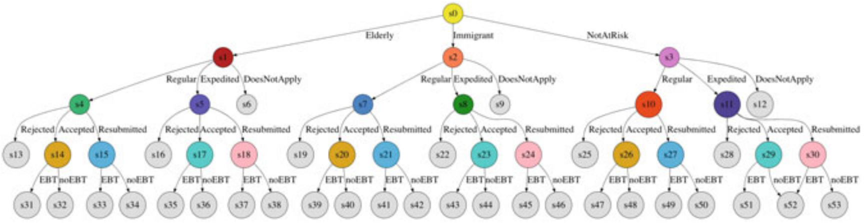

s0
Elderly
Immigrant
NotAtRisk
s1
s2
s3
Regular Expedited DoesNotApply
Regular Expedited DoesNotApply
Regular Expedited DoesNotApply
s4
s5
s6
s7
s8
s9
s10
s11
s12
Rejected Accepted Resubmitted
s13 s14 s15 s16 s17 s18 s19 s20 s21 s22 s23 s24 s25 s26 s27 s28 s29 s30
EBT noEBT
s31 s32 s33 s34 s35 s36 s37 s38 s39 s40 s41 s42 s43 s44 s45 s46 s47 s48 s49 s50 s51 s52 s53

Fig. 4.4 Event tree depicting the outcomes of the benefit application process

work that provides a better way of accommodating the information provided by the expert.

The nodes of our event tree are called situations $s_i \in S$ indexed according to temporal precedence; they represent different outcomes faced by applicants travelling through the system. The edges represent the probabilities of different outcomes of each possible event occurring. We can elicit the probability of observing a unit travelling down each edge of the tree. The probability of a unit travelling down each of those edges should sum to one for each situation. The root-to-sink paths on the tree can be thought of as all possible outcomes of the application procedure. Situations with the same colour on the tree represent events whose outcomes have the same probabilities. In Fig. 4.5, leaf nodes showing terminating outcomes are depicted in light grey.

The tree structure is naturally flexible just like the BN and can easily be modified to accommodate natural language suggestions. For instance, suppose the expert would like to add in a variable: the outcome of an interview process for regular applicants (the expedited process is waived.) Adapting the model simply requires adding two edges representing the outcome of the interview being successful or rejected to the set of situations in which an applicant applies through the regular route $\{s_4, s_7, s_{10}\}$. This simple adjustment in the tree structure would require adding a node to the BN as well as updating the conditional probability tables for the children nodes.

Another feature of the event tree structure is that the context-specific independences are expressed directly in the tree structure. In this example, elderly applicants are often less likely to apply for benefits because the dollar amount is often too small a motivation for the perceived difficulty of the application. Immigrants are also less likely to apply because, although citizenship is required to apply for benefits, citizens with undocumented family members may fear citizenship repercussions of applying for assistance.

These context-specific probabilities are modelled through the colourings of the positions of the Chain Event Graph (CEG), rather than requiring separate BN models with context-specific conditional independence relationships. To read conditional independence relationships from the graph, we begin by saying that two situations are in the same stage $u_j \in U$ if they are the same colour. In order to draw a condensed representation of the graph, we define positions $w_k \in W$ as sets of situations that have the same colour and the same downstream sub-trees. This allows us to merge stages for a more compact chain event graph representation, called the Chain Event Graph (CEG), depicted in Fig. 4.5.

In the same spirit as the Markov condition for BNs, we can read statements of the form 'the future is independent of the past given the present'. Given that a unit reaches a position, what happens afterwards is independent not only of all developments through which it was reached, but also of the positions that logically cannot happen. These conditional independence statements can be read off the graph just as they can for BNs. To illustrate this process, we need new definitions that identify the sets of positions in the graph.

Formally, let $\mathscr{C} = (W, E)$ denote a coloured CEG on a set of positions W and edges E. Then, we call a set of positions $W' \subseteq W$ a fine cut if disjoint union of events

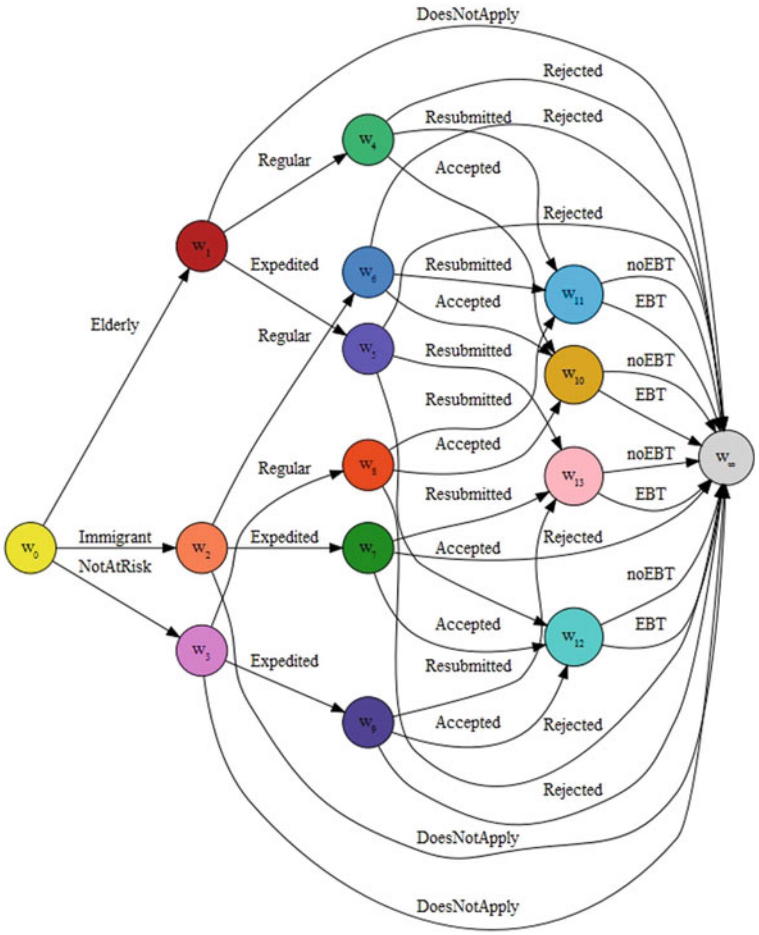


Fig. 4.5 Chain Event Graph representation of the benefits application process

centred on these vertices is the whole set of root-to-leaf paths. That is, none of the positions $w \in W'$ is up- or downstream of another, and all of the root-to-sink paths on $\mathscr{C}$ must pass through one of the positions in W'.

Furthermore, a set of stages $u \in U$ denoted $W' \subseteq U$ is a cut if the set of positions in the colouring $w \in u|u \in U$ is a fine cut. The definitions of fine cut and cut help us to differentiate the "past" from the "future" in the graph.

A cut-variable denoted X_W can be thought of as an indicator variable used to define the present. Formally, X_W is the corresponding set of positions, W in a cut or a fine cut, and X_W is measurable with respect to the probability space defined by the CEG.

Then, we can define a vector of random variables whose vertices are located upstream or downstream. Denote the "past" random variables as $Y_{\prec W} = (Y_w | w$

upstream of W) and the “future” by $Y_{W\preceq} = (Y_{w'} | w'$ downstream of $W)$. Then we can formally define the conditional independences in a CEG:

Theorem 4.2 *Let $\mathscr{C} = (W, E)$ be a CEG and let $W' \subseteq W$ be a set of positions, then for any cut-variable $X_{W'}$, we find*

1. *If W' is a fine cut then $Y_{\prec W'} \perp\!\!\!\perp Y_{W'\preceq} | X_{W'}$.*
2. *If W' is a cut then $Y_{\prec W'} \perp\!\!\!\perp Y_{W'} | X_{W'}$.*

Proof can be found in Smith and Anderson (2008).

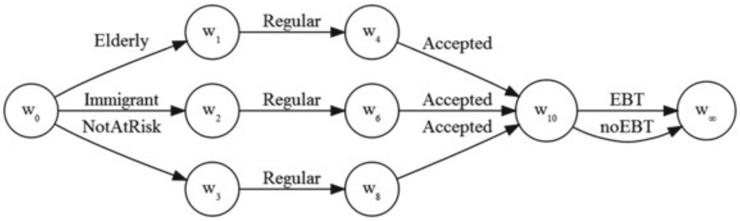
Theorem 4.2 explains how to read conditional independence from the CEG structure. The next step is to validate the structure. Just as for the BN, natural language questions from the semigraphoid axioms elucidate the conditional independence relationships. At each cut, consider the conditional independence between each pair of upstream and downstream variables. For instance, given that eligible applicants apply for benefits, does knowing whether or not they are part of an at-risk population provide any additional information about whether or not they apply for expedited benefits? By perfect decomposition, does knowing that the candidate received application assistance provide any information about whether or not they will receive the electronic benefits given that they had the correct documentation and passed the interview? Does knowing that they had application assistance provide any additional information about whether or not they passed the interview given that they had the correct documentation? These queries validate the model and may prompt further adaptations.

In the BN, Theorem 4.1 provides a systematic way to check all of the conditional independence relationships. Thwaites and Smith (2015) proposed a new d-separation theorem for CEGs. In a BN, the ancestral graph helps to address these queries. The ancestral graph has no direct analogue in the CEG. Instead, following Thwaites and Smith (2015) the CEG admits a pseudo-ancestral representation. Pseudo-ancestral graphs depict the nodes of interest and all the upstream variables, consolidating the downstream variables. Moralizing the graph in a BN corresponds to removing the colourings of the CEG.

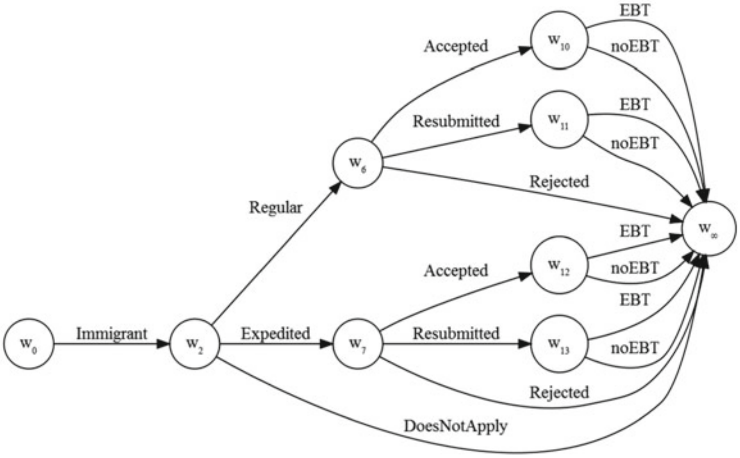
Is the ability to complete a transaction on the EBT card independent of whether the applicant is a member of an at-risk population given that they completed a successful regular application? The pseudo-ancestral graph as shown in Fig. 4.6a shows the probability that $\Lambda = \{\text{Regular, Accepted}\}$. Being a part of the at-risk population is independent of being able to utilize an EBT card because we see that all the possible pathways must pass through w_{10}, identifying it as a single vertex composing a fine cut.

On the other hand, suppose we want to test the independence of the application verdict from the selected method of application for at-risk immigrant population. This ancestral graph can be given by the CEG in Fig. 4.6b. These are not independent because there is no single vertex composing a fine cut.

One of the strengths of the CEG model is that it does not require any algebra, but instead can be elicited entirely using coloured pictures. CEGs are of particular use for problems that exhibit some asymmetry. After validating the structure, populating the model with data or elicited probabilities provides a full statistical model that can



(a) A pseudo-ancestral CEG representing an independence between the query.



(b) A pseudo-ancestral CEG representing a dependence between the query.

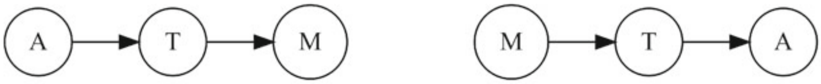
Fig. 4.6 Two uncoloured pseudo-ancestral CEGs

be used for inference, details can be found in Collazo et al. (2017). The CEG offers a class of models that is more general than BNs, enabling modellers to represent context-specific independences.

The CEG is a powerful model particularly well suited to expert elicitation, as experts often convey information in a story, which naturally expands to an event tree.

4.3.3 Multi-regression Dynamic Model

Our next two examples of customized classes of graphical models consider the problem of assessing participation in the Summer Meals Program (SMP). SMP meal sites are designated as either open or closed. Open sites do not have a set population like in



(a) The correct summary MDM graph (b) An inadmissible MDM summary graph

Fig. 4.7 Two DAGs with equivalent BN representations, but unique multi-regression dynamic model representations

a school or particular program, but rather are open to the public and thus dependent on walk-ins for the bulk of participation.

Although the need in the summer is severe, participation in the program remains relatively low. Advocates generally agree that the two biggest obstacles to program participation are a lack of awareness about the program and unavailable transportation to the site. These factors affect meal participation which fluctuates throughout the 3 months of summer holidays. Available data for meal participation records how many meals were served through the program each day for about 3 months in the summer. Transportation data records the number of available buses. Awareness can be measured through testing data that records when participants queried a government information line to receive information about where the closest sites serving meals are.

Advocates would most like to capture the effect that awareness of SMP has on available transportation, and that transportation in turn has on meal participation. To simplify the elicitation, additional obstacles like low summer school enrolment, poor food quality, and insufficient recreational actives are not considered as primary drivers of meal participation levels. The relationship between awareness and available transportation is well documented, as is the relationship between transportation and meal participation (Wilkerson and Krey 2015).

The advocates emphasize drastic shifts in awareness, transportation, and meal participation throughout the summer months. On public holidays and weekends, there is a lack of public transportation and a corresponding sharp decline in meals. This temporal aspect of the problem prompts the modeller to consider a time-series representation as the most natural class of graphical model.

To emphasize the importance of selecting a time-series representation over a BN, consider the limitations of the standard BN model. Suppose the advocates agree on the general structure shown in the DAG in Fig. 4.7a, as children and parents must know about the meal before they take transportation to the meal. Then, in turn, they must travel to the meal before receiving the meal. However, if the graph is interpreted as a BN, then Fig. 4.7a only encodes the conditional independence relationship $M \perp\!\!\!\perp A \mid T$, which does not capture the ordering expressed by the advocates. To further stress this point, Fig. 4.7 shows a DAG with the reverse ordering that encodes equivalent conditional independence relationships when interpreted as a BN. As we will see below, if these are summary graph of MDMs whose edges represent the strengths given in the model definition in Eq. 4.3, then the models are distinguishable.

The experts remark that a media campaign and corresponding surge in awareness prompts a corresponding increase in the number of people travelling to meal sites. These aspects of the problem, taken with those discussed above prompt us to consider each of the elements as time series. In order to capture the linear relationship between variables that the experts have expressed, we also define the edges of the graph to correspond to regression coefficients between each parent and child.

Assuming linear relationships exist between awareness and transportation and transportation to the meal site and actual participation, the system can be described as regressions in a time-series vector $\boldsymbol{Y}_t = \{Y_t(1), Y_t(2), Y_t(3)\}$. We denote the time series of the key measurements: awareness by $Y_t(1)$, available transportation by $Y_t(2)$, and summer meals participation by $Y_t(3)$. This model corresponds to another example from our toolbox of alternative representations: the multi-regression dynamic model, the general definition of which is shown below.

Definition 4.3 A collection of time series $\boldsymbol{Y}_t = \{Y_t(1), \dots, Y_t(i), \dots, Y_t(n)\}$ can be considered a the multi-regression dynamic model (MDM) if the observation equations, system equation, and initial information as given below adequately describes the system. Each series in the MDM can be represented by an observation equation of the form:

$$Y_t(r) = \boldsymbol{F}_t(r)'\boldsymbol{\theta}_t(r) + v_t(r) \qquad v_t(r) \sim (0, V_t(r)),\ \ 1 \leq r \leq n$$

where $\theta_t = \{\theta_t(1), \dots, \theta_t(n)\}$ are the state vectors determining the distribution of $Y_t(r)$. $\boldsymbol{F}_t(r)$ is a known function of $\boldsymbol{y}_t(r)$ for $1 \leq r$, that is, each observation equation only depends on the past and current observations rather than the future ones. $V_t(r)$ are known as scalar variance observations. These can be estimated from available data or else elicited from experts. The indexing over r encodes the strict ordering of the nodes that is so key to this problem.

The system equation is given by

$$\boldsymbol{\theta}_t = G_t\boldsymbol{\theta}_{t-1} + \boldsymbol{w}_t \qquad \boldsymbol{w}_t \sim (\boldsymbol{0}, \boldsymbol{W}_t)$$

where $\boldsymbol{G}_t = \text{blockdiag}\{G_t(1), \dots, G_t(n)\}$. Each $G_t(r)$ represents a $p_r \times p_r$ matrix. For a linear MDM we can use in this example, we can take $\boldsymbol{G}_t$ to be the identity matrix. The term $\boldsymbol{w}_t$ represents the innovations of the latent regression coefficients, that is, the difference between the observed and forecasted values. $\boldsymbol{W}_t = \text{blockdiag}\{W_t(1), \dots, W_t(n)\}$, where each $W_t(r)$ has dimensions $p_r \times p_r$, where p_r is the number of parent of $Y_t(r)$. Lastly, the initial information is expressed as

$$(\boldsymbol{\theta}_0|y_0) \sim (\boldsymbol{m}_0, C_0)$$

where $\boldsymbol{m}_0$ is a vector of mean measurements of the observation and C_0 is the variance–covariance matrix where $C_0 = \text{blockdiag}\{C_0(1), \dots, C_0(n)\}$.

This means that $(\boldsymbol{Y}_t(r)|\boldsymbol{Y}^{t-1}, \boldsymbol{F}_t(r), \boldsymbol{\theta}_t(r))$ follows some distribution with mean $\boldsymbol{F}_t(r)^t\boldsymbol{\theta}_t(r)$ and variance $V_t(r)$.

Modelling this behaviour requires dynamic linear models in which the parents are the regression coefficients for each series. For our example in Fig. 4.7a, the system and observation model equations are

$$\begin{aligned}
\boldsymbol{\theta}_t(1) &= \boldsymbol{\theta}_{t-1}(1) + w_t(1) & Y_t(1) &= \theta_t^{(1)}(1) + v_t(1)\\
\boldsymbol{\theta}_t(2) &= \boldsymbol{\theta}_{t-1}(2) + w_t(2) & Y_t(2) &= \theta_t^{(1)}(2) + \theta_t^{(2)}(2)Y_t(1) + v_t(2)\\
\boldsymbol{\theta}_t(3) &= \boldsymbol{\theta}_{t-1}(3) + w_t(3) & Y_t(3) &= \theta_t^{(1)}(3) + \theta_t^{(2)}(3)Y_t(2) + v_t(3)
\end{aligned}$$

The strengths of the parents are given by the regression coefficients $\theta_t^{(2)}(2)$ for $Y_t(2)$ and $\theta_t^{(2)}(3)$ for $Y_t(3)$. The initial information $\{\boldsymbol{\theta}_0\}$ can be elicited from the domain experts or taken from previous data observations.

Suppose after the experts agree on the structure, the modeller examines the one-step ahead forecasts and notices errors on some days. Examining these days might prompt the experts to recognize that the days of interest correspond to days with a heat advisory. They suggest that the heat index throughout the summer also affects meal participation. This structural change can be quickly integrated into the system by adding observation and system equations and initial information for and updating the system for downstream node. Because the ordering in the MDM is strict, and the heat index is a parent of meal participation, we will relabel meal participation as $Y_t(4)$ and the heat index as its parent $Y_t(3)$.

$$\begin{aligned}
&\boldsymbol{\theta}_t(4) = \boldsymbol{\theta}_{t-1}(4) + w_t(4)\\
&Y_t(3) = \theta_t^{(1)}(3) + v_t(3)\\
&Y_t(4) = \theta_t^{(1)}(4) + \theta_t^{(3)}(4)Y_t(2) + \theta_t^{(2)}(4)Y_t(3) + v_t(3)
\end{aligned}$$

In this new model, the regression coefficients $\theta_t^{(3)}(4)$ and $\theta_t^{(2)}(4)$ for meal participation $Y_t(4)$ indicate the strengths of the edges in the summary graph in Fig. 4.8.

In this way, the natural language expressions of the domain experts can be used to adjust the model.

Generally, particular observations of $Y_t(r)$ are denoted as $y_t(r)$. The MDM ensures two critical conditional independence relationships. The first holds that if

$$\perp\!\!\!\perp_{r=1}^{n} \boldsymbol{\theta}_{t-1}(r)|\boldsymbol{y}^{t-1} \tag{4.1}$$

then

$$\perp\!\!\!\perp_{r=1}^{n} \boldsymbol{\theta}_{t}(r)|\boldsymbol{y}^{t} \tag{4.2}$$

where $\boldsymbol{y}^{t-1}(i) = \{y_1(i), \ldots, y_{t-1}(i)\}$ and

$$\boldsymbol{\theta}_t(r) \perp\!\!\!\perp Y^t(r+1), \ldots, Y^t(n) | Y^t(1), \ldots, Y^t(r) \tag{4.3}$$

Equation (4.2) tells us that if the parameters$\{\boldsymbol{\theta}_{t-1}(r)\}$ are independent of each other given the past data $\{\boldsymbol{y}^{t-1}\}$ then $\{\boldsymbol{\theta}_t(r)\}$ is also independent of $\{\boldsymbol{y}^t\}$. By induction, we can see that given the initial parameters $\{\boldsymbol{\theta}_0(r)\}$ are independent, then they remain independent as the series unfolds.

For the beginning example, we need to ensure that $\boldsymbol{\theta}_0(1) \perp\!\!\!\perp \boldsymbol{\theta}_0(2) \perp\!\!\!\perp \boldsymbol{\theta}_0(3)$. Awareness is measured by the amount of public media generated, transportation is a measure of public transportation available, and the participation rate is the number of meals served every day in the summer. The domain experts agree that these can be independent of each other. Additionally, Eq. 4.3 ensures the following conditional independence relationships:

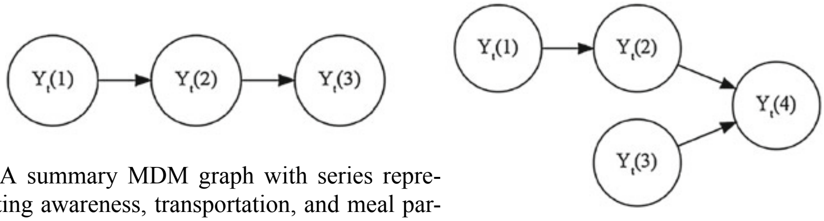
$$\boldsymbol{\theta}_t(1) \perp\!\!\!\perp \{y^{t-1}(2), y^{t-1}(3)\} | y^{t-1}(1)$$
$$\boldsymbol{\theta}_t(2) \perp\!\!\!\perp y^{t-1}(3) | \{y^{t-1}(1), y^{t-1}(2)\}$$

An analogue of the d-separation theorem for MDMs identifies part of the topology of the graph that ensures that these conditional independence statements hold.

Theorem 4.3 *For MDM* $\{\boldsymbol{Y}_t\}$ *if the ancestral set* $\boldsymbol{x}_t(r) = \{y^t(1), \ldots, y^t(r)\}$ *d-separates* $\boldsymbol{\theta}_t(r)$ *from subsequent observations* $\{y^t(r+1), \ldots, y^t(n)\}$ *for all* $t \in T$, *then the one-step ahead forecast holds :*

$$p(\boldsymbol{y}_t | \boldsymbol{y}^{t-1}) = \prod_r \int_{\boldsymbol{\theta}_t(r)} p\{\boldsymbol{y}_t(r) | \boldsymbol{x}^t(r), \boldsymbol{y}^{t-1}(r), \boldsymbol{\theta}_t(r)\}\, p\{\boldsymbol{\theta}_t(r) | \boldsymbol{x}^{t-1}(r), \boldsymbol{y}^{t-1}(r)\} d\boldsymbol{\theta}_t$$

This one-step ahead forecast factorizes according to the topology of the graph, allowing us to examine the plots of each of the series. For our example, the one-step ahead forecast factorizes:



(a) A summary MDM graph with series representing awareness, transportation, and meal participation respectively.

(b) A MDM summary graph with additional series for heat index.

Fig. 4.8 A summary MDM graph after refining elicitation with experts including the original variables plus a new series with the heat index A MDM summary graph with additional

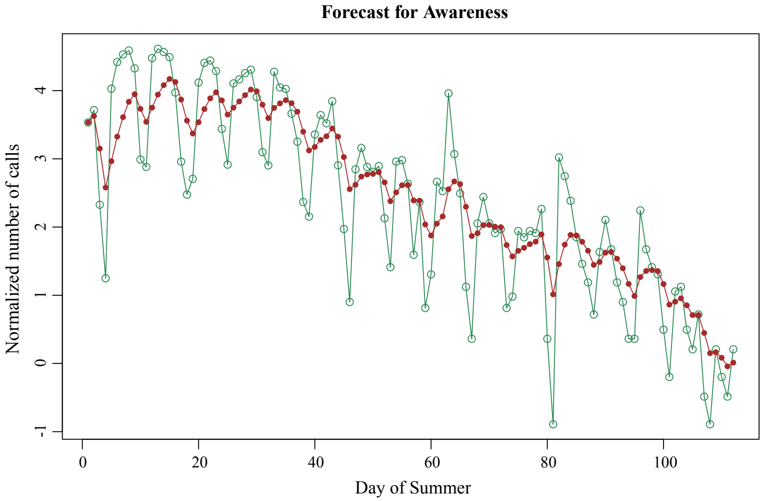


Fig. 4.9 The logarithmic plot of awareness (as measured by calls to ask for meal site locations) throughout the summer months. The open green dots are actual observations; the filled brown dots are the one-step ahead forecast

$$
\begin{aligned}
p(\boldsymbol{y}_t|\boldsymbol{y}) = &\int_{\boldsymbol{\theta}_t(1)} p\{\boldsymbol{y}(1)_t|\boldsymbol{y}^{t-1}(1)\boldsymbol{\theta}_t(1)\}\, p\{\boldsymbol{\theta}_t(1)\}d\boldsymbol{\theta}_t(1) \\
\times &\int_{\boldsymbol{\theta}_t(2)} p\{\boldsymbol{y}(2)_t|\boldsymbol{y}(1)^t, \boldsymbol{y}^{t-1}(2), \boldsymbol{\theta}_t(2)\}\, p\{\boldsymbol{\theta}_t(2)|\boldsymbol{y}^{t-1}(1), \boldsymbol{y}^{t-1}(2)\}d\boldsymbol{\theta}_t(2) \\
\times &\int_{\boldsymbol{\theta}_t(3)} p\{\boldsymbol{y}(3)_t|\boldsymbol{\theta}_t(1), \boldsymbol{\theta}_t(2), \boldsymbol{y}^{t-1}(3), \boldsymbol{\theta}_t(3)\}\, p\{\boldsymbol{\theta}_t(3)|\boldsymbol{y}^{t-1}(1), \boldsymbol{y}^{t-1}(2), \boldsymbol{y}^{t-1}(3)\}d\boldsymbol{\theta}_t(3)
\end{aligned}
$$

Examining plots of the errors of each forecast can help determine what further structural adjustments should be made. For instance, in Fig. 4.9, awareness has a cyclical nature, as people are less likely to text for an address of a meal site on weekends and holidays. This model can be adapted to include seasonal shifts using the equations from West and Harrison (1997).

The implementation of this problem as an MDM rather than a BN maintains the strength of the relationships between each series and its regressors, respecting the natural language expression of the system by the domain experts. An additional feature of the MDM is that this representation renders the edges causal in the sense carefully argued in Queen et al. (2009). For our model, note that while the two DAGs in Fig. 4.7 both represent $A_t \perp\!\!\!\perp M_t$, and are thus indistinguishable, the arrows in the MDM representation are unambiguous. The MDM offers a dynamic representation of a system in which the regressors influence a node contemporaneously.

4.3.4 Flow Graph

Structures can be adapted to meet additional constraints, such as conservation of a homogeneous mass transported in a system. However, these constraints motivate employing yet another graph with different semantics to transparently express the expert structural judgements. To illustrate how we might derive this from a natural language expression of a problem, consider the following example from the Summer Meals Program (SMP).

SMP provides no-cost meals to children under 18 at schools and community-based organizations during the summer months. SMP relies on food being procured from vendors, prepared by sponsors, and served at sites. Participation in the program is low, nationally 15% percent of eligible children use the program (Gundersen et al. 2011). Sponsors, entities who provide and deliver meals, are reimbursed at a set rate per participant, but sponsors often struggle to break even. One of the key possible areas for cost cutting is the supply chain of the meals. Community organisers hypothesize different interventions on each of these actors might help make the program more sustainable such as follows:

- A school district serving as a sponsor (Austin ISD) is having trouble breaking even. What happens when they partner with an external, more financially robust sponsor (City Square) to provide meals to the school. What is the effect on the supply chain of meals to the Elementary and Intermediate schools?
- Several smaller sponsors (among them the Boys and Girls Club) are having trouble breaking even and decide to create a collective to jointly purchase meals from a vendor (Revolution Foods). How does the presence of the new collective alter the flow of meals to the two Boys and Girls Club sites?
- Two sites, say apartment complexes A and B, are low-performing, and the management decides to consolidate them. What is the long-term effect on a system?
- What happens when a sponsor, City Square, changes vendors from Revolution Foods to Aramark?
- What happens when one sponsor, Austin ISD, no longer administers the program and another sponsor, Boys and Girls Club takes responsibility for delivering food to the Intermediate and high Schools?

Hearing the domain expert describe what types of intervention they would like to be able to model can elucidate the critical elements of the structure. In this example, the effect of the supply and transportation of meals through the network is key to the types of behaviour the modeller hopes to capture. This problem can be framed as a set quantity of meals moving through the system. Key model assumptions must always be checked with the domain expert. In this case, one of the key assumptions is that the number of children who are in need of meals and are likely to attend the program is relatively stable throughout the summer. This is a reasonable assumption, particularly when modelling a set population such as students in summer school or extracurricular programming. Community advocates verify that the assumption is reasonable because all of these sites and sponsors need a relatively set population in order to break even on the program.

Additionally, to estimate the effect of the addition or removal of actors in the system, it is important to assume that the number of meals for children in need is conserved. Thus, if a sponsor and subsequent sites leave the program, then those children will access food at another sponsor's meal sites, provided transportation is available. This assumption allows us to model particular interventions of interest, where combining, removing, or adding actors to the system is of particular interest. The dynamics of this particular problem involve the switching of ownership—what happens when the path flow of meals through the system changes—either a sponsor buys a meal from a different vendor, or a site turns to a different sponsor to supply their meals. This is a key component of the problem, but unfortunately it renders a key component of the problem intractable for the BN as shown below. However, Smith and Figueroa (2007) discovered a methodology for re-framing this problem as a tractable variant of a BN that simultaneously remains faithful to the dynamics of the problem described above.

If we began modelling the process as a BN, we might begin by first identifying the actors involved. A scenario for the key players in the city of Austin, Texas may consist of the following players at the vendor, sponsor, and site level. Levels are denoted by $z(i, j)$ where i indicates the level (vendor, sponsor, or site), and j differentiates between actors on a particular level. In this example, the players are

- $z(1, 1)$ Revolution Foods
- $z(1, 2)$ Aramark
- $z(2, 1)$ City Square
- $z(2, 2)$ Austin Independent School District
- $z(2, 3)$ Boys and Girls Club
- $z(3, 1)$ Apartment complex A
- $z(3, 2)$ Apartment complex B
- $z(3, 3)$ Elementary School
- $z(3, 4)$ Intermediate School
- $z(3, 5)$ High School
- $z(3, 6)$ Boys and Girls Club site A
- $z(3, 7)$ Boys and Girls Club site B.

These actors compose the nodes of the network; the edges represent the flow of meals between entities. For instance, vendor Aramark $z(1, 2)$ prepares meals for sponsors at Austin ISD, $z(2, 2)$, who in turn dispenses them at the Intermediate School, $z(3, 4)$. We assume that each day, a set number of meals runs through the system. This list of actors can be readily obtained from natural language descriptions of the problem. Eliciting this information would simply require the modeller to ask the domain experts to describe the flow of meals through each of the actors in the system. This structural elicitation and resultant graph in Fig. 4.10 are transparent to the expert, an advantage of customized modelling.

As the modeller begins to check the relationships encoded in the graphical model elicited in Fig. 4.10, the missing edges between actors in a given level means that

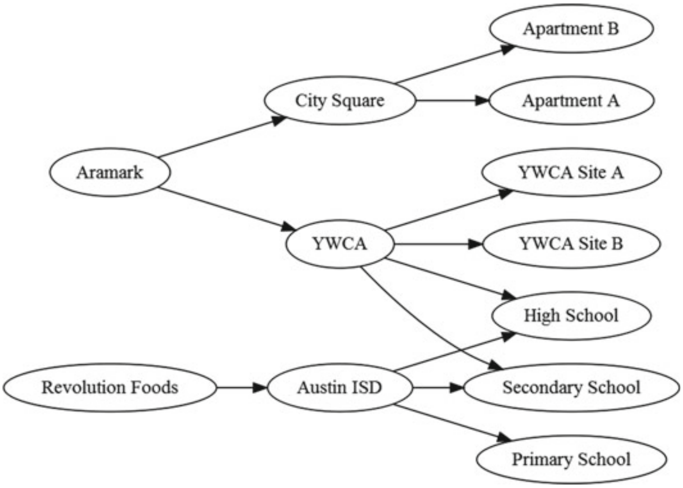


Fig. 4.10 Flow graph showing transfer of meals from vendors $z(1, j)$, to sponsors $z(2, j)$, to sites $z(3, j)$

each of the sponsors is unaffected by the meals being transported to and from the other sponsors. However, this is not realistic for closed sites because the experts have told us that knowing the number of meals served at all but one sponsor gives us perfect information about the remaining sponsor, as we know the number of meals served by sponsors remains constant! For instance, if we know how many meals are prepared by Aramark, $z(1, 1)$, then we have perfect information about how many are prepared by Revolution Foods, $z(1, 2)$, because meals are conserved at each level, implying a directed line from $z(1, 1)$ to $z(1, 2)$. Modelling this process graphically, as in Fig. 4.10, induces severe dependencies in the network if we consider the graph to be a BN. Thus, the problem as the experts have expressed it cannot be represented as a BN.

By decomposing the information in Fig. 4.10 into paths as shown in Smith and Figueroa (2007), we can apply the methodology of dynamic Bayesian networks. Denote $\boldsymbol{\phi}_t'[l] = (\phi_t(l, 1), \phi_t(l, 2), \ldots, \phi_t(l, n_l))$, where $l = \{1, 2, 3\}$ as the node states vector for each of the three levels, where $\phi_t(l, j_l)$ represents the mass owned by player $z(l, j_l)$ during time t. This probabilistic representation allows the modeller to retain the advantages of the clear representation in Fig. 4.10 to draw information about the system from the experts as well as the computational convenience of the BN machinery.

The full methodology for translating the hierarchical flow graph to the dynamic Bayesian Network (DBN) representation is given in Smith and Figueroa (2007); here, we simply state the elements of the model that would need to be a part of the probability elicitation. Information about the numbers of meals held by each entity at each day during the summer can be represented by a time-series vector $\boldsymbol{X}_t' = (\boldsymbol{X}_t'[1], \boldsymbol{X}_t'[2], \boldsymbol{X}_t'[3])$, representing the number of meals at the vendor, sponsor, and site levels, respectively. Next, we represent the paths of meals travelling from vendor to meal site as aggregates of the product amounts. The paths in this diagram are

$$\begin{aligned}
&\pi(1) = \{z(1,1), z(2,1), z(3,2)\} && \pi(2) = \{z(1,1), z(2,1), z(3,1)\}\\
&\pi(3) = \{z(1,1), z(2,3), z(3,6)\} && \pi(4) = \{z(1,1), z(2,3), z(3,7)\}\\
&\pi(5) = \{z(1,1), z(2,3), z(3,5)\} && \pi(6) = \{z(1,1), z(2,3), z(3,4)\}\\
&\pi(7) = \{z(1,2), z(2,2), z(3,5)\} && \pi(8) = \{z(1,2), z(2,2), z(3,4)\}\\
&\pi(9) = \{z(1,2), z(2,2), z(3,3)\}
\end{aligned} \tag{4.4}$$

Fully embellishing this model involves eliciting the core states, the underlying drivers of the number of meals passing through each of the actors. These can be readily adapted to reflect the beliefs of different domain experts. For instance, different school districts often follow different summer school schedules, so if the advocates were interested in applying the model to a different region, it would simply require updating the core state parameters. The information about the path flows is most readily supplied through available data about the number of meals prepared, transported, and served throughout the summer.

As with the MDM, we can read the conditional independence relationships in the model that result from the definition. The dynamic linear model is essentially a Markov chain, so we should check that the flow of items in the network only depends on the previous iteration. If not, then the model must be adapted to express a Markov chain with memory. Furthermore, we must check to see that the past observations of how much stuff is in the model at each level are independent of future amounts given all of the governing state parameters for that particular time step. We can check this by plotting the one-step ahead forecast as we did for the MDM.

4.4 Discussion

The case studies in Sect. 4.3 show how drawing the structure from the experts' natural language description motivates the development of more flexible models that can highlight key features of a domain problem. The SBP example shows that a BN is appropriate when the expert describes a problem as a set of elements that depend on each other. The SNAP application example highlights the advantages of a tree-based approach when the experts describe a series of events and outcomes. The open SMP example shows how additional restrictions on the BN structure can draw out the contemporaneous strengths between elements of the model that is crucial to the experts' description. Lastly, the flow of meals in a system shows how working with the accessible representation of meal flow in a system can be translated into a valid structure while remaining faithful to the assumptions expressed by the expert.

Table 4.1 Examples of customized graphical models

Name	Description	When to use	Applications
(Dynamic) Bayesian network	Directed acyclic graph of random variables	Systems naturally expressed as dependence structure between random variables	Biological networks Smith (2010), ecological conservation Korb and Nicholson (2009)
(Dynamic) chain event graph	Derived from event tree coloured to represent conditional independence	Asymmetric problems, problem description is told as a series of unfolding events	Healthcare outcomes Barclay et al. (2014), forensic evidence Collazo et al. (2017)
Chain graphs	Hybrid graph with directed and undirected edges	Problem description has both directional and ambiguous relationships	Mental health Cox and Wermuth (1993), social processes David (2014)
Flow graph	Hierarchical flow network	Supply and demand problems, homogeneous flows	Commodity supply Smith and Figueroa (2007)
Multi-regression dynamic model	Collection of regressions where the parents are the regressors	Contemporaneous effects between time series	Marketing Smith (1993), traffic flows Queen et al. (2009), neural fMRI activity Costa et al. (2015)
Regulatory graph	Graph customized to regulatory hypotheses	Need to test a regulatory hypothesis	Biological control mechanisms Liverani and Smith (2015)

A summary table is shown in Table 4.1 citing additional examples of applications of these bespoke graphical models we have used in the past. References are given for two classes of models, chain graphs, and regulatory graphs that are not explored in this chapter. This is of course a small subset of all the formal graphical frameworks now available. These case studies and applications in the table are examples of possible customized models.

Generally, allowing these representations to capture dynamics uniquely to a given application cultivates more suitable representations. Just as the d-separation theorem allows us to reason about conditional independence in the BN, analogous theorems elucidate the dependence structure of custom representations. Each of these examples of elicited structure has its own logic which can be verified by examining the conditional independence statements and confirming with the expert that the model accurately conveys the expert's beliefs.

Carefully drawing structure from an expert's natural language description is not an exact science. We have offered a few guidelines for when to use particular models summarized in the flowchart in Fig. 4.11. The examples discussed here are far from exhaustive, and Fig. 4.11 also highlights areas of open research. Spirtes and Zhang

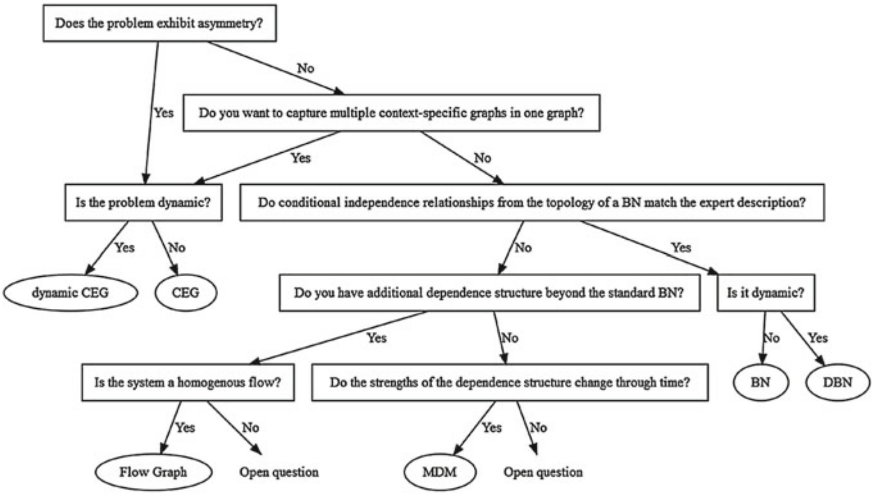

Does the problem exhibit asymmetry?
Yes
No
Do you want to capture multiple context-specific graphs in one graph?
Yes
No
Is the problem dynamic?
Do conditional independence relationships from the topology of a BN match the expert description?
Yes
No
dynamic CEG
CEG
No
Yes
Do you have additional dependence structure beyond the standard BN?
Is it dynamic?
Yes
No
No
Yes
Is the system a homogenous flow?
Do the strengths of the dependence structure change through time?
BN
DBN
Yes
No
Yes
No
Flow Graph
Open question
MDM
Open question

Fig. 4.11 Flowchart to guide picking an appropriate structure

(2016) confirm that determining what new classes of models might be more appropriate than a BN for a given domain. A full protocol for choosing one customizing model over another remains to be formalized. While software for BN elicitation is ubiquitous, robust software for these alternative models is under development.

The premise of drawing the structure from a natural language description rather than tweaking a model to fit an existing structure represents a substantial shift in how modellers elicit structure. Furthermore, inference on each of these novel representations engenders customized notions of causation, as each of the full probability representations of customized models admits its own causal algebras. The causal effects following intervention in a BN are well studied, and these methods can be extended to custom classes of models discussed here. A thorough investigation of causal algebras is beyond the scope of this chapter, but it offers further motivation for careful attention to structure in the elicitation process. In a later work, we will demonstrate how each structural class has its own causal algebra and that for causation to be meaningful the underlying structure on which it is based needs to properly reflect domain knowledge. The work of customized structure elicitation is a relatively poorly explored space. We hope this chapter excites others to develop new tools to make problem descriptions more powerful and reliable.

Acknowledgements The authors would like to thank The Alan Turing Institute EPSRC Grant EP/L016710/1 and the Leverhulme Bridges for their support.

References

Andersson, S. A. (2001). An alternative markov property for chain graphs. *Scandinavian Journal of Statistics, 28*(1), 33–85.

Barclay, L. M., Hutton, J. L., & Smith, J. Q. (2014). Chain event graphs for informed missingness. *Bayesian Analysis*, *9*(1), 53–76.

Barclay, L. M., Collazo, R. A., Smith, J. Q., Thwaites, P. A., & Nicholson, A. E. (2015). The dynamic chain event graph. *Electronic Journal of Statistics*, *9*(2), 2130–2169.

Barons, M. J., Wright, S. K., & Smith, J. Q. (2018). Eliciting probabilistic judgements for integrating decision support systems. In *Elicitation* (pp. 445–478). Springer.

Bedford, T., & Cooke, R. (2001). *Probabilistic risk analysis: Foundations and methods*. University of Cambridge Press.

Bedford, T., & Cooke, R. M. (2002). Vines—A new graphical model for dependent random variables. *Annals of Statistics*, *30*(4), 1031–1068.

Bouckaert, R. R., & Studeny, M. (1995). Chain graphs: Semantics and expressiveness. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty* (pp. 69–76). Springer.

Chilton, M., & Rose, D. (2009). A rights-based approach to food insecurity in the United States. *American Journal of Public Health*, *99*(7), 1203–11.

Collazo, A. R., Görgen, C., & Smith, J. Q. (2017). *Chain Event Graphs*.

Cooke, R. M. (1991). *Experts in uncertainty: Opinion and subjective probability in science*. New York: Oxford University Press.

Costa, L., Smith, J., Nichols, T., Cussens, J., Duff, E. P., & Makin, T. R. (2015). Searching multiregression dynamic models of resting-state fMRI networks using Integer programming. *Bayesian Analysis*, *10*(2), 441–478.

Cox, D. R. & Wermuth, N. (2014). *Multivariate dependencies: Models, analysis and interpretation*.

Cox, D. R., & Wermuth, N. (1993). Linear Dependencies represented by Chain Graphs. *Statistical Science*, *8*(3), 204–283.

Dowler, E. A., & O'Connor, D. (2012). Rights-based approaches to addressing food poverty and food insecurity in Ireland and UK. *Social Science & Medicine*, *74*(1), 44–51.

Gundersen, C., Kreider, B., & Pepper, J. (2011). The economics of food insecurity in the United States. *Applied Economic Perspectives and Policy*, *33*(3), 281–303.

Hagan, A. O.', & Oakley, J. (2014). *SHELF: the Sheffield elicitation framework*.

Hanea, A. M., Kurowicka, D., & Cooke, R. M. (2006). Hybrid method for quantifying and analyzing Bayesian Belief Nets. *Quality and Reliability Engineering International*, *22*, 709–729.

Hanea, A. M., McBride, M. F., Burgman, M. A., & Wintle, B. C. (2018). Classical meets modern in the IDEA protocol for structured expert judgement. *Journal of Risk Research*, *21*(4), 417–433.

Kaye, L., Lee, E., & Chen, Y. Y. (2013). Barriers to Food stamps in New York State: A perspective from the field. *Journal of Poverty, 17*(1), 13–28.

Korb, K. B., & Nicholson, A. E. (2009). *Bayesian Artificial Intelligence*.

Liverani, S., & Smith, J. Q. (2015). Bayesian selection of graphical regulatory models. *International Journal of Approximate Reasoning*, *77*, 87–104.

Loveless, T. A. (2010). Food stamp/Supplemental Nutrition Assistance Program (SNAP) Receipt in the Past 12 Months for Households by State: 2010 and 2011 American Community Survey Briefs.

Nolen, E., & Krey, K. (2015). The effect of universal-free school breakfast on milk consumption and nutrient intake. *Food Studies: An Interdisciplinary Journal*, *5*(4), 23–33.

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Elsevier.

Pearl, J., & Verma, T. S. (1995). A theory of inferred causation. *Studies in Logic and the Foundations of Mathematics, 134*, 789—-811.

Pearl, J. (2009). *Causality* (2nd ed.). New York, USA: Cambridge University Press.

Queen, C. M., & Albers, C. J. (2009). Intervention and causality: Forecasting traffic flows using a dynamic Bayesian network. *Journal of the American Statistical Association, 104*(486), 669–681.

Richardson, T., & Spirtes, P. (2002). Ancestral graph Markov models. *The Annals of Statistics*, *30*(4), 962–1030.

Rougier, J., & Crucifix, M. (2018). Uncertainty in climate science and climate policy. *Climate Modelling*, 361—-380.

Shafer, G. (1996). *The Art of causal conjecture*. The MIT Press.

Smith, J. Q. (1993). Multiregression dynamic models. *Journal of the Royal Statistical Society: Series B*, *55*(4), 849–870.

Smith, J. Q. (2010). *Bayesian decision analysis: Principles and practice*. Cambridge: Cambridge University Press.

Smith, J. Q., & Anderson, P. E. (2008). Conditional independence and chain event graphs. *Artificial Intelligence*, *172*(1), 42–68.

Smith, J. Q., & Figueroa, L. J. (2007). A causal algebra for dynamic flow networks. *Advances in Probabilistic Graphical Models*, *213*, 39–54.

Spirtes, P., & Zhang, K. (2016). Causal discovery and inference: Concepts and recent methodological advances. *Applied Informatics*, *3*(1), 3.

Steffen, S. L. (1996). *Graphical models*. Oxford: Clarendon Press.

Thwaites, P. A., & Smith, J. Q. (2015). A new method for tackling asymmetric decision problems. (Id).

West, M., & Harrison, J. (1997). *Bayesian forecasting and dynamic models*, New York.

Wilkerson, R. L., & Krey, K. (2015). Associations between neighborhoods and summer meals sites : Measuring access to federal summer meals programs. *Journal of Applied Research on Children: Informing Policy for Children at Risk, 6*(2).

Chapter 5
Bayesian Modelling of Dependence Between Experts: Some Comparisons with Cooke's Classical Model

David Hartley and Simon French

Abstract A Bayesian model for analysing and aggregating structured expert judgement (SEJ) data of the form used by Cooke's classical model has been developed. The model has been built to create predictions over a common dataset, thereby allowing direct comparison between approaches. It deals with correlations between experts through clustering and also seeks to recalibrate judgements using the seed variables, in order to form an unbiased aggregated distribution over the target variables. Using the Delft database of SEJ studies, compiled by Roger Cooke, performance comparisons with the classical model demonstrate that this Bayesian approach provides similar median estimates but broader uncertainty bounds on the variables of interest. Cross-validation shows that these dynamics lead to the Bayesian model exhibiting higher statistical accuracy but lower information scores than the classical model. Comparisons of the combination scoring rule add further evidence to the robustness of the classical approach yet demonstrate outperformance of the Bayesian model in select cases.

5.1 Introduction

Algorithmic approaches for combining judgements from several experts have evolved over the years. Initially, techniques were either simple averaging, known as opinion polling, or in essence Bayesian (French 1985, 2011). However, the Bayesian approach did not prove practical and fell by the wayside, whilst the opinion polling techniques gained traction. In practice, Cooke's development of a performance-weighted opinion polling approach, known as the classical model (Cooke 1991, 2007), dominated among the mathematical approaches to eliciting and aggregating expert judgement data and remains the exemplar in this field. Non-mathematical

D. Hartley (✉) · S. French
Department of Statistics, University of Warwick, Coventry CV4 7AL, UK
e-mail: d.s.hartley@warwick.ac.uk

S. French
e-mail: simon.french.50@gmail.com

A. M. Hanea et al. (eds.), *Expert Judgement in Risk and Decision Analysis*,
International Series in Operations Research & Management Science 293,
https://doi.org/10.1007/978-3-030-46474-5_5

approaches (designated “behavioural” approaches) to combining experts’ assessments have also been applied in many contexts. Here, typically, a group of experts discuss and agree on some form of consensus probability distribution within a structured framework (Garthwaite et al. 2005). There are benefits and risks to either behavioural or mathematical aggregation techniques, both practically and philosophically; however, both are possible and the choice in practice on which to use is context-dependent (EFSA 2014).

Bayesian approaches (Hartley and French 2021) for SEJ start with the formalisation of a prior probability representing the decision-maker’s belief ahead of hearing from the experts. Experts’ judgements are then treated as *data*, and appropriate likelihood functions are created to represent the information inferred from their stated judgements. Bayes’ theorem is applied to combine the prior with the elicited judgements on the uncertainty, to give the decision-maker’s posterior perspective given the experts’ statements. Calculation of the, potentially very complex, likelihood function was one of the key challenges that made early Bayesian models intractable.

Bayesian methods are starting to become more tractable with the advent of more effective computational approaches, particularly Markov Chain Monte Carlo (MCMC) (Wiper and French 1995; Clemen & Lichtendahl 2002; Lichtendahl 2005; Albert et al. 2012; Billari et al. 2014). At the same time, many of the principles early Bayesian models sought to highlight, e.g. expert to expert correlation, have not been explicitly tackled within existing non-Bayesian models. Thus the time is right to more formally assess these new Bayesian frameworks versus current approaches in an aim to build their credibility with decision-makers.

One of the key characteristics of Bayesian models is that they can utilise a parametric structure and thus infer a final posterior parametric distribution to represent the decision-maker’s belief given the experts’ judgements. This is a motivating factor for considering Bayesian frameworks for mathematical aggregation in SEJ. Opinion pooling techniques result in non-parametric representations of the consensus output. SEJ outputs are often used as inputs to broader parametric models and thus having the consensus in a parametric form can be very powerful. Another motivating factor for considering Bayesian models, in addition to the ability to encode more complex dynamics such as expert to expert correlation, is the ability to specifically incorporate prior knowledge into the process. If we are deploying SEJ in contexts where there is a well-defined decision-maker, the Bayesian approach can help understand how her unique position changes given the experts’ inputs. Note, in some cases, SEJ will be used to act as a “rational scientist”. In these cases, informative priors may be inappropriate and the model can be adjusted to take “naive” priors, whereby nearly all the information encapsulated in the posterior comes from the experts’ judgements as the priors have been selected to be intentionally uninformative.

Bayesian models are often structurally context-dependent, and many models have been utilised in only a small number of settings, eliminating the possibility of a broad meta-analysis. Building on the work of (Lichtendahl 2005; Albert et al. 2012) and (Billari et al. 2014), within this chapter, we have considered a Bayesian framework applied retrospectively to existing SEJ studies. This allows us to generate predictions against a common data set and compare existing models accordingly. We recognise this does necessitate some compromises within the Bayesian paradigm which may limit efficacy (such as no input during expert elicitation on parametrisation or prior selection), however, does set a benchmark for the use of generalised Bayesian models within SEJ.

It is important to note that any Bayesian approach may suggest a different procedure to the elicitation and documentation of SEJ studies (Hartley and French 2021; EFSA 2014; Cooke et al. 2000). Our aim in this chapter is to demonstrate some practical applications of the Bayesian framework utilising the database[1] of studies compiled by Cooke and Goossens (2008) and provide some performance comparisons with the classical model. Performance assessments are used to build the case for the feasibility of generalised Bayesian frameworks and to provide evidence that such a framework could be a credible choice for a decision-maker. Whilst not a primary focus within this chapter, we shall to a lesser extent note some of the more procedural elements that are important when considering Bayesian approaches.

5.2 Overview of the Bayesian Model

Bayesian approaches treat experts judgements as *data* and then create appropriate likelihood functions to represent the information implicit in their statements. The main complexity in applying Bayesian methods relates to:

- the experts' ability to encode their knowledge probabilistically and their potential for overconfidence (Clemen & Lichtendahl 2002; O'Hagan et al. 2006; Hora 2007; Lin and Bier 2008);
- shared knowledge and common professional backgrounds which drives correlation between expert' judgements (Shanteau 1995; Mumpower and Stewart 1996; Wilson 2016; Hartley and French 2021);
- correlation that may exist between the experts judgements and the decision-makers own judgements (French 1980);
- the effects of other pressures which may drive bias. These may arise from conflicts of interests, fear of being an outlier, concern about future accountabilities, competition among the experts themselves, more general psychological biases, and emotional and cultural responses to context (Hockey et al. 2000; Skjong and Wentworth 2001; Lichtendahl and Winkler 2007; French et al. 2009; Kahneman 2011).

The Bayesian perspective makes it clear that one needs to think about correlation between experts' judgements due to shared knowledge; other approaches to aggregating expert judgements do not. As any statistician knows, ignoring dependencies

[1] This database is constantly growing as studies are completed. To give an indication of scale, when the Eggstaff OOS validation analysis (Eggstaff et al. 2014) was conducted, 62 datasets were evaluated. These sets included 593 experts and 754 seed variables which resulted in 6,633,508 combinations and 67,452,126 probability judgements. A subset of these are considered within this chapter for cross-validation of the Bayesian approach.

between data leads to overconfidence in estimates. The same is true here, although we have noted that allowing for correlations between experts has been a considerable hurdle to the development of practical Bayesian methods.

The Bayesian framework we have employed simplifies some of the inherent complexity by breaking the post-processing into four distinct steps:

- Expert clustering
- Distribution fitting
- Recalibration
- Aggregation

The method is applied to judgements in the form used by the classical model. Here, estimates are elicited for both the *target* variables of interest and for *seed* variables, for which the analyst conducting the study knows the values a priori but the experts do not. These variables can be used as a calibration dataset within the Bayesian paradigm but are used by the classical model in order to calculate the performance weighting scores. All elicitations are made against a standard set of quantiles (typically three–0.05, 0.5, 0.95 or five–0.05, 0.25, 0.5, 0.75, 0.95).

We will not give a full mathematical exposition of our Bayesian framework here; however, we shall outline some of the key components behind each of the above steps to help with the analysis later on. For a full mathematical background, please consult Hartley and French (2021).

5.2.1 Expert Clustering

One of the risks leading to overconfidence in a final posterior comes from the shared knowledge or common professional backgrounds that experts may have which drives correlation. As we outlined before, finding such an underlying correlation and correcting for it is often a challenge for Bayesian models. One approach to bypass the issue of directly calculating complex correlation matrices would be to identify the sources of the underlying similarity in estimation and with this knowledge cluster experts into homogeneity groups in which all experts with similar historic knowledge are grouped together. As part of the aggregation exercise, this knowledge could be utilised to reduce the risk of overconfidence (Albert et al. 2012; Billari et al. 2014). One approach to forming these groups would be to attempt to elicit information about potential sources of common knowledge, in addition to the quantiles, from the experts. This approach is appealingly simple and would require only a procedural update. In practice, however, this elicitation is likely to be challenging as sources of this correlation may be opaque, even to the experts themselves. Thus, algorithmic approaches, which attempt to infer these groupings, could be considered.

The framework we have employed, similar to (Billari et al. 2014), utilises algorithmic clustering techniques in order to group and re-weight experts. Given the classical model data structure, there is a choice of data set to use for the clustering exercise, the target variables, the seed variables, or a combination thereof. We have chosen here to use the seed variables. If there is underlying correlation between experts, driven by their shared knowledge, then this correlation should be apparent in their seed variable estimations. If there is no such link on the seed variables, then we would argue that there is limited risk of overconfidence on the target variables. This is clearly true only if the seed variables are within the same domain as that of the target variables, i.e. shared knowledge of experts in rare genetic conditions within hamsters does not imply shared knowledge in the risk of a bolt breaking within a suspension bridge. Representativeness of seed variables is similarly a core tenet underlying the use of these variables within the classical model. Please note that definition of meaningful seed variables is not easy, and there are those that would question the use of these variables altogether, although extensive cross-validation literature on Cooke's model does demonstrate their value (Colson and Cooke 2017; Eggstaff et al. 2014; Lin and Cheng 2009; Flandoli et al. 2011). We will leave this aside for now, however, and note that similar to Billari et al. (2014), the target variables could have been used in their place. Given seed variable estimations, it is easy to apply any number of clustering algorithms to the seed variable space (in which each expert is a point) in order to generate the expert groupings. We recommend utilising either hierarchical clustering, due to its efficacy over sparse datasets (and easy comprehension by decision-makers) or mixture models, specifically, Dirichlet process mixture models, due to their limited assumptions about the number of groupings a priori and their ability to integrate easily with the broader Bayesian framework (Billari et al. 2014).

5.2.2 Distribution Fitting

Within Bayesian frameworks, it is common for distribution fitting to be utilised in order to apply parametric models in the post-analysis of experts' assessments. This both makes the computation simpler and aligns with the assumption, in many practical applications, that underlying phenomena are parametric in nature. One of the benefits of a Bayesian approach is this parametric form. Often the outputs of an SEJ study can feed further analysis and having a fully parametrised posterior distribution can make calculations of future models much simpler. Opinion pooling method outputs are typically non-parametric.

Due to the complexity of eliciting experts' expectations on parameters, it is often preferable to elicit on observables first and then parametrise *post hoc*. Ideally, this would be done in conjunction with the experts (similar to behavioural SEJ approaches) to ensure that they are comfortable with the final statement about their beliefs; however, as we are applying this analysis retrospectively, this is not feasible. Thus, a choice must be made of which parametrisation to use. The aim of any fitting process must be to select a distribution which minimises the discrepancy to quantiles elicited from the experts in order to ensure that the fitted distribution reasonably reflects their underlying beliefs. Commonly used distributions within SEJ are the Gaussian distribution (Albert et al. 2012; Billari et al. 2014), the log-normal distribution (de Vries and van de Wal 2015) or a piecewise distribution which is uniform on the interior quantiles and exponential on the tails (Clemen & Lichtendahl 2002). All of these distributions have advantages and disadvantages, e.g. in fitting the log-normal distribution, assumptions must be made on the un-elicited minimum and on the exponential behaviour post the top quantile, or, in the Gaussian, assumptions of symmetry about the mean. To this extent, we have chosen to utilise a two-piece Gaussian distribution (a Gaussian distribution with different variances above and below the median). This choice allows exact fitting to the expert quantiles, with minimal points of discontinuity and no assumptions on the extremities. It is, however, admittedly an *ad hoc* choice and our framework is generic, and thus could be applied to many parametrisations. The impact of different parametrisations on the final decision-maker posterior is an area for further research. If this Bayesian framework were to be applied to a study from the offset (rather than utilising data *post hoc* as we are doing here), then discussions about the appropriate distributions to use should be had with the experts.

5.2.3 Recalibration

Bayesian models typically consider the topic of recalibration differently to frequentist approaches. In the Bayesian model, as probability is subjective and thus a property of the observer (typically the decision-maker) of the system, it appears reasonable, for any such observer to consider all the information at hand in forming their final posterior distribution. An example of such information may be any bias which the experts have exhibited in historic judgements. Many potential drivers of bias, such as anchoring (Kahneman et al. 1982; Kahneman 2011), can be minimised through elicitation procedure (Cooke et al. 2000). Others, such as consistent over/under confidence, are often still visible (Burgman 2015). Thus if expert A, from a pool of experts, has historically been systematically overconfident, a Bayesian decision-maker may choose to broaden the tails in expert A's elicited judgement distributions, before aggregating with other experts, in order to truly reflect the decision-maker's belief of the uncertainty. Please note that there is significant resistance to this form of recalibration in certain areas with the argument that you should not adjust the experts forecasts as this creates an ownership problem (effectively the forecasts are no longer the experts' once you have adapted them, they belong to the analyst) and an accountability issue accordingly.[2] We would argue that the use of recalibration is context-dependent. In expert judgement problems with a single decision-maker, it would potentially be remiss to ignore any such information about potential additional uncertainty. Regardless, the model we have used is modular in design and recalibration could be included or excluded as appropriate given the context of the

[2] If it is assumed that experts are operating as coherent subjective Bayesians (Finetti 1974; De Finetti 1957), then there are mathematical inconsistencies with certain forms of recalibration (Kadane and Fischhoff 2013; Lichtenstein et al. 1982). However, there is evidence of incoherence among expert judgements within the Delft data, even on the small number of elicited quantiles. The exact form of calibration we are employing is also explicitly excluded from the mathematical analysis in Kadane and Fischhoff (2013).

problem at hand. Significant overconfidence is apparent in many studies within the Delft database; thus, this analysis has included recalibration. Further work should be conducted to empirically assess the impact of this recalibration. One approach to this would be to conduct the same analysis outlined in this chapter using both calibrated and un-calibrated Bayesian approaches. This is left for further research.

Seed variables, used for the performance weighting calculation within the classical model framework, can be used for the quantification of bias adjustments. These variables, elicited from the experts with true realisations known by the facilitator a priori, allow an analyst to identify if there are any systemic biases between prediction and realisation which need to be eliminated. The approach we use for this, taken originally from Clemen & Lichtendahl (2002), is to identify "inflation" factors which are multiplicative parameters inferred from the seed variable estimates and their realisations. In the case of a study in which three quantiles are elicited, there are three multiplicative parameters. The first of these is a positioning inflation parameter that assesses if there is consistent over or under forecasting of the *median* assessment. The other two parameters are then multiplicative dispersion parameters. These are calculated on the distance (or in the case of the two-piece Gaussian, the standard deviation), defined by the gap between the median and the upper/lower estimates, respectively. In this way, the dispersion inflation parameters control for any systemic bias in over- or underestimating the *uncertainty* in the judgements the experts give. Posterior estimates for these multiplicative[3] inflation factors can be inferred for each expert by starting with the assumption that they are well calibrated and then utilising the seed variables provided in the study as data and passing each through Bayes' rule. In practice, we also strengthen the analysis by allowing an inflation factor dependency structure between experts. We infer this through hierarchical models and MCMC as per Clemen & Lichtendahl (2002).

5.2.4 Aggregation

Once the set of expert homogeneity groups (H) and a final set of individual experts' judgements on the target variables (which have been recalibrated and fit to appropriate parametric distributions) have been confirmed, we can combine these to create a final posterior through aggregation. In order to do this, we utilise a hierarchical model, first proposed in Albert et al. (2012), which includes a novel approach to capturing the dependencies between experts. The aggregation model assumes that each expert's parameterised beliefs, derived from the elicited quantiles, are linked to that of the other experts in their group through a common shared group distribution. Each group will have a parametrised distribution, with parameters defined by the

[3] Utilising multiplicative inflation factors in this way does put constraints on the scales of variables (both seed and target) as it assumes that all variables are of a similar order of magnitude. If we imagine some variables are logged, then this form of recalibration would not work. This is currently a restriction with this framework and more research is required into potential solutions, although one possible approach is outlined in Wiper and French (1995).

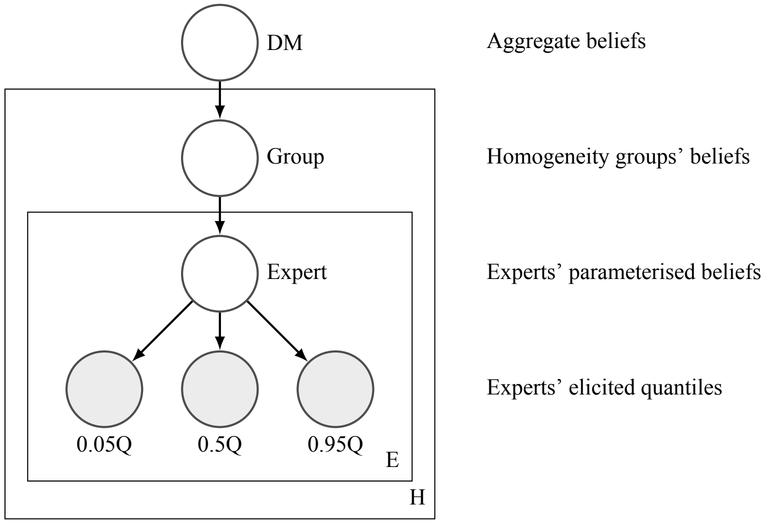

DM
Group
Expert
0.05Q
0.5Q
0.95Q
E
H
Aggregate beliefs
Homogeneity groups' beliefs
Experts' parameterised beliefs
Experts' elicited quantiles

Fig. 5.1 A belief hierarchy for aggregation of expert judgement with homogeneity groups (DM - decision-maker)

combined beliefs of its expert members. The groups likewise are linked to each other via a common shared universal distribution that of the Supra-Bayesian. The final combined posterior distribution represents the updated decision-maker judgement and is calculated through MCMC. A simple diagram of this model is shown in Fig. 5.1.

The motivation for this expert partition is that rather than explicitly calculating the correlation matrix, the grouping approach is used to appropriately weigh the impact of each expert in the final model, offsetting overconfidence effects driven by correlation.

One of the advantages of this approach is that the hierarchical model can capture both the underlying consensus and diversity between experts. Opinion pooling methods do not attempt to assess consensus of opinion. Additionally, hierarchical Bayesian models of this nature allow inference not only the posterior distributions of the target variables but also all of the other latent elements within the model, such as inter-group dependencies. To this extent, it is possible to recover after the analysis has been completed, all of the homogeneity groups' beliefs as well as the parametrisations used for the individual experts. This gives the analyst a diagnostic tool to help understand how uncertainty has propagated through the model.

In order to combine each of the above four steps within our framework, we utilise MCMC. We have chosen to do our grouping utilising agglomerative hierarchical clustering (Charrad et al. 2014) within this chapter. This is to ensure deterministic group definitions given the significant number of predictions made. This means we have a two-step process, one step to create the necessary clusters and the second to do the parametrisation, recalibration and aggregation, which are all done within a single piece of MCMC code. If we utilise a Bayesian hierarchical clustering algorithm (such as Dirichlet process mixture models), then this allows us to do all four steps within each iteration of the MCMC algorithm. This is philosophically appealing as it means we only use each piece of data once (seed variable information is used twice in the two-step model) but proves less stable with very small datasets and less intuitive to the decision-maker. To this extent, the choice of whether to use a one-step or two-step process is context-dependent. For a full mathematical exposition of the Bayesian framework, and to understand how the four elements are combined together, a review of Hartley and French (2021) is recommended.

Given this framework, it is interesting to understand how results compare to those of Cooke's classical model. Thanks to the open availability of historic Expert Judgement Studies, and the underlying data that Roger Cooke has kindly provided, we see how results differ from classical model outputs. We will start by doing a deeper dive into a couple of specific examples within geology and environmental resource management before looking more generically across the breadth of studies within the Delft database.

5.3 Effusive Eruption

Following the eruption of the Icelandic volcano, Eyjafjallajökull, in 2010 a scientific emergency group (SAGE) was appointed by the UK government. One of the tasks of this group was to consider the potential of future eruption scenarios that may impact the UK, and volcanic eruptions were subsequently added to the UK National Risk Register. One of the key scenarios adopted by the UK National Risk Register was considering the eruption of the Grimsvötn volcano (commonly known as the Laki Eruption due to its presence within the Laki crater) which occurred in 1783–84. This volcano had a huge impact on Europe, particularly in Iceland where 60% of the grazing livestock died (predominantly by Fluorosis) and 20% of the Icelandic population were also killed as a result of illness, famine and environmental stress. This eruption was considered to represent a "reasonable worst-case scenario" for future eruptions.

Risk to the UK from such a scenario recurring would be in the form of volcanic gases, aerosols, acid rain and deposition of acids. These factors can have significant environmental impact (due to deposition on vegetation, buildings and potential impact to groundwater), or impact on transport, particularly aviation (as we saw with Eyjafjallajkull), where sulphur dioxide and sulphuric acid can cause damage to airframes and turbines, engine corrosion, or put crew and passengers at risk of exposure. To model this complexity, meteorological (weather and atmospheric transport) models, in addition to chemistry models, are considered. In order to support this modelling and determine a set of prior values for some of the source characteristics, an expert judgement study was conducted in 2012 (Loughlin et al. 2012) (note: this study followed an earlier study conducted on the same topic in 2010 (EFSA 2010)).

Structurally, the elicitation was conducted with 14 multidisciplinary experts. Experts were from academia, research institutes and other institutes with operational responsibilities. These experts were able to cover all of the modelling fields described earlier (meteorology, atmospheric dispersion, chemistry) in addition to specific volcanology expertise. Quantitative responses were captured for 8 seed variables, alongside 28 target variables (22 volcanological in nature and 6 related to plume chemical processes). Not all questions were answered by all experts, with number of responses for each variable coming from between 5 and all 14 experts. For comparisons between the Bayesian framework and the classical performance-weighted model, we shall consider only the 10 target variables which were responded to by all experts and for which details are captured within the Delft database (Cooke and Goossens 2008).

Seed variables experts were asked to quantify were related to the historic Laki eruption, e.g.

- What was the area of the Laki Lava flows in km^2?
- What was the estimated production of Laki in CO_2 megatonnes?

With true realisations of $500\,\text{km}^2$ and 349 megatonnes, respectively, an example target variable question was:

- What is the likelihood that in the next Laki-like eruption there is an episode which releases 10 times more SO_2 on the same timescale as the peak eruption episode during Laki?

With other questions similarly linked back to the Laki eruption, this link is important as it helps ensure that the seed variables are truly representative of the target variables and are thus suitable for use within the recalibration exercise inherent within the Bayesian model (and likewise for appropriate performance weighting in the classical model).

Across the total 112 seed variable estimations, if we were to a priori assume that experts were well calibrated/statistically accurate, we should expect to see 11–12 of the seed variable realisations sitting outside the range given by the 0.05 and 0.95 quantiles provided by the experts. Individual experts would expect to have no more than one judgement where the true realisation sits outside of these bounds. In practice, actually 64% (72) of the true realisations fell outside of the 90th percentile bounds given by the experts. For individual experts, between 37.5% and 100% of realisations fell outside of the confidence bounds given (Fig. 5.2). These results are potentially shocking to the uninitiated and may appear to point to a lack of true "expertise" of the experts in the panel, in practice, however, these types of numbers are not uncommon for judgements within SEJ (Burgman 2015) and reflect the complexity of the underlying dynamics within the contexts in which SEJ operates (hence, the need for judgement in the first place). What it does point to, however, is a cautionary note for decision-makers that experts can often be, and in this case are demonstrated to be, systemically overconfident in their judgements. This furthers the case for recalibration, without which, further uncertainty driven by this overconfidence will be ignored.

Running the classical model over this dataset results in 3 experts getting a weighting (Expert 10–53%, Expert 14–31%, Expert 12–16%) and 11 experts being removed

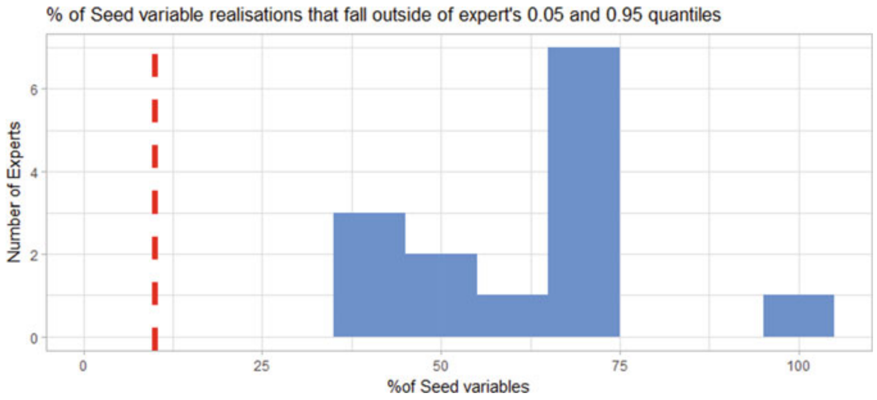


Fig. 5.2 Across seed variables within the Laki effusive eruption scenario study, experts demonstrate significant overconfidence in judgements. All experts have more variables outside of the 0.05 and 0.95 quantiles than would be expected for high statistical accuracy. Red line indicates the expected % of variables for a perfectly calibrated expert

from the final CM optimised decision-maker quantile calculation altogether.[4] To compare the impact of this to the Bayesian framework, we can first consider the homogeneity groups that are created as a result of the first step within the model, the clustering exercise. Running this process identifies five core homogeneity groups within the expert pool (Fig. 5.3), of which two are formed of a single expert and three

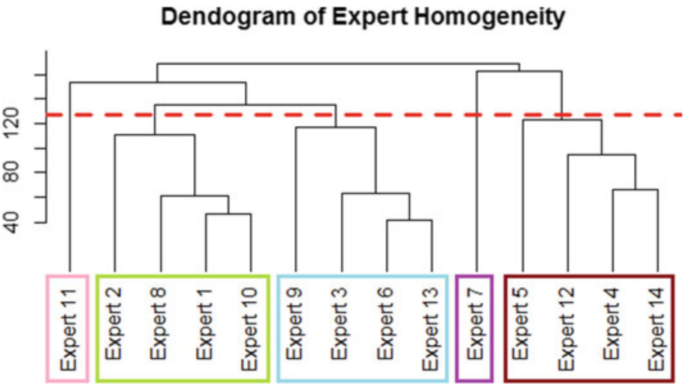


Fig. 5.3 Experts are clustered into five homogeneity groups (coloured boxes) demonstrated by a horizontal cut on the dendogram (red line). These five groupings, based on seed variable responses, cluster the experts into three groups of four individuals and two outliers (Experts 11 and 7) who sit within their own homogeneity groups as they regularly offer differing opinions to the other experts

[4] Please note that, whilst not included in the final calculation of the quantiles within Cooke's methods optimised volcanology, these experts' assessments are still involved in determining the intrinsic range of the random variables.

are each of four experts. The two experts who are grouped within their own pools have done so as a direct result of a significant divergence in judgement between themselves and the remaining groups as it pertains to the seed variables. Thus, the Bayesian model identifies that there is the potential for discrepancy in opinion on the target variables that should be considered and upweights these individuals relative to their peers. In this way, the Bayesian model is capturing the diversity of thinking across the experts. Please note that, at this stage, no judgements have been recalibrated; thus, we do not yet know whether this diversity is a result of different mental models by these experts or due to miscalibration. The recalibration exercise ensures that experts are well calibrated before aggregation, and thus we minimise the risk of simply up-weighting a "poor" forecaster.

Expert judgements are subsequently passed through the distribution fitting, recalibration and aggregation processes described earlier to create a single decision-maker posterior distribution. It is important here to reflect on the context of the analysis that we are conducting, and hence the decision-maker we are trying to model. In the case of this volcanology study, there is not an individual decision-maker whose belief is being updated by the experts. The study is being conducted in order to arrive at a consensus distribution which reflects that of a rational scientist. As a result of this, we need to be thoughtful about the choice of priors that we use in our model. As outlined in Hartley and French (2021), in rational scientist scenarios, there is no individual decision-maker with a significant a priori belief thus we recommend using diffuse priors to minimise the impact of the analyst on the output. The only exception to this is on the median inflation factor, for which we set a tight prior, centred around 1. This assumes that experts are well calibrated on their median and ensures that there are only minor changes feasible to the centrally elicited quantile. However, the model is given freedom to adjust the upper and lower tails to mitigate the overconfidence seen earlier. If extensive changes to the median were allowed in the model, it could be argued that the judgements no longer reflect that of the expert, and thus the aim of achieving a rational scientific consensus would be compromised (note, in the context of an individual updating their beliefs, further recalibration of the median may be appropriate). Numerically, the set of priors considered for our model here are as outlined in Hartley and French (2021), where rational scientist consensus is also the goal. These priors were consistent across all of the analysis within this chapter. More extensive exposition of the considerations of priors within Bayesian SEJ models is outlined in Hartley and French (2021).

Before getting on to discussions regarding the uncertainty bounds provided by the Bayesian/Cooke's models, it is first interesting to assess differences between the posterior median for the Bayesian model and Cooke's optimised decision-maker's 0.5 quantile. In many contexts, final decision-makers will look to a point estimate from which to base their next best action. As the Bayesian model is trying to consider both the consensus and diversity in opinion, the hierarchical nature enforces unimodality in the posterior distribution. This posterior mode (which, due to the parametrisation used, will be located at the median) reflects the most likely single value a decision-maker would use to represent a point estimate. Whilst we recognise that ignoring uncertainty in this way is counter to the goals of risk management for which SEJ is

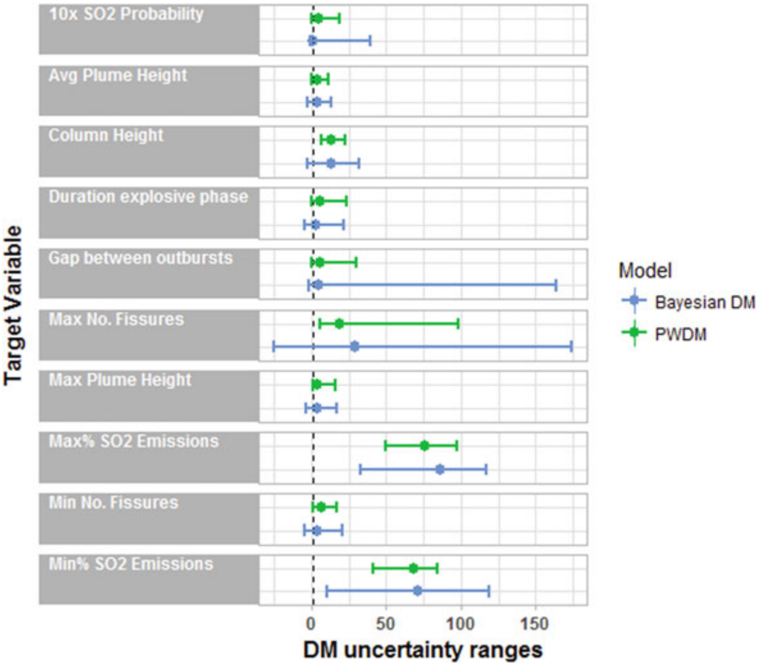


Fig. 5.4 The Bayesian model produces median estimates similar to that of the PWDM and always within the PWDM uncertainty bounds. The Bayesian decision-maker, however, suggests a higher level of underlying uncertainty

typically employed, the use of point estimates is a common decision-making reality and thus worth assessing. For brevity, outputs from Cooke's classical model will henceforth be referred to as the PWDM (Performance-Weighted Decision-Maker with optimisation).

Figure 5.4 outlines the final uncertainty ranges for each of the 10 target variables and each of the ascribed models.[5] It is important to note that across all of the distributions the median for the Bayesian model sits within the uncertainty bounds of the PWDM. This is reassuring. Given the extent to which the PWDM has been utilised in practical studies, if there were fundamental concerns on this number these are likely to have been surfaced before. This suggests that a decision-maker considering either model is not likely to make a significantly different decision based on the expected value alone. There is a noticeable difference, however, in the ranges given by the two models. The PWDM has consistently narrower bounds than the Bayesian decision-

[5] Table 5.3 and Fig. 5.13 in the Supplementary material outline the same responses but also include the results from considering an equal-weighted linear opinion pool (EWDM), omitted initially for brevity and clarity. Comparisons versus the EWDM are considered later when assessing the distributional forms as this provides greater clarity on the difference in modelling approaches than simply the uncertainty bounds alone. All EWDM results have been taken directly from the tool EXCALIBUR used to calculate the PWDM optimised DM.

maker. This is as we would expect and is caused by two predominant factors. Firstly, the PWDM selection criteria, by design, are optimising for statistical accuracy *and* information, and will often trade minor reductions in statistical accuracy for significantly improved information. This occurs due to the fact that information is a slower responding function than statistical accuracy. Secondly, the Bayesian decision-maker is recalibrating the experts' judgements. Given that experts have demonstrated overconfidence, the decision-maker has correspondingly increased uncertainty ranges.

Please note that some of the posterior uncertainty ranges for the target variables within Fig. 5.4 and Table 5.3 modelled using the Bayesian framework are demonstrating infeasible values (e.g. for variable 6, the maximum number of fissures, as a result of the recalibration exercise, negative values have appeared. For obvious reasons, it is not possible to have negative values for such a variable). The Bayesian framework does allow variable constraints to be considered, and the key is the constant of proportionality within Bayes' rule. This constant allows us to apply such bounds *post hoc*, by removing the infeasible area and rescaling accordingly. The other approach would be to consider the framework utilising distributions which are inherently constrained, for example, utilising a beta distribution if the target variable is a percentage as this is naturally constrained to the interval [0, 1]. Neither adjustment has been performed here as the focus is on highlighting the impact of different modelling approaches at a macro-level although these considerations are very important when applying the framework in practice.

Whilst the median and the uncertainty bounds themselves are critical, it is also important to understand the shape of the final decision-maker distribution for each model. Figures 5.5 and 5.6 outline two such distributions, selected as these show different behaviours of the models. The equal-weighted decision-maker (EWDM) distribution has also been added to these slides for comparison. The equal-weighted decision-maker is the result of a linear opinion pool with identical weighting given to each expert.

Target variable 3 (Fig. 5.5) demonstrates common behaviour of the Bayesian model versus Cooke's performance-weighted approach and the equal-weighted decision-maker, notably, a single modal point with a Gaussian decay in either direction (rather than multi-modality), narrower shoulders, and a broader support. One of the outlined aims of the Bayesian framework is to identify underlying consensus in opinion from the experts, and thus this distributional shape is by design and reflects a starting assumption that the Supra-Bayesian decision-maker's belief is of this form. Note that this is a decision to be made in setting up the model, and the framework is generic in nature to support many other possible parametrisations. Broader support is driven by the recalibration portion of the model and the overconfidence displayed by the experts on the seed variables. If experts were systemically under confident, we would expect to see narrower tails on the Bayesian model. In practice, overconfidence is much more common (Burgman 2015).

Target variable 10 (Fig. 5.6) demonstrates a slightly different picture; here, once again the modal point of the Bayesian model is similar to that of the PWDM, and the unimodal shape is maintained. However, in this instance, the EWDM is demonstrating a slightly different picture of the uncertainty. Both the PWDM and the Bayesian model

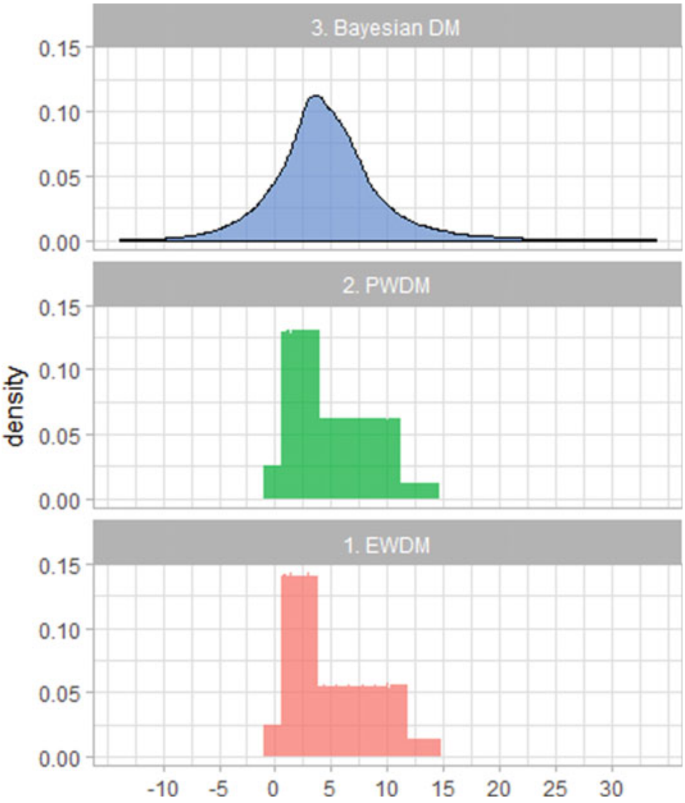


Fig. 5.5 Final distributions for target variable 3: 'After the initial explosive phase (i.e. first days), what is the likely average sustained plume height for gases above the vent for the remainder of the active episode?' Note the similar shapes between the PWDM and EWDM distributions. The Bayesian model demonstrates a slightly higher modal point, and more uniform shape as it is focussed on the underlying consensus in opinion. Note also the larger support of the Bayesian decision-maker as this recalibrates for overconfidence

put a very significant amount of the density at the modal point with little probability to a value below this and a limited but positive probability of more extreme values. The EWDM, however, has significantly less mass around the modal point, a larger probability of a lower realisation and more significant density in the upper tail. It is a positive sign that similar distributional shapes are visible here for the Bayesian and PWDM as there is no a priori reason that this should be the case and suggests that they may both be pointing to similar underlying consensus between experts.

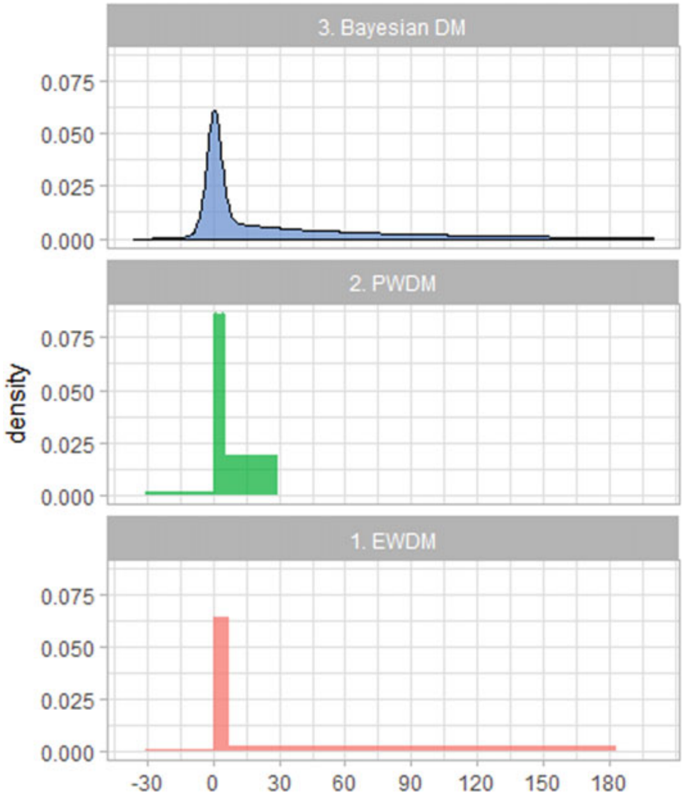


Fig. 5.6 Final distributions for target variable 10: 'What is the typical gap between major gas outburst episodes?' Note, similar to the PWDM (although not the same extent) the Bayesian decision-maker puts more of the distributional density in the region just greater than the modal point. Whilst the Bayesian model includes the greater range suggested by the EWDM, it tapers off much faster

5.4 Invasions of Bighead and Silver Carp in Lake Erie

Forecasting the likelihood of, and damage caused by, invasive non-indigenous species within many natural environments is difficult and poses a problem for those responsible for natural resource management. Similar to the context outlined before, often, the data necessary to build comprehensive decision models is incomplete, and thus expert judgement can be used to supplement what data is available. A recent study, (Wittmann et al. 2015, 2014; Zhang et al. 2016), utilised expert judgement through the classical model to forecast the impacts of Asian carp in Lake Erie. Asian carp is non-indigenous and currently believed not to be established within the lake. Assessments were made to quantify potential aspects of the Asian carp population (biomass, production and consumption) as well as impacts to existing fish species, in the instance that these carp become established within the lake. Establishment could occur as a result of contamination of bait, release by humans or through waterway connections linked to currently established populations.

Structurally, the study comprised of 11 experts, each of whom was asked to assess 84 variables (20 seeds and 64 targets) within the elicitation questionnaire. In practice, for 5 of the seed variables, actual realisations did not become available and for 1 expert, only 11 of the seed questions were responded to. Hence, within this analysis, to ensure consistency across modelling approaches, these 5 seed variables and this 1 expert have been removed, to leave 15 seed variables and 10 experts. Please note that this selection choice differs from the original paper in which the expert was left in the study but a further four seed variables were removed. Elicitations were made against the standard three quantiles (0.05, 0.5, 0.95).

The clustering of experts defined by seed variable responses suggested three core homogeneity groups within the expert pool. The largest group consisted of six members (experts 3, 4, 7, 8, 9 and 10), the second group three members (experts 1, 2 and 5) and finally expert six sat within their own group as their responses consistently differed to those of the remaining groups, suggesting that they may be using a different set of reference data or mental models through which to base their judgements. Supporting a decision-maker in identifying why a particular homogeneity grouping may have arisen could be difficult as the space in which the clustering is performed may be high dimensional (in this case 15 dimensions). Principal component analysis can be used to reduce dimensionality and create a lower dimensional visual representation of the variation in responses between experts. Figure 5.7 outlines some key PCA outputs for this study and identifies the emergent groups in a visual way.

Similar to the prior study outlined, overconfidence was common in the experts across the Lake Erie study. 47 of the 150 (31.3%), seed variable/judgement combinations had realisations sitting outside of the bounds given by the experts. Significantly

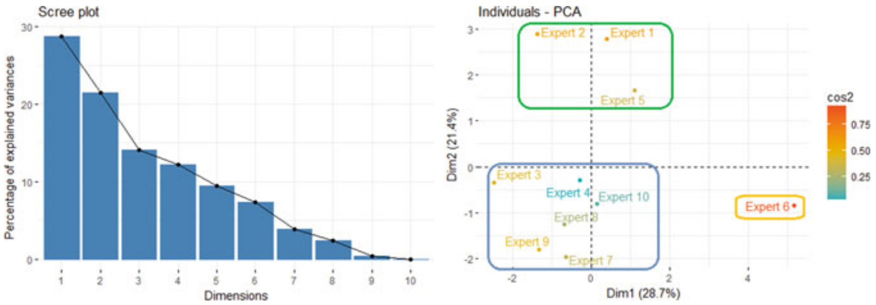


Fig. 5.7 A scree plot of the principal component analysis demonstrates that the first two identified components explain 50.1% of the variance across the original 15 dimensions within the seed variable space. When we isolate these two components and look at where the individual experts sit, the homogeneity groups identified by the model (coloured boxes) emerge. Expert 6 is separated from the remainder and thus is within their own homogeneity group, as they systemically give differentiated responses to the other experts. Note that groupings here are for visualisation purposes only; actual clustering occurs over the full set of seed variable dimensions

Table 5.1 Biomass levels (t/km^2) predicted for the Lake Erie study sole invader scenario

	EWDM			PWDM			BDM		
Target variable	0.05	0.5	0.95	0.05	0.5	0.95	0.05	0.5	0.95
Peak biomass									
Bighead carp	0.0	2.4	17.2	1.6	8.9	25.9	0.7	4.2	13.0
Silver carp	0.0	2.3	17.0	1.6	8.8	25.9	0.7	4.1	11.9
Equilibrium biomass									
Bighead carp	0.0	1.2	9.1	0.4	3.0	12.2	0.3	2.0	6.2
Silver carp	0.0	1.1	8.0	0.4	3.0	12.2	0.3	2.3	6.8

more than the ~15 (10%), we would have expected assuming the experts were all well calibrated. Unlike the effusive eruption example given earlier, however, the range of calibration across the experts was broad. Expert 4 demonstrated strong statistical accuracy, only 1 of the realisations (7%) fell outside of the judgement bounds they gave. As expected this translates into significantly less recalibration within the Bayesian model. Expert 4 had the lowest recalibration parameters within the group. Classical model analysis of the study put all weighting to Expert 4, thereby effectively removing all other experts' judgements from the quantile aggregation within the PWDM optimised decision-maker. Note that, as before, all experts are still included in the calculation of the intrinsic ranges.

The key finding of the original elicitation was that given the right starting condition, there is significant potential for the establishment of Asian carp within Lake Erie. In particular, they have the potential to achieve a biomass level similar to some already established fish species currently harvested commercially or recreationally (yellow perch, walleye, rainbow smelt and gizzard shad). These findings remain when considering the final posteriors proposed by the Bayesian model. The Bayesian model estimations of peak biomass levels for bighead and silver carp, in scenarios where they are the sole invader, suggest higher medians than those predicted by the EWDM but lower than those predicted by the PWDM for both bighead and silver carp (Table 5.1).

Uncertainty ranges within the Bayesian model are slightly narrower than either of the other two. Equilibrium biomass estimates in the same scenarios were lower than peak biomass levels but displayed consistent behaviour between models. The PWDM estimates that the median equilibrium values were approximately 1/3 of the peak value compared to approximately 1/2 in the Bayesian or EWDM models. In the joint invasion scenario (where both bighead and silver carp are established), the Bayesian estimation of the equilibrium biomass was marginally higher than both the PWDM and the EWDM (Table 5.2). Final quantile estimations of the proportion of the total biomass that is bighead carp in the joint invasion scenario were identical in the PWDM and Bayesian models and marginally higher than that of the EWDM (Table 5.2 and Fig. 5.8).

Table 5.2 Estimates for the Lake Erie joint invasion scenario

	EWDM			PWDM			BDM		
Target Variable	0.05	0.5	0.95	0.05	0.5	0.95	0.05	0.5	0.95
Equilibrium biomass	0.0	2.2	12.3	0.4	3.0	12.2	0.6	3.6	10.4
Proportion Bighead carp	0.0	0.3	0.9	0.1	0.5	0.9	0.1	0.5	1

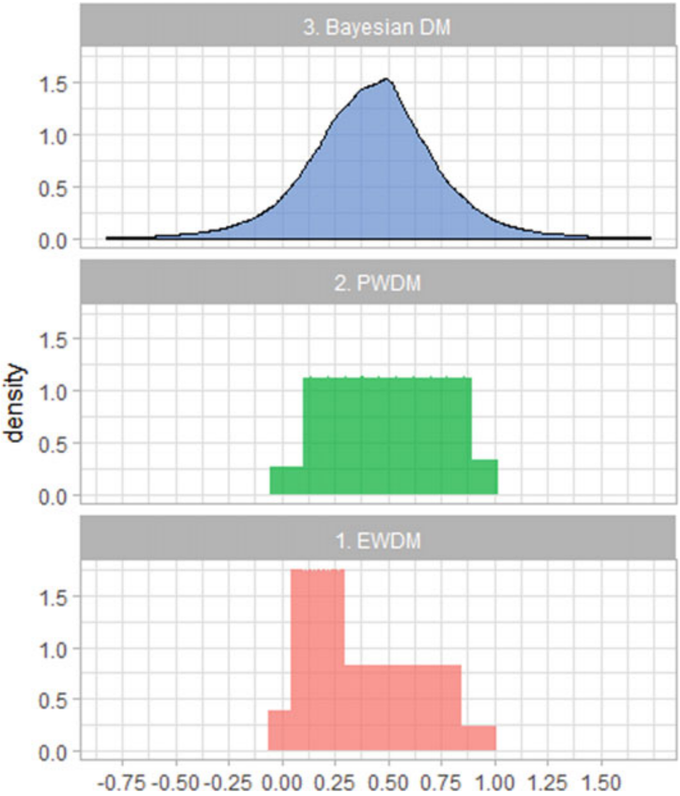


Fig. 5.8 Final distributions for Target variable 10 in the Lake Erie study: ‘What is the proportion of the total biomass that is bighead within the joint invasion scenario?’ The Bayesian model and the PWDM suggest a marginally higher proportion of the biomass will be the non-indigenous bighead carp than EWDM predictions. Note the narrower shoulders and broader tails of the Bayesian model, consistent with other target variable estimations

Overall, similar to what was seen in the effusive eruption case, the quantities of interest resulting from the Bayesian model do not vary significantly (where significance is defined as implying a radically different conclusion from the judgement data) from those of the PWDM, and in this case the EWDM. All models have suggested that there is significant potential for the establishment of tangible biomass of these carp, in relation to existing fish populations, although each model has demonstrated a slightly different posterior distribution of the uncertainty as they emphasise different underlying elements of the judgements.

This is reassuring for a new model, such as the Bayesian framework, as existing models have been used and tested extensively. If radically different values had been found, significant justification would be required.

In both this and the earlier effusive eruption example, we have seen that the median estimate was similar in the performance-weighted and Bayesian approaches. If we look at a broader subset of the Delft studies, specifically a subset where all variables are on a uniform scale to ensure the recalibration algorithm is applicable, we can assess the final decision-maker medians for each of the target variables. In total, there are 20 considered studies with 548 forecasted target variables. In Fig. 5.9, we can see that this similarity between median estimates is true more broadly as there is a strong correlation (0.82) between the final median estimates of the two approaches (0.99, removing outliers).

A log scale is used in Fig. 5.9 to allow us to compare across studies. Whilst variables are on a consistent scale within a single study, they may be on very different

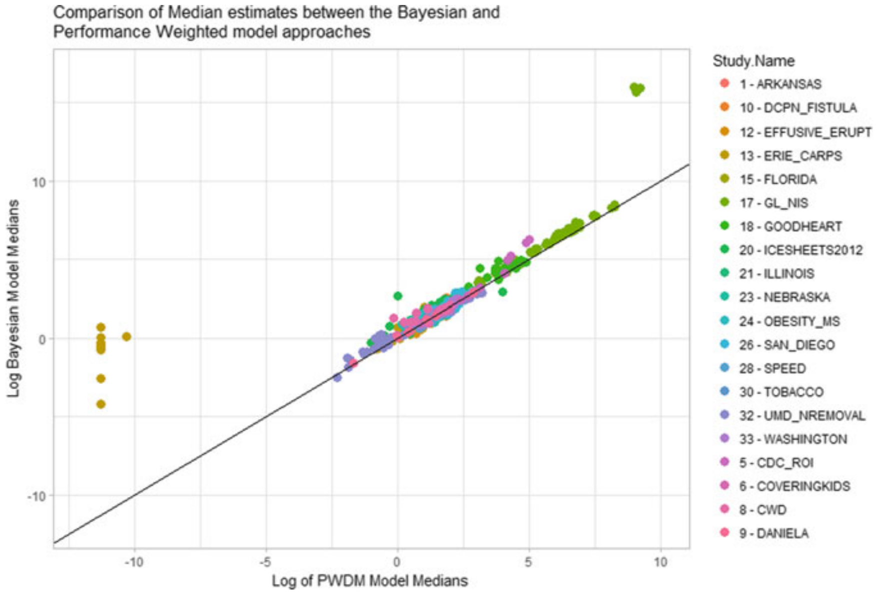


Fig. 5.9 Median estimates from the final decision-maker distributions are highly correlated between the Bayesian and performance-weighted approaches across studies within the Delft database

scales in different studies. There are a small number of outliers at either end of the plot, from the Lake Erie study and the GL_NIS study, where the Bayesian model has a final median of a significantly different order of magnitude to the PWDM. Those at the lower end, from Lake Erie, have been driven by the fact that the performance weighting approach selected a single expert who had a value for these variables many orders of magnitude lower than some of their compatriots who were also included in the Bayesian aggregation. On the upper end, the discrepancy is driven by a small number of target variables within the GL_NIS study whose estimated values were many orders of magnitude higher than other target variables. This will have broken the constraint of scale uniformity from the recalibration process within the Bayesian model and potentially projected higher than realistic values here. Aside from this small number of outliers, however, there is broadly good consistency between approaches on this median value. Whilst important for decision-makers, the median is not the only element that is being considered by those utilising the output of expert judgement studies, with the way that uncertainty is being expressed also of critical importance. A recent study on the impact of melting ice sheets, and the subsequent commentary papers, emphasised some of these considerations.

5.5 Ice Sheet Example

Climate change is one of the major issues of the current age and as such is an area where strong scientific insight is fundamental to building the case for necessary political decision-making and public behavioural change. Unsurprisingly, despite the wealth of geological datasets and sophisticated models, understanding the complexity in and predicting the outcome of many climate change problems relies heavily on expert judgement. Willy Aspinall, who performed the effusive eruptions elicitation study, alongside Jonathan Bamber, conducted a glaciological study to predict the impact of melting ice sheets, due to global warming, on rising sea levels (Bamber and Aspinall 2013).

This research has been extensively cited and has positively contributed to the ongoing debate regarding the appropriate use of expert judgement within the geological community. One commentary by de Vries and van de Wal (2015), and subsequent discussion papers (Bamber et al. 2016; de Vries and van de Wal 2016), assessed and questioned a number of key elements of applying the classical model in this context. In particular:

- The correct way of assessing the lack of consensus in the interpretation of post-processing the experts' answers.
- The reduction in the "effective" number of experts caused by the classical model weighting process.
- The choice of underlying distributions.

Please note. One of the other topics raised in the commentary paper was regarding the choice of variables to elicit from the experts. In this case, the primary elicited variables reflected the experts' predictions on the impact to sea level rise from three separate ice sheets (East Antarctic, West Antarctic and Greenland) and were then combined *post hoc* to create a total sea level rise estimate utilising a Monte Carlo model. Questions were raised whether this would accurately reflect the experts' underlying belief of the final target variable as the choice of model can impact the total uncertainty bounds. Hence, it was suggested that the total sea level rise estimates should have been elicited explicitly. This is a question of study design and we will not tackle it here, except to comment that it is very common for expert judgement to be used both to make judgements on final decision-making variables, or variables which are then inputs into a broader model. The choice selected here may be largely context-dependent. One of the design elements of the Bayesian model (a fully parameterised posterior) is a support tool for decision-makers and analysts utilising the output of expert judgement studies as priors in other models.

Many of these topics are not unique to glaciology and have been commented on elsewhere in the literature with respect to the classical model. The Bayesian hierarchical model, by design, takes a philosophically different approach to each of these areas than the classical model. Thus, whilst it will not address all of the comments posed in de Vries and van de Wal (2016), it would be interesting to consider how the application of the Bayesian aggregation model to the same data performs relative to the performance-weighted approach.

In the effusive eruption and Lake Erie example, we compared some of the forecasts for target variables, however, made no comment as to the validity of the final estimates nor how this varies between models. To be confident in any forecast a decision-maker should have prior validation of a model's results. To this extent, we shall not assess the two models over the target variables within the ice sheet studies as we have done before but will look for ways of assessing how well the models perform (in this context and the previous studies) using some cross-validation techniques.

5.6 Cross-Validation

Measurable target variable realisations are uncommon within expert judgement studies due to the inherent rarity of events assessed or the lack of ethical means of collecting data. These are the same drivers which lead to the studies in the first instance. Consequently, standard models of assessing forecast accuracy, e.g. Out-of-sample validation, are rarely feasible. There are two primary concerns when designing a validation framework for expert judgement studies. Firstly, how to generate a significant sample of testing data for which there are both modelled aggregate judgements and realisations, and secondly, what testing methodology to use to assess the validity of the final distributions on this testing set.

Validation within SEJ models is relatively new, and there is much work to do in order to formally define an agreed-upon approach. Several different methods of building the testable set have been proposed, all of which fall into the context of cross-validation. Cross-validation involves taking only judgements about seed variables (considered as these are the variables against which both judgements have been made and true realisations are known) and permuting through certain subsets of these, using each subset as a training set and then modelling the remaining subset. Clemen (2008) proposed such a technique using a method known as ROAT (Remove One At A Time). Here, each seed variable is removed from the training set one at a time, and all remaining variables are used to train the model, i.e. if there are S seed variables in the data set, each training set will be of size, S-1, and there will be S final forecasts. ROAT is a fast method of cross-validation as relatively few judgements need to be made; however, it was demonstrated that this method could have an inherent bias against a performance-weighted decision-maker (Cooke 2008). Other methods of cross-validation considered have utilised bigger training subsets (Colson and Cooke 2017; Lin and Cheng 2009; Flandoli et al. 2011) and (Cooke et al. 2014), although questions arose over the implementation of a couple of these studies as the numbers quoted did not align with those from Cooke's modelling platform EXCALIBUR (Cooke 2016; Cooke and Solomatine 1992).

Arguably the most comprehensive cross-validation of PWDM was outlined in Eggstaff et al. (2014). Within this cross-validation model, the authors considered every permutation of the seed variable partitions from 1 to S-1 (the code was also vetted against EXCALIBUR). Modelling this level of data showed strong support for the advantages of a PWDM model over an EWDM model but relied on an extremely large number of forecasts which would be a struggle to replicate at scale for other modelling approaches. Given that the number of subsets of a set of size n is 2^n, for a single study of 10 variables, this would create 1022 forecasted subsets (both the empty set and the complete set are removed). Given that each subset forecast within the Bayesian model can take a few minutes to complete, computation cost of this number of forecasts is very high (speed is definitely a distinct advantage of the PWDM over Bayesian MCMC approaches). Colson and Cooke (2017), whilst building on the work of Eggstaff et al. (2014), have recently recommended considering all permutations of training subsets 80% of the size of the original set of seed variables. This creates a manageable sized set of forecasts to perform whilst overcoming some of the biases in the ROAT methodology. For all subsequent analysis within this chapter, we have utilised this 80% methodology. For a study of size 10, there are 45 training subsets of size 8 and 90 resultant forecasts (two for each model run). If 80% was non-integer, we have shrunk the training set size to the nearest integer, and where necessary the minimum number of variables removed was set to 2 to ensure the methodology was not applying a ROAT process.

Methodologies for assessing the accuracy of the forecasts on the given testable sets also vary, and there is further opportunity for research and consolidation on an agreed approach here. One simple method that is considered in many studies is to ignore the uncertainty bounds within the model and simply assess the median within the distribution, utilising a metric such as the mean average percentage error (MAPE), assuming that this represents the most likely value a decision-maker would use in practice. This gives an indicative value on the discrepancy between the forecasts and the actual realisations for these point estimates but does not assess the full richness of the analysis conducted.

The aim of other methods is to quantitatively assess how representative the full distribution is compared to the observed phenomena, within the test sets. One such method is to consider a reapplication of the classical model itself (Eggstaff et al. 2014; Colson and Cooke 2017). Here, each modelling type is considered an "expert", the testable set is the set of seed variables and target variables are omitted. The performance measures (statistical accuracy and information) are then calculated for each model across all of the forecasted variables considered. Typically, for a given study, each subset is assessed in this way and then aggregate statistical accuracy and information scores are calculated for the total study by taking the mean or median of the scores for each subset. Geometric means are also calculated but the arithmetic mean is the value most commonly utilised.

Applying this cross-validation technique to the three studies, we have discussed within this chapter and using the 80% subset rule results in 527 separate subsets and 1529 individual forecasts. Statistical accuracy and information scores for these forecasts are then calculated within R. The R code was validated by taking a sample of these forecasts, rebuilding it from scratch within EXCALIBUR and ensuring consistency of the output. All numbers for Cooke's classical model (PWDM, and EWDM when relevant) have been drawn directly from Eggstaff et al. (2014) supplementary material, kindly provided by Roger Cooke, in order to ensure consistency. Mean statistical accuracy scores, Fig. 5.10, show that the Bayesian model (0.53 effusive eruption, 0.54 Lake Erie, 0.57 ice sheets) scored higher in each study than the PWDM (0.29, 0.45, 0.31). Conversely, Cooke's model (1.6, 0.85, 1.01) performed better than the Bayesian model (0.98, 0.38, 0.63) according to the information criteria outlined. This highlights exactly the behaviour we might expect to see, given the distributions we saw earlier, the fatter tails of the Bayesian model as a result of calibration, and the inherent trade-off made within Cooke's model.

Perhaps more surprising is the performance relative to the EWDM, which has also been included in Fig. 5.10 to provide another reference point. Across the studies

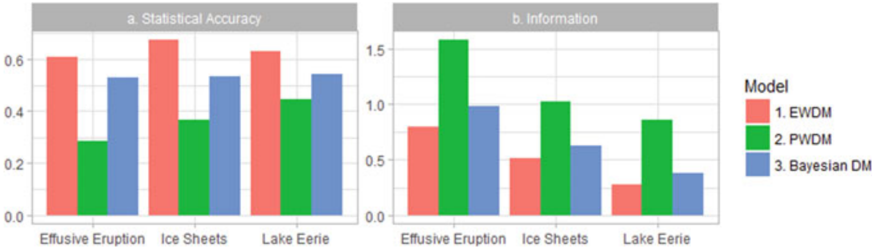


Fig. 5.10 Arithmetic mean of the statistical accuracy and information scores for each tested model across the three studies previously discussed. The Bayesian model typically demonstrates better statistical accuracy but lower information than the PWDM, as we would expect. Perhaps surprisingly, given the broader support, in these studies, the Bayesian model demonstrates higher informativeness than the EWDM. This is due to the Bayesian model having narrower shoulders. Please note that information is a relative measure and absolute informativeness numbers are not relevant cross studies and should only be considered across models within a single study

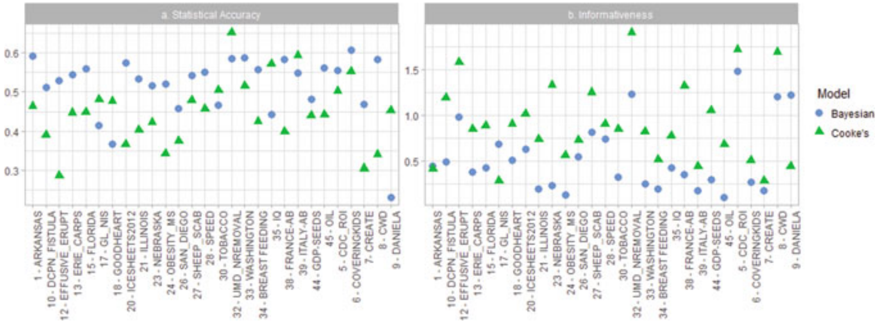


Fig. 5.11 Statistical accuracy and information plots for each analysed study within the Delft database. The Bayesian model demonstrates consistently higher statistical accuracy than the PWDM but lower information scores

outlined, the Bayesian model had higher information scores than the EWDM (0.80, 0.28, 0.52) but lower statistical accuracy scores (EWDM; 0.61, 0.63, 0.35). This may seem counterintuitive, as the Bayesian model has been specifically recalibrated, whereas the EWDM has not. The reason for this behaviour is the consensus focus that the Bayesian model has, rather than diversity which is emphasised in the EWDM approach. Despite the fatter tails, by looking for a consensus view, the Bayesian model typically has narrower shoulders than the EWDM, as we have seen in some of the earlier distributions, e.g. Fig. 5.5. Narrower shoulders are likely to reduce the statistical accuracy score but increase information. In this way, the Bayesian model is also trading off between statistical accuracy and information. If we were to only apply the recalibration component of the Bayesian framework and then aggregate utilising the EWDM, we should expect to see the highest statistical accuracy scores but the lowest information of any of the models discussed so far.

Expanding the cross-validation technique to the broader set of studies within the Delft database can help us ascertain whether we see the above behaviour consistently. As the recalibration within the Bayesian model cannot currently deal with variables on different scales, a subset of 28 studies which were utilised by Eggstaff and within the Delft database were considered. In each considered study, all of the variables were of similar order of magnitude. Study names align with those in the original paper. In total, this equated to 2706 forecasted subsets and 6882 individual variable forecasts (Fig. 5.11).

The Bayesian model outperformed the PWDM on statistical accuracy in 71.4% (20 out of the 28) studies based on the arithmetic mean. The PWDM outperformed the Bayesian model in mean informativeness in 93% (26 of the 28) of the cases. This is reassuring as it demonstrates that the model behaves consistently across studies relative to the PWDM and aligns with what we saw earlier. One of the studies (Study 9—DANIELA) is clearly an outlier with an extremely low statistical accuracy and high information for the Bayesian model. This is because there was a convergence issue with this model, believed to be due to the combination of a low number of experts (4) and seed variables (7, given the holdout sample, only 5 of which would be included in each subset). Whilst more work is necessary to understand the impact of the number of seed variables on expert judgement models, performance weighting guidelines suggest that at least 10 seed variables are considered. Bayesian models with recalibration will similarly require minimum numbers to reach appropriate convergence which meaningfully reflects underlying expert bias.

The above analysis highlights that the Bayesian model and PWDM model are trading off between statistical accuracy and information to different degrees. We would argue the choice of which model to use in practice for a decision-maker may depend on the context in which the study is being performed and the sensitivity of the decision they are making to either information or statistical accuracy. To get a better sense, however, whether the trade-off that the Bayesian model is making is reasonable, we can consider the combination score, as per Cooke's performance weighting method. Here, the statistical accuracy and information scores are multiplied together to give a combined score. This metric for cross-validation is based on the same motivations that lie behind performance weighting. It is thus important to ensure that it is not biased towards a performance-weighted decision-maker. More research is needed to confirm this is the optimal unbiased cross-validation approach. We note this challenge and agree that more work should be done to define a set of cross-validation metrics and processes that are independently ratified, model agnostic and applied consistently to such studies. In the short term, however, this does remain the best available approach and gives us access to a body of knowledge built in the previously listed studies for comparison. Rather than considering the aggregate combined score, which may mask some of the underlying behaviour, we will consider the combined score of each forecasted subset for each study. Figure 5.12 plots the combined score of the PWDM versus the Bayesian decision-maker and an $x = y$ line to help identify relative performance.

This plot highlights a number of interesting elements about the performance of the two models across these subsets. Firstly, of keynote, is that across many of the studies outlined, a significant portion of the subset forecasts sit above the line $x = y$ (e.g. Study 23 or Study 35). This implies that for these studies the PWDM has outperformed the Bayesian model on aggregate whilst considering such a combination measure. Whilst this might appear disheartening for the Bayesian framework, it provides further evidence of the robustness of a performance-weighted approach, which should be admired for its consistent ability to stand up to scrutiny and can provide further reassurance for those who have relied on this model over the past 3 decades. On the positive side for the Bayesian framework, however, is that there are studies in which the mass of points have been more balanced (e.g. Study 1, Study 27 and Study 28) and a few studies in which the Bayesian model appears to be a better predictor across the given subsets (e.g. Study 1, Study 20, Study 8). In fact, there is only one study (Study 23) in which the Bayesian model did not outperform the PWDM on some subset of the seed variables when we consider a combination metric. In total across the 2706 subsets, the Bayesian model outperformed the performance-weighted model in approximately a third of cases (912) with the PWDM demonstrating higher combination scores in 1794 subsets. It is reassuring for the Bayesian approach that there

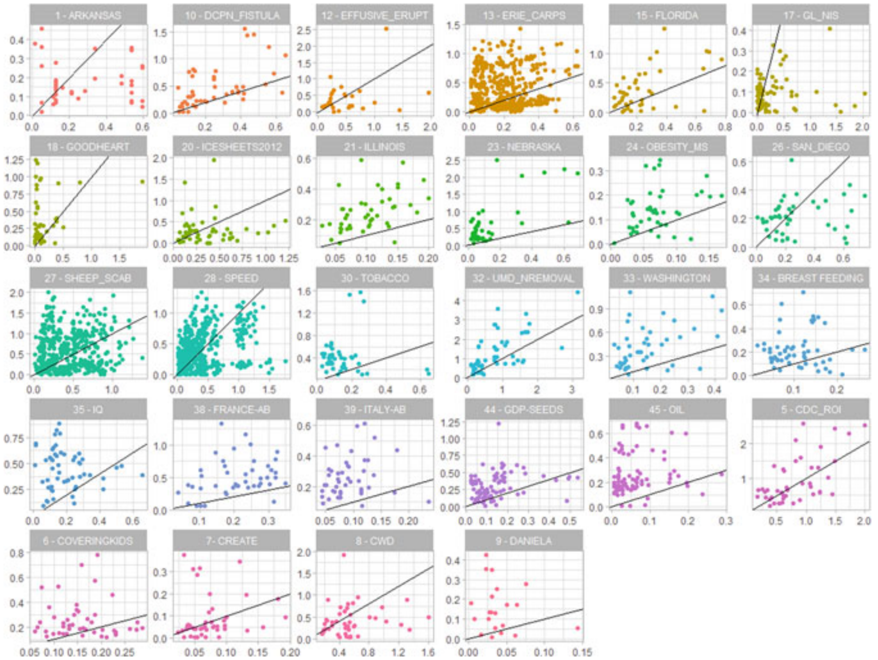


Fig. 5.12 A plot of the combination scores for each analysed study subset. The performance-weighted model (y-axis) demonstrates higher combination scores than the Bayesian model (x-axis) as a significant mass of the points are above the x = y line. There are studies, however, e.g. Study 20 (the ice sheets example), where the Bayesian model typically has higher combination scores

is a substantial number of cases where the model can meet the aims of providing a consensus distribution which is fully paramaterised, whilst performing well against the PWDM when considering a combined statistical accuracy and information score. To be a fully viable model, however, more research is required to understand the drivers of what causes certain combinations to perform better in the Bayesian context than others. One potential option, originally posited in Hartley and French (2021), is that performance here could be linked to the number of experts/seed variables present within the study. This assessment is left for future research.

5.7 Discussion

This chapter has outlined the application of the Bayesian approach to aggregating expert judgements, and its ability to supplement existing models, by:

- Assessing the extent to which experts display systematic over or under confidence.

- Minimising potential overconfidence for the decision-maker that arises from the impact of correlation between expert judgements driven by shared knowledge and common professional backgrounds
- Emphasising the underlying consensus between experts whilst reflecting the diversity of judgements.
- Providing a fully parametrised posterior distribution that is easy to integrate into further analysis.

The framework has been assessed in detail against a small number of studies and then at a macro-level across many studies within the Delft database.

This analysis has shown that such new Bayesian frameworks can be practical, unlike many preceding Bayesian approaches, and can be implemented without a significant overhead in defining complex priors. Utilising relatively diffuse priors (consistent across studies) has been shown to provide results on a similar order of magnitude to current approaches. This would also support the potential of applications of the Bayesian approach in contexts where the aggregate distribution is designed to emulate a rational scientist's perspective in addition to those where a specific decision-maker , potentially with significant a priori beliefs and consequently tighter priors, exists.

The outputs of a Bayesian model of expert judgement have been compared across studies to the performance weighting approach of Cooke's classical model. This comparison has shown that the resultant outputs of the Bayesian approach typically do not vary substantially from the performance-weighted approach when only the median point is considered, however, emphasise a different perspective of the uncertainty. Consistent with other analysis of the Bayesian approach (Hartley and French 2021), the Bayesian model displays a unimodal posterior, with narrower shoulders than an equal-weighted approach (as it emphasises underlying consensus) and has fatter tails than the performance-weighted approach (as it usually highlights systemic overconfidence of experts).

Through cross-validation, we have shown that, as we might expect a priori given its structure, the Bayesian model demonstrates higher statistical accuracy than the performance-weighted approach, but lower informativeness. This suggests that based on the decision-making context the potential sensitivity to each of these metrics may impact the choice of model considered.

Finally, by considering the single combined score metric (the product of the information and statistical accuracy), we have seen that the performance-weighted approach once again stands up to scrutiny and outperforms the Bayesian framework, when configured in this particular way, in the majority (circa 2/3) of cases. There are, however, a substantial number of cases (circa 1/3) for which the Bayesian model outperforms the performance-weighted approach lending credibility to the usage of the Bayesian model in general.

Overall, this chapter has demonstrated that the goal of a practical generic Bayesian framework for mathematical aggregation of expert judgement is feasible and can produce reasonable results when compared to current best in class approaches even when considered broadly with a single set of parametrisations/priors. Much more work is required to assess:

- The impact of the number of seed variables/experts.
- Different parametrisations and priors within the generic framework.
- Approaches for dealing with variables on different scales.
- The drivers of out/underperformance relative to performance-weighted approaches.

However, we have now shown that there is sufficient evidence that the application of resources to assessing these areas is justified.

The performance-weighted approach outlined by Cooke clearly remains the exemplar in this space for many applications; however, we now have a Bayesian approach which can provide a different perspective, add value for decision-makers with specific needs and which we hope will continue to evolve and challenge the performance-weighted method.

5.8 Supplementary Material

Table 5.3 Target variable predicted quantiles for the effusive eruption study across models

	EWDM			PWDM			BDM		
Target variable	0.05	0.5	0.95	0.05	0.5	0.95	0.05	0.5	0.95
10x SO2 probability	0.02	2.76	32.19	0.24	5.03	19.2	−0.41	0.99	39.44
Column height	5.24	13.61	22.68	6.52	12.87	22.22	−2.96	13.47	31.67
Avg plume height	0.6	3.79	11.92	0.58	4.05	11.26	−2.95	4.22	13.06
Max plume height	0.53	3.93	13.2	0.85	3.87	15.78	−3.63	3.98	16.74
Max% SO2 emissions	53.04	86.18	99.74	50	75.63	97.45	32.63	86.15	116.77
Min% SO2 emissions	25.09	70.89	89.96	40.81	68.2	84.61	10.27	71.21	118.77
Max no. fissures	3.88	26.15	472.4	6.08	18.45	98.43	−25.05	28.77	174.32
Min no. fissures	0.12	2.95	13.6	1.19	6.6	16.83	−4.31	3.55	20.27
Duration explosive phase	0.14	2.88	15.84	0.34	5.64	23.79	−4.3	2.86	21.44
Gap between outbursts	0.12	7.17	183.4	0.51	5.71	29.87	−2.17	4.37	163.99

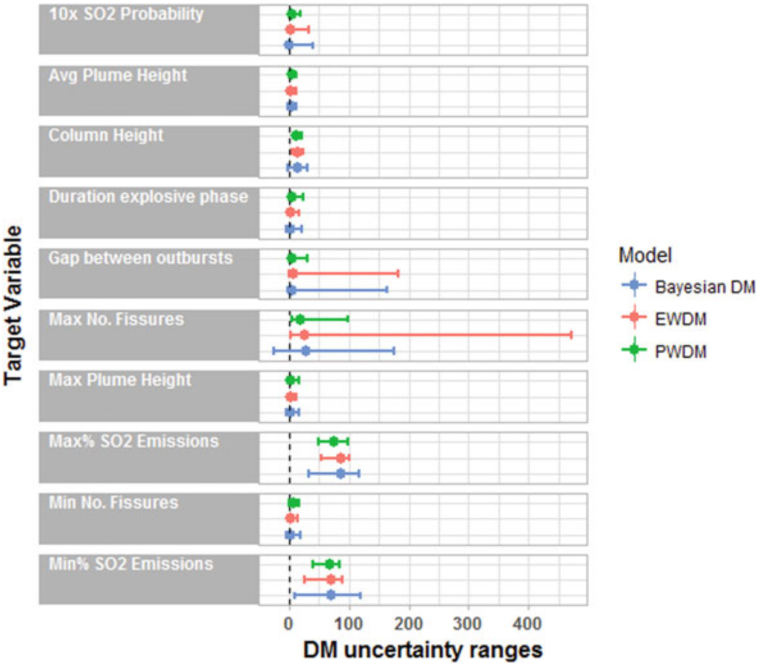


Fig. 5.13 Replication of Fig. 5.4 including the EWDM. The Bayesian model displays posterior uncertainty ranges consistently broader than the PWDM, however, displays uncertainty bounds both broader and narrower than the EWDM

References

Albert, I., Donnet, S., Guihenneuc-Jouyaux, C., Low-Choy, S., Mengersen, K., & Rousseau, J. (2012). Combining expert opinions in prior elicitation. *Bayesian Analysis* (pp. 503–532).

Bamber, J. L., & Aspinall, W. P. (2013). An expert judgement assessment of future sea level rise from the ice sheets. *Nature Climate Change, 3*(4), 424.

Bamber, J. L., Aspinall, W. P., & Cooke, R. M. (2016). A commentary on how to interpret expert judgment assessments of twenty-first century sea-level rise by Hylke de Vries and Roderik SW van de Wal" *Climatic Change, 137*(3–4), 321–328.

Billari, F. C., Graziani, R., Melilli, E. (2014). Stochastic population forecasting based on combinations of expert evaluations within the Bayesian paradigm. *Demography, 51*(5), 1933–1954.

Burgman, M. A. (2015). *Trusting judgements: How to get the best out of experts*. Cambridge University Press.

Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A. & Charrad, M.M. (2014). Package nbclust. *Journal of Statistical Software, 61*, 1–36

Clemen, R. T. (2008). Comment on Cooke's classical method. *Reliability Engineering & System Safety, 93*(5), 760–765.

Clemen, R. T., & Lichtendahl, K. C. (2002). *Debiasing expert overconfidence: A Bayesian calibration model*. PSAM6: San Juan, Puerto Rico.

Colson, A. R., & Cooke, R. M. (2017). Cross validation for the classical model of structured expert judgment. *Reliability Engineering & System Safety, 163*, 109–120.

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

Cooke, R. M. (Ed.). (2007). Expert judgement studies. *Reliability Engineering and System Safety*.

Cooke, R. M. (2008). Response to discussants. *Reliability Engineering & System Safety, 93*(5), 775–777.

Cooke, R. M. (2016). Supplementary Online Material for Cross Validation of Classical Model for Structured Expert Judgment.

Cooke, R. M., & Goossens, L. H. J. (2000). Procedures guide for structured expert judgement in accident consequence modelling. *Radiation Protection Dosimetry, 90*(3), 303–309.

Cooke, R. M., & Goossens, L. H. (2008). TU Delft expert judgment data base. *Reliability Engineering & System Safety, 93*(5), 657–674.

Cooke, R. M., & Solomatine, D. (1992). *EXCALIBUR integrated system for processing expert judgements version 3.0*. Delft: Delft University of Technology and SoLogic Delft.

Cooke, R. M., Wittmann, M. E., Lodge, D. M., Rothlisberger, J. D., Rutherford, E. S., Zhang, H., & Mason, D. M. (2014). Out of sample validation for structured expert judgment of Asian carp establishment in Lake Erie. *Integrated Environmental Assessment and Management, 10*(4), 522–528.

De Finetti, B. (1974). *Theory of Probability*. Chichester: Wiley.

De Finetti, B. (1975). *Theory of Probability*. Chichester: Wiley.

de Vries, H., & van de Wal, R. S. W. (2015). How to interpret expert judgment assessments of 21st century sea-level rise. *Climatic Change, 130*(2), 87–100.

de Vries, H., & van de Wal, R. S. W. (2016). Response to commentary by JL Bamber, WP Aspinall and RM Cooke. *Climatic Change, 137*(3–4), 329–332.

Eggstaff, J. W., Mazzuchi, T. A., & Sarkani, S. (2014). The effect of the number of seed variables on the performance of Cookes classical model. *Reliability Engineering & System Safety, 121*, 72–82.

EFSA. (2010). Statement of EFSA on the possible risks for public and animal health from the contamination of the feed and food chain due to possible ash fall following the eruption of the Eyjafjallaj kull volcano in Iceland. *EFSA Journal, 8*, 1593.

EFSA. (2014). Guidance on expert knowledge elicitation in food and feed safety risk assessment. *EFSA Journal*.

Flandoli, F., Giorgi, E., Aspinall, W. P., & Neri, A. (2011). Comparison of a new expert elicitation model with the Classical Model, equal weights and single experts, using a cross-validation technique. *Reliability Engineering & System Safety, 96*(10), 1292–1310.

French, S. (1980). Updating of belief in the light of someone else's opinion. *Journal of the Royal Statistical Society, A143*, 43–48.

French, S. (1985). Group consensus probability distributions: a critical survey (with discussion). In J. M. Bernardo, M. H. DeGroot, D. V. Lindley & A. F. M. Smith *Bayesian Statistics 2, North-Holland* (pp. 183–201).

French, S. (2011). Aggregating expert judgement. *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales, 105*(1), 181–206.

French, S., Maule, A. J., & Papamichail, K. N. (2009). *Decision behaviour, analysis and support*. Cambridge: Cambridge University Press.

Garthwaite, P. H., Kadane, J. B., & O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association, 100*(470), 680–701.

Hartley, D., & French, S. (2021). A Bayesian method for calibration and aggregation of expert judgement. *Journal of Approximate Reasoning, 130*, 192–225.

Hartley, D., French, S. (2018). Elicitation and calibration: A Bayesian perspective. *Elicitation: The science and art of structuring judgement* (pp. 119–140).

Hockey, G. R. J., Maule, A. J., Clough, P. J., & Bdzola, L. (2000). Effects of negative mood on risk in everyday decision making. *Cognition and Emotion, 14*, 823–856.

Hora, S. (2007). Eliciting probabilities from experts. In W. Edwards, R. F. Miles, & D. Von Winterfeldt, *Advances in decision analysis: From foundations to applications* (pp. 129–153). Cambridge: Cambridge University Press.

Kadane, J. B., & Fischhoff, B. (2013). A cautionary note on global recalibration. *Judgment and Decision Making, 8*(1), 25–27.

Kahneman, D. (2011). *Thinking, fast and slow*. London: Penguin, Allen Lane.

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

Lichtendahl, K. C. (2005). Bayesian models of expert forecasts. Ph.D. thesis.

Lichtendahl, K. C., & Winkler, R. L. (2007). Probability elicitation, scoring rules, and competition among forecasters. *Management Science, 53*(11), 1745–1755.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic & A. Tversky *Judgement under uncertainty* (pp. 306–334). Cambridge: Cambridge University Press.

Lin, S.-W., & Bier, V. M. (2008). A study of expert overconfidence. *Reliability Engineering and System Safety, 93*, 711–721.

Lin, S.-W., & Cheng, C.-H. (2009). The reliability of aggregated probability judgments obtained through Cooke's classical model. *Journal of Modelling in Management, 4*(2), 149–161.

Loughlin, S. C., Aspinall, W. P., Vye-Brown, C., Baxter, P. J., Braban, C., Hort, M., et al. (2012). Large-magnitude fissure eruptions in Iceland: Source characterisation. *BGS Open File Report, OR/12/098*, 231pp. Retrieved from http://www.bgs.ac.uk/research/volcanoes/LakiEruptionScenarioPlanning.html.

Mumpower, J. L., & Stewart, T. R. (1996). Expert judgement and expert disagreement. *Thinking and Reasoning, 2*(2–3), 191–211.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, R., Garthwaite, P. H., Jenkinson, D., et al. (2006). *Uncertain judgements: Eliciting experts' probabilities*. Chichester: Wiley.

Shanteau, J. (1995). Expert judgment and financial decision making. *Risky business: Risk behavior and risk management*. B. Green: Stockholm, Stockholm University.

Skjong, R., & Wentworth, B. H. (2001). Expert judgement and risk perception. In *Proceedings of the eleventh (2001) international offshore and polar engineering conference*. Stavanger, Norway: The International Society of Offshore and Polar Engineers.

Wilson, K. J. (2016). An investigation of dependence in expert judgement studies with multiple experts. *International Journal of Forecasting*.

Wiper, M. W., & French, S. (1995). Combining experts' opinions using a normal-Wishart model. *Journal of Forecasting, 14*, 25–34.

Wittmann, M. E., Cooke, R. M., Rothlisberger, J. D., & Lodge, D. M. (2014). Using structured expert judgment to assess invasive species prevention: Asian Carp and the Mississippi-Great Lakes Hydrologic Connection. *Environmental Science & Technology, 48*(4), 2150–2156.

Wittmann, M. E., Cooke, R. M., Rothlisberger, J. D., Rutherford, E. S., Zhang, H., Mason, D. M., et al. (2015). Use of structured expert judgment to forecast invasions by bighead and silver carp in Lake Erie. *Conservation Biology, 29*(1), 187–197.

Zhang, H., Rutherford, E. S., Mason, D. M., Breck, J. T., Wittmann, M. E., Cooke, R. M., et al. (2016). Forecasting the impacts of silver and bighead carp on the Lake Erie food web. *Transactions of the American Fisheries Society, 145*(1), 136–162.

Chapter 6
Three-Point Lifetime Distribution Elicitation for Maintenance Optimization in a Bayesian Context

J. René van Dorp and Thomas A. Mazzuchi

Abstract A general three-point elicitation model is proposed for eliciting distributions from experts. Specifically, lower and upper quantile estimates and a most likely estimate in between these quantile estimates are to be elicited, which uniquely determine a member in a flexible family of distributions that is consistent with these estimates. Multiple expert elicited lifetime distributions in this manner are next used to arrive at the prior parameters of a Dirichlet Process (DP) describing uncertainty in a lifetime distribution. That lifetime distribution is needed in a preventive maintenance context to establish an optimal maintenance interval or a range thereof. In practical settings with an effective preventive maintenance policy, the statistical estimation of such a lifetime distribution is complicated due to a lack of failure time data despite a potential abundance of right-censored data, i.e., survival data up to the time the component was preventively maintained. Since the Bayesian paradigm is well suited to deal with scarcity of data, the formulated prior DP above is updated using all available failure time and right-censored maintenance data in a Bayesian fashion. Multiple posterior lifetime distribution estimates can be obtained from this DP update, including, e.g., its posterior expectation and median. A plausible range for the optimal time-based maintenance interval can be established graphically by plotting the long-term average cost per unit time of a block replacement model for multiple posterior lifetime distribution estimates as a function of the preventive maintenance frequency. An illustrative example is utilized throughout the paper to exemplify the proposed approach.

Keywords Expert judgment · Generalized two-sided power distribution · Dirichlet process · Bayesian inference

J. R. van Dorp (✉) · T. A. Mazzuchi
School of Engineering and Applied Science, The George Washington University, Washington D.C., USA
e-mail: dorpjr@gwu.edu

T. A. Mazzuchi
e-mail: mazzu@gwu.edu

A. M. Hanea et al. (eds.), *Expert Judgement in Risk and Decision Analysis*,
International Series in Operations Research & Management Science 293,
https://doi.org/10.1007/978-3-030-46474-5_6

6.1 Introduction

Maintenance optimization has been a focus of research interest for quite sometime. Dekker (1996) provides an elaborate review and analysis of applications of maintenance optimization models. *"... many textbooks on Operations Research use replacement models as examples"* (Dekker 1996). More recently, Mazzuchi et al. (2007) present a review of mathematical decision models to optimize condition-based maintenance and time-based maintenance. The block replacement model utilized in this paper falls in the latter category. Often the main bottleneck in the implementation of maintenance optimization procedures is the lack of the life length distributions of system components due to scarcity of component failure data. This phenomenon is inherent to an efficient preventive maintenance environment aimed at avoiding those component failures in the first place. One approach to overcome this lack of failure data is to determine the lifetime distribution based on the additional use of multiple expert judgment estimates followed by the combination thereof using a linear opinion pool (see, e.g., Cooke 1991).

As a rule experts classify into two usually unrelated groups: (1) substantive experts (also known as technical experts or domain experts) who are knowledgeable about the subject matter at hand and (2) normative experts possessing knowledge of the appropriate quantitative analysis techniques (see, e.g., Dewispelare et al. 1995; Pulkkinen and Simola 2000). In principle, in the absence of data, normative experts are tasked with specifying distributions that are consistent with the substantive expert's judgment. Substantive experts in the case of eliciting lifetime distributions for preventive maintenance optimization are those reliability engineers that have maintained the components in question.

These substantive experts, however, may not be statistically trained. To facilitate elicitation, graphically interactive and statistical elicitation procedures for distribution modeling have been developed. For example, AbouRizk et al. (1991) developed software with a graphical user interface (GUI) to ease fitting of beta distributions using a variety of methods, and DeBrota et al. (1989) have developed software for fitting bounded Johnson S_B distributions. Wagner and Wilson (1996) introduced univariate Bézier distributions (or curves), which are a variant of spline functions, and the software tool $PRIME$ with a GUI to specify them. These methods all focus on the indirect elicitation of a continuous density function. van Noortwijk et al. (1992) suggested, in contrast, the elicitation of a discrete lifetime distribution using a method referred to as the histogram technique. van Dorp (1989) developed a software implementation of this technique, also with a GUI interface, but most importantly with expert feedback analysis with the intent to reduce elicitation bias. The histogram technique is reminiscent of the fixed interval method suggested in Garthwaite et al. (2005) or the Roulette-method implemented in $MATCH$, a more recent web-based tool developed by Morris et al. (2014) for eliciting probability distributions that implements the SHeffield ELicitation Framework (SHELF) (Oakley and O'Hagan 2018).

Herein, the three-point elicitation of a lower quantile x_p, a most likely estimate η and an upper quantile estimate x_r is suggested, given a specified support $[a, b]$, for the elicitation of a lifetime distribution from reliability engineers. In the SHELF protocol, *plausible ranges* are first directly elicited from the substantive experts. Through a behavioral aggregation process using a facilitator, minimum lower and upper bounds $L_{\min}$ and $U_{\max}$ are established. The range $(L_{\min}, U_{\max})$ next serves as the common support in SHELF (Oakley and O'Hagan 2018) when fitting a bounded distribution to an individual expert's elicited quantiles in that group process. In Cooke's classical method (Cooke 1991), an *intrinsic range* is determined using the $k\%$ overshoot rule, where k is an integer value selected by a normative expert, e.g., $k = 1, 5$ or 10. Denoting the minimum (maximum) quantile elicited among a group of substantive experts by $x_{\min}$ ($x_{\max}$), the intrinsic range next follows as (L, U) where

$$L = x_{\min} - (x_{\max} - x_{\min}) \cdot k/100 \text{ and } U = x_{\max} + (x_{\max} - x_{\min}) \cdot k/100.$$

For the elicitation procedure suggested herein, however, it shall be demonstrated that such fixed common support $[a, b]$ can be set arbitrarily large, e.g., by a normative expert. The latter avoids having to use, e.g., (i) the overshoot percentage approach in Cooke's classical method (Cooke 1991) to arrive at that common support, or (ii) the elicitation of a common support from multiple experts, as is case for the SHELF protocol (Oakley and O'Hagan 2018) when eliciting a bounded distribution.

As in SHELF (Oakley and O'Hagan 2018), the quantile levels p and r of the lower quantile x_p and upper quantile x_r are free to be specified, although in the illustrative example used herein the values are set at $p = 0.20$ and $r = 0.80$, with the requirement that the most likely estimate η falls between these quantiles (otherwise, lower quantile levels can be used). In *MATCH* (Morris et al. 2014), for example, the elicitation of quartiles and tertiles has been proposed. The larger the tail probabilities implied by the quantile levels, the more likely it is that substantive experts may have experienced observations in those tails. The smaller these implied tail probabilities, on the other hand, the larger the probability mass, and its range, that is specified by the substantive experts. While both are desirable objectives, the specification of these quantile levels in any elicitation method that utilizes such lower and upper quantile estimates involves a trade-off between these two objectives.

Using the quantile estimates x_p and x_r with the most likely estimate η, $a < x_p < \eta < x_r < b$, a five-parameter Generalized Two-Sided Power ($GTSP$) distribution (Herrerías-Velasco et al. 2009) is fitted, consistent with these estimates, with probability density function (pdf)

$$f(x|\Theta) = C(\Theta) \times \begin{cases} \left(\frac{x-a}{\eta-a}\right)^{m-1}, & \text{for } a < x < \eta, \\ \left(\frac{b-x}{b-\eta}\right)^{n-1}, & \text{for } \eta \leq x < b, \end{cases} \tag{6.1}$$

where $\Theta = \{a, \eta, b, m, n)$ and for $m, n > 0$,

$$C(\Theta) = \frac{mn}{(\eta - a)n + (b - \eta)m}.$$

The uniform ($n = m = 1$), the triangular distribution ($n = m = 2$), and the Two-Sided Power distribution ($n = m$) (van Dorp and Kotz 2002) are members within the $GTSP$ family. Herrerías-Velasco et al. (2009) suggested the family of $GTSP$ distributions as a more flexible alternative to the four-parameter beta distribution with support (a, b) and pdf

$$\begin{cases} f(x|a, b; \alpha, \beta) = \frac{(b-a)^{1-(\alpha+\beta)}}{B(\alpha,\beta)} \times (x - a)^{\alpha-1}(b - x)^{\beta-1}, \\ a < x < b, \ \alpha, \beta > 0, \end{cases} \tag{6.2}$$

where the beta constant $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta)$. For example, a lower and upper quantile $a < x_p < x_r < b$, with specified density support (a, b), uniquely determines the beta pdf (6.2) (see, van Dorp and Kotz 2000; Shih 2015), whereas the $GTSP$ family allows for the additional specification of a most likely value η, $a < x_p < \eta < x_r < b$. Moreover, the moment ratio coverage diagram in Fig. 6.1, plotting kurtosis β_2 against $\sqrt{|\beta_1|}$ with the convention that $\sqrt{|\beta_1|}$ retains the sign of the third central moment, demonstrates a larger coverage for $GTSP$ distributions (6.1) in the unimodal domain than the beta distributions (6.2). This added flexibility allows one to numerically solve for the unique member in the $GTSP$ family of distributions that matches the expert judgment estimates x_p and x_r and most likely value

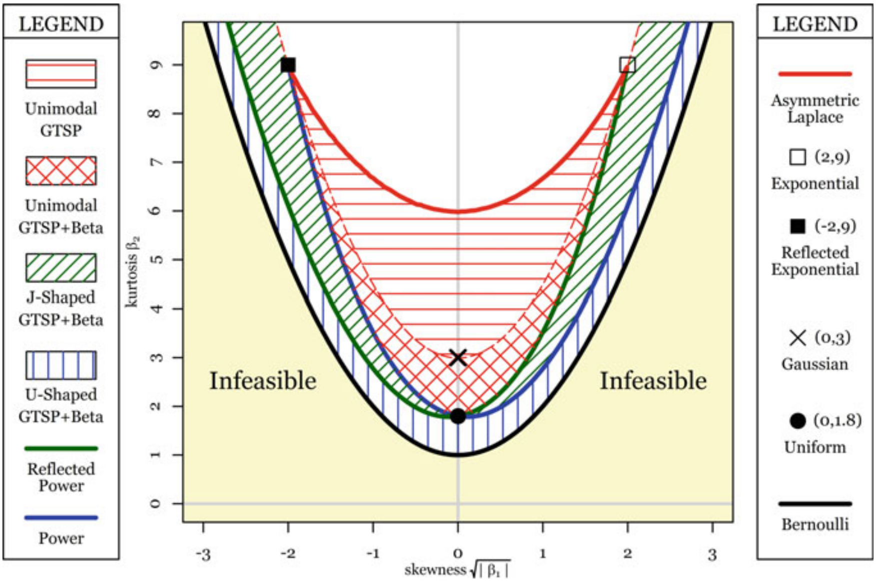


Fig. 6.1 Moment ratio ($\sqrt{\beta_1}$, β_2) coverage diagram for $GTSP$ distributions (6.1) and *beta* distributions (6.2)

estimate η, $a < x_p < \eta < x_r < b$. Note that in the J-shaped and U-shaped domains Fig. 6.1 demonstrates the same coverage for the pdfs (6.1) and (6.2).

In Sect. 6.2, the $GTSP$ fitting procedure to the expert judgment estimates x_p and x_r and most likely value estimate η, $a < x_p < \eta < x_r < b$ is presented. The uniqueness of the GTSP member distribution that exactly matches those estimates is proven in the Appendix. In Sect. 6.3, the use of an equal-weighted mixture of elicited $GTSP$ distributions is suggested to construct a prior cumulative distribution function "parameter" for a Dirichlet Process (DP) (see, e.g., Ferguson 1973). Of course, alternative weighting schemes for experts, such as, e.g., performance-based weights as in Cooke's classical method (Cooke 1991), can be utilized in this procedure, but barring additional information setting equal weights is minimally informative from an expert weighting perspective. Furthermore, in the Bayesian context herein, the prior information, hopefully, will ultimately be outweighed by observed data. Next, for parametric convenience, a generalized trapezoidal distribution (GT) (van Dorp and Kotz 2003) is fitted to this mixture of expert elicited/fitted distributions, although this is not a requirement for the application of the prior formulation approach for the DP. The uncertainty in the DP is governed by a single DP parameter denoted $\alpha(\mathbb{R}^+)$. A conservative procedure is proposed in Sect. 6.3 to estimate this DP parameter $\alpha(\mathbb{R}^+)$ from the expert elicited $GTSP$ distributions while allowing for the largest uncertainty in the DP given these elicited distributions. Larger uncertainty in the DP ensures a larger effect of all available data (both failure and maintenance data) when updating the prior information in the DP in a Bayesian fashion.

Posterior component lifetime probability distribution estimates are obtained in Sect. 6.4 using the updating procedures for DP's described in Susarla and Van Ryzin (1976). Utilizing multiple posterior lifetime distribution estimates, including the mean and median estimate of the DP, a plausible range for the optimal maintenance interval is established graphically in Sect. 6.5 by plotting the long-term average cost per unit time per estimate in a block replacement setup presented in Mazzuchi and Soyer (1996) as a function of the preventive maintenance frequency. Throughout the paper, an illustrative example demonstrates and visualizes the analysis procedures using a series of figures. Some concluding remarks are provided in Sect. 6.6.

6.2 Three-Point GTSP Distribution Elicitation

In the procedure below and given a fixed support (a, b) for the pdf (6.1) one can, without loss of generality, first standardize the elicited x_p, x_r and η values to standardized values y_p, y_r and θ in the interval $(0, 1)$ using the linear transformation $(x - a)/(b - a)$. Utilizing that same linear transformation, the pdf (6.1) reduces to

$$f(y|m, n, \theta) = \frac{mn}{(1-\theta)m + \theta n} \times \begin{cases} \left(\frac{y}{\theta}\right)^{m-1}, & \text{for } 0 < y < \theta, \\ \left(\frac{1-y}{1-\theta}\right)^{n-1}, & \text{for } \theta \le y < 1, \end{cases} \tag{6.3}$$

where $0 < \theta < 1$, n, $m > 0$. While the most likely value is elicited directly, the standardized lower and upper quantile estimates y_p and y_r are needed to indirectly elicit the power parameters m and n of the pdf in (6.3), hence, the requirement $0 < y_p < \theta < y_r < 1$. From (6.3), one directly obtains the cumulative distribution function

$$F(y|\Theta) = \begin{cases} \pi(\theta, m, n)\left(\frac{y}{\theta}\right)^m, & \text{for } 0 \leq y < \theta, \\ 1 - [1 - \pi(\theta, m, n)]\left(\frac{1-y}{1-\theta}\right)^n, & \text{for } \theta \leq y \leq 1, \end{cases} \tag{6.4}$$

with mode (or anti-mode) probability $Pr(X \leq \theta) = \pi(\theta, m, n)$, where

$$\pi(\theta, m, n) = \frac{\theta n}{(1-\theta)m + \theta n}. \tag{6.5}$$

From (6.4) and given the quantile estimates y_p, y_r, the following set of non-linear equations (the quantile constraints) needs to be solved to obtain the power parameters m and n in (6.3), (6.4) and (6.5):

$$\begin{cases} F(y_p|\theta, m, n) = \pi(\theta, m, n)\left(\frac{y_p}{\theta}\right)^m = p, \\ F(y_r|\theta, m, n) = 1 - [1 - \pi(\theta, m, n)]\left(\frac{1-y_r}{1-\theta}\right)^n = r. \end{cases} \tag{6.6}$$

In the appendix, it is proven that the quantile constraint set in (6.6) defines a unique implicit function $m^\bullet = \xi(n)$, where $\xi(\cdot)$ is a continuous strictly increasing concave function in n, and the implicit function $\xi(n)$ has the following tangent line at $n = 0$:

$$M(n|p, \theta) = n \times \frac{\theta}{1-\theta} \times \frac{1-p}{p},$$

where for all values of $n > 0$, $M(n|p, \theta) \geq \xi(n)$. Thus, $\xi(n) \downarrow 0$ as $n \downarrow 0$ and the point $(m^\bullet = \xi(n),\ n)$ satisfies the first quantile constraint in (6.6) for all $n > 0$. As a result, when $n \downarrow 0$ the density $f(y|m^\bullet = \xi(n), n, \theta)$ given by (6.3) converges to a Bernoulli distribution with probability mass p at $y = 0$ and probability mass $1 - p$ at $y = 1$. Analogously, it can be proven that the quantile constraint set defines a unique implicit function $n^\bullet = \zeta(m)$, where $\zeta(\cdot)$ is a continuous strictly increasing concave function in m, and the implicit function $\zeta(m)$ has the following tangent line at $m = 0$:

$$N(m|r, \theta) = m \times \frac{1-\theta}{\theta} \times \frac{r}{1-r},$$

where for all values of $m > 0$, $N(m|r, \theta) \geq \zeta(m)$. Thus $\zeta(m) \downarrow 0$ as $m \downarrow 0$ and the point $(m,\ n^\bullet = \zeta(m))$ satisfies the second quantile constraint in (6.6). As a result, when $m \downarrow 0$ the density $f(y|m, n^\bullet = \zeta(m), \theta)$ given by (6.3) converges to a Bernoulli distribution with probability mass r at $y = 0$ and probability mass $1 - r$ at $y = 1$. From these conditions above, it follows that the quantile constraint set in (6.6) has a unique solution $(m^*,\ n^*)$ where $m^*,\ n^* > 0$. A detailed proof is provided in the Appendix.

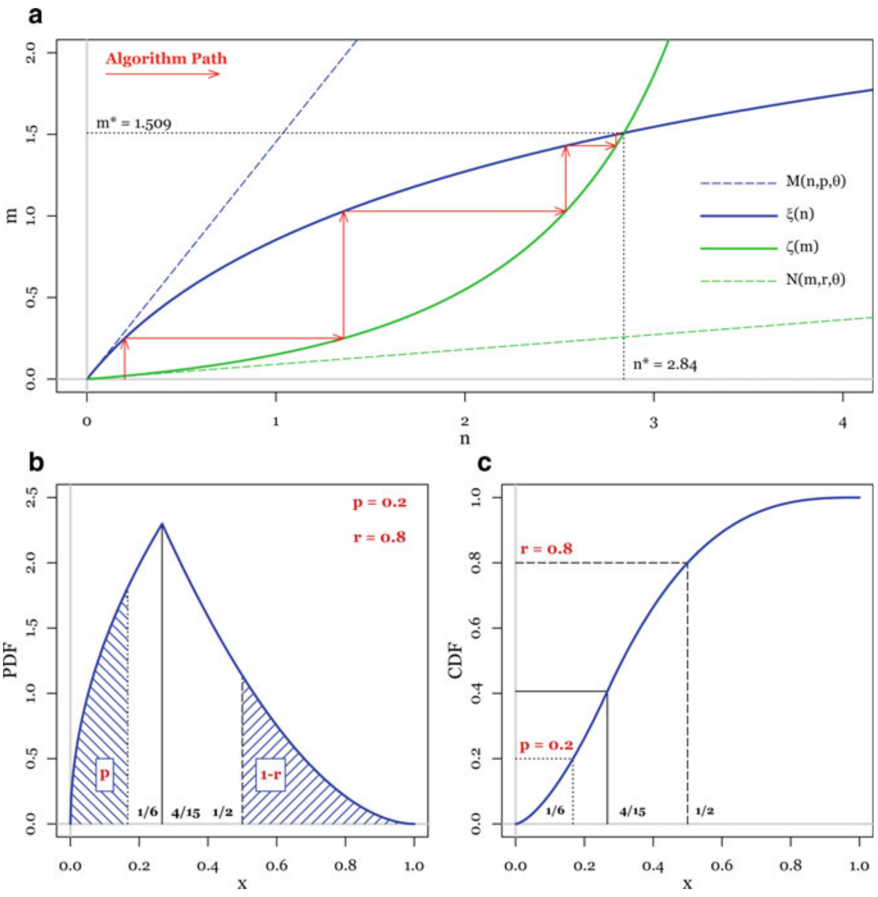


Fig. 6.2 **a**: Implicit functions $\xi(n)$ (in blue) and $\zeta(m)$ (in green) and algorithm path (in red) for the example data (6.7) demonstrating existence and uniqueness of the intersection point between implicit functions $\xi(n)$ and $\zeta(m)$ for $m > 0$ and $n > 0$; **b**: $GTSP$ pdf solution (8); **c**: $GTSP$ cdf solution (8)

The unique solution $m^\bullet = \xi(n)$ for a fixed value $n > 0$ may be solved for by employing a standard root finding algorithm such as, for example, $GoalSeek$ in Microsoft Excel or $uniroot$ in R for a given value of n. Similarly, the unique solution $n^\bullet = \zeta(m)$ may be solved for given a fixed value of $m > 0$. The following direct algorithm now solves the quantile constraints in (6.6):

Step 1: Set $n^\bullet = \delta > 0$.
Step 2: Calculate $m^\bullet = \xi(n^\bullet)$ (satisfying the first quantile constraint in (6.6)).
Step 3: Calculate $n^\bullet = \zeta(m^\bullet)$ (satisfying the second quantile constraint in (6.6)).
Step 4: If $\left|\pi(\theta, m^\bullet, n^\bullet)\left(\frac{y_p}{\theta}\right)^{m^\bullet} - p\right| < \epsilon$ Then Stop Else Goto Step 2.

Figure 6.2a plots the implicit functions $\xi(n)$ and $\zeta(m)$ in a single graph with n on the horizontal axis and m on the vertical axis for the example data:

$$y_p = 1/6, \theta = 4/15, y_r = 1/2, \; p = 0.2, \; r = 0.8. \tag{6.7}$$

At the intersection of $\xi(\cdot)$ and $\zeta(\cdot)$ in Fig. 6.2a, both quantile constraints (6.6) are met for the example data (6.7). The algorithm path toward the unique solution $(m^*, \; n^*)$ of the quantile constraints (6.6) for the example data (6.7) is indicated by a red arrowed line in Fig. 6.2a with solution

$$m^* \approx 1.509 \text{ and } n^* \approx 2.840. \tag{6.8}$$

The $GTSP$ pdf in (6.3) with power parameters in (6.8) is plotted in Fig. 6.2b. The $GTSP$ cdf in (6.4) with power parameters in (6.8) is plotted in Fig. 6.2c.

6.3 Prior Distribution Construction

The expert data for the illustrative example is provided in Table 6.1. Table 6.1A provides the solutions for the power parameters for Experts 1 through 3 given a fixed support [0, 30]. Please observe from Table 6.1A that the standardized data for Expert 1 coincides with the example data (6.7). To demonstrate that the support of the $GTSP$ distribution can be chosen arbitrarily large, by, e.g., a normative expert, given the elicitation procedure described in Sect. 6.2, Table 6.1B (Table 6.1C) provides the solutions for the power parameters for Experts 1 through 3 given a fixed support [0, 100] ([0, 1000]). One observes from Table 6.1A–C that enlarging the support results in larger values for the right-branch power parameter n, whereas the values for the left-branch power parameter m are similar in size. However, being able to choose the support arbitrarily large does not preclude one from utilizing, e.g., (i) the overshoot percentage approach in Cooke's classical method (Cooke 1991) to arrive at a common support for the multiple expert elicited lifetime distributions or (ii) the facilitator approach in the Sheffield protocol (Oakley and O'Hagan 2018) and use that protocol's elicited "plausible range" for that common support.

Figure 6.3 plots the pdfs and cdfs for the three different support solutions provided in Table 6.1. In addition, Fig. 6.3 highlights the locations of the different quantiles in Table 6.1 using vertical dashed and dotted lines and the locations of the most likely values using vertical solid lines. One observes for this example from these vertical lines that the most likely value estimated by Expert 2 (in green in Fig. 6.3) is less than the estimated lower quantile estimates for Expert 1 (in blue in Fig. 6.3) and for Expert 3 (in red in Fig. 6.3). One observes from Fig. 6.3b, d, f that the $GTSP$ solutions provided in Table 6.1 match the quantile estimates at their quantile levels 0.20 and 0.80. One also observes that the most like values for the three expert elicited/fitted $GTSP$ distributions hover around their medians. By visually comparing Fig. 6.3c–d with Fig. 6.3e–f, one observes no concernable difference between the pdfs and cdfs with support [0, 100] and [0, 1000], respectively. A slight difference is observed when comparing the pdfs in Fig. 6.3a (with support [0, 30]) with those in Fig. 6.3c, e. Finally, Fig. 6.3 plots (in light blue) the pdf and cdf of a mixture of

Table 6.1 Expert data for illustrative example with supports A: [0, 30]; B: [0, 100]; C: [0, 1000]

A	**Expert 1**	**Expert 2**	**Expert 3**		**Expert 1**	**Expert 2**	**Expert 3**
a	0	0	0		0	0	0
p	0.2	0.2	0.2	a	0.2	0.2	0.2
r	0.8	0.8	0.8	r	0.8	0.8	0.8
b	**30**	**30**	**30**	b	**1**	**1**	**1**
x_p	5	2	6	y_p	1/6	1/15	1/5
η	8	4	9	θ	4/15	2/15	3/10
x_r	15	7	12	y_r	1/2	7/30	2/5
m	1.509	1.269	2.328	m	1.509	1.269	2.328
n	2.838	7.733	5.755	n	2.838	7.733	5.755
B	**Expert 1**	**Expert 2**	**Expert 3**		**Expert 1**	**Expert 2**	**Expert 3**
a	0	0	0		0	0	0
p	0.2	0.2	0.2	a	0.2	0.2	0.2
r	0.8	0.8	0.8	r	0.8	0.8	0.8
b	**100**	**100**	**100**	b	**1**	**1**	**1**
x_p	5	2	6	y_p	0.05	0.02	0.06
η	8	4	9	θ	0.08	0.04	0.09
x_r	15	7	12	y_r	0.15	0.07	0.12
m	1.592	1.289	2.363	m	1.592	1.289	2.363
n	13.397	29.565	26.029	n	13.397	29.565	26.029
C	**Expert 1**	**Expert 2**	**Expert 3**		**Expert 1**	**Expert 2**	**Expert 3**
a	0	0	0		0	0	0
p	0.2	0.2	0.2	a	0.2	0.2	0.2
r	0.8	0.8	0.8	r	0.8	0.8	0.8
b	**1000**	**1000**	**1000**	b	**1**	**1**	**1**
x_p	5	2	6	y_p	0.005	0.002	0.006
η	8	4	9	θ	0.008	0.004	0.009
x_r	15	7	12	y_r	0.015	0.007	0.012
m	1.612	1.294	2.372	m	1.612	1.294	2.372
n	148.727	310.055	286.556	n	148.727	310.055	286.556

the individual expert elicited/fitted $GTSP$ pdfs and cdfs using equal weights in this mixture (referred to in the caption of Fig. 6.3 as the mixture distribution).

In the next section, the equi-weight mixture cdf and the elicited/fitted expert cdfs can be used directly to formulate the prior parameters of a Dirichlet process (DP), see, e.g., Ferguson (1973). Although this not being a requirement for the prior DP formulation approach in the next section, first a generalized trapezoidal (GT) distribution (see, van Dorp and Kotz 2003) is fitted to the equi-weight mixture pdf in Fig. 6.3a for parametric convenience. The pdf and cdf of the GT distribution are provided in (6.9) and (6.10). For $\eta_1 = \eta_2 = \eta$ and $\alpha = 1$, the GT family of distributions (6.9) reduces to the $GTSP$ family of distributions (6.1).

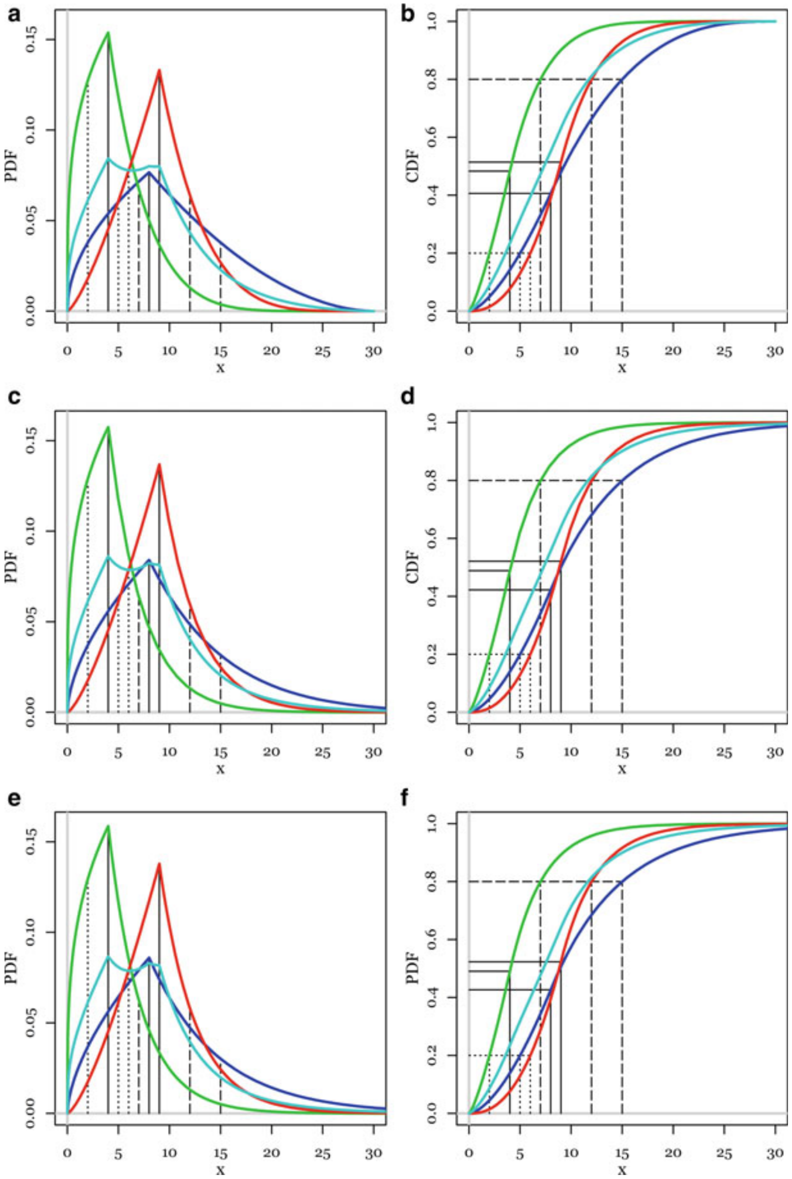


Fig. 6.3 *GTSP* pdfs and cdfs matching the expert data in Table 6.1 given different specified support ranges. Expert 1's distributions in dark blue, Expert 2's distributions in green, Expert 3's distributions in red, and equi-weighted mixture distributions in light blue. Lower quantiles in Table 6.1 indicated using dotted black lines, upper quantiles using dashed black lines and most likely values using solid black lines. **a**–**b**: *GTSP* pdfs and cdfs with support [0, 30]; **b**–**c**: *GTSP* pdfs and cdfs with support [0, 100]; **e**–**f**: *GTSP* pdfs and cdfs with support with support [0, 1000]

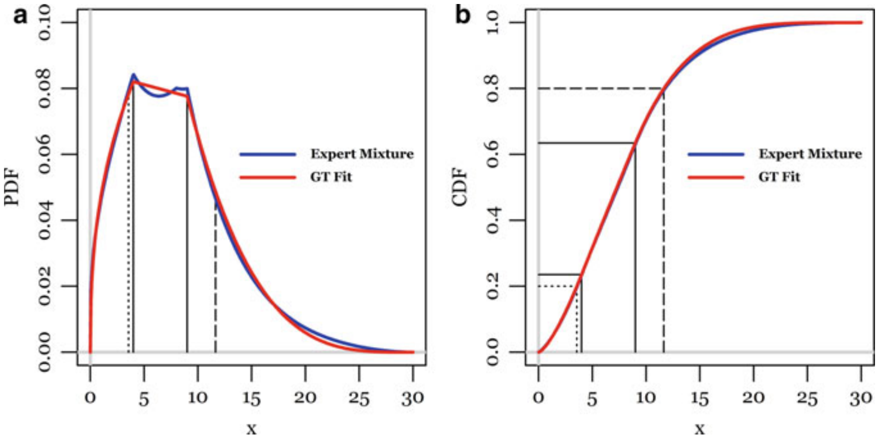


Fig. 6.4 Equi-weight mixture distribution of expert elicited $GTSP$ distributions (in blue) and Generalized Trapezoidal (GT) distribution (6.10) (in red) with parameters (6.11) fitted to this equi-weight mixture distribution—previously depicted in light blue in Fig. 6.3a–b. **a**: pdfs, **b**: cdfs

$$f(x|a,\eta_1,\eta_2,b,m,n,\alpha)=\frac{2mn}{2\alpha(\eta_1-a)n+(\alpha+1)(\eta_2-\eta_1)mn+2(b-\eta_2)m}$$
$$\times\begin{cases}\alpha\left(\frac{x-a}{\eta_1-a}\right)^{m-1}, & \text{for } a\le x<\eta_1,\\ \left\{(\alpha-1)\frac{\eta_2-x}{\eta_2-\eta_1}+1\right\}, & \text{for } \eta_1\le x<\eta_2\\ \left(\frac{d-x}{d-\eta_2}\right)^{n-1}, & \text{for } \eta_2\le x<b.\end{cases} \tag{6.9}$$

$$F(x|a,\eta_1,\eta_2,b,m,n,\alpha)=$$
$$\begin{cases}\frac{2\alpha(b-a)n_3}{2\alpha(\eta_1-a)n+(\alpha+1)(\eta_2-\eta_1)mn+2(b-\eta_2)m}\left(\frac{x-a}{\eta_1-a}\right)^{m}, & \text{for } a\le x<\eta_1,\\ \frac{2\alpha(b-a)n_3+2(x-b)n_1n_3\left\{1+\frac{(\alpha-1)}{2}\frac{(2c-b-x)}{(c-b)}\right\}}{2\alpha(\eta_1-a)n+(\alpha+1)(\eta_2-\eta_1)mn+2(b-\eta_2)m}, & \text{for } \eta_1\le x<\eta_2,\\ 1-\frac{2(d-c)n_1}{2\alpha(\eta_1-a)n+(\alpha+1)(\eta_2-\eta_1)mn+2(b-\eta_2)m}\left(\frac{d-x}{d-\eta_2}\right)^{n}, & \text{for } \eta_2\le x<b.\end{cases} \tag{6.10}$$

Figure 6.4 plots the GT pdf and cdf (in red) with parameters

$$a=0,\eta_1=4,\eta_2=9,b=30,m=1.394,n=4.466,\alpha=1.056, \tag{6.11}$$

together with the equi-weight mixture pdf and cdf of the expert elicited/fitted $GTSP$ distribution (in blue)—previously depicted in light blue in Fig. 6.3a–b.

The GT fitted parameters (6.11) were obtained by setting $\eta_1 = 4$ (the minimum of the most likely estimates of the experts in Table 6.1), by setting $\eta_2 = 9$ (the maximum of the most likely estimates of the experts in Table 6.1) and using a least squares approach between the equi-weight mixture cdf in Fig. 6.3 and the GT cdf (6.10) to obtain the remaining parameters m, n and α in (6.10). The vertical solid lines in Fig. 6.4 identify the locations of η_1 and η_2 in Fig. 6.4, whereas the vertical dashed and dotted lines identify the locations of the 20th and 80th percentiles in Fig. 6.4. From Fig. 6.4b, one visually observes a close fit between the two cdfs.

6.3.1 Constructing a Dirichlet Process as a Prior for the cdf $F(x)$

A Dirichlet process (Ferguson 1973) may be used to define a prior distribution for a cdf $F(x)$ for every time $x \in (0, \infty) = \mathbb{R}^+$. It is referred to as a random process since it is indexed by time x. Ferguson (1973) showed that for a DP with parameter measure function $\alpha(\cdot)$, where

$$0 < \alpha[(a, b)] < \alpha(\mathbb{R}^+) < \infty,\ (a, b) \subset \mathbb{R}^+,$$

the random variable $F(x)$ follows a beta distribution given by (6.2) with support [0, 1] and parameters

$$\alpha[(0, x)], \text{and}\, \alpha\{[x, \infty)\} = \alpha(\mathbb{R}^+) - \alpha[(0, x)]. \tag{6.12}$$

Thus with (6.12) and (6.2) one obtains

$$\begin{aligned} E[F(x)|\alpha(\cdot)] &= \frac{\alpha[(0, x)]}{\alpha(\mathbb{R}^+)}, \\ V[F(x)|\alpha(\cdot)] &= \frac{\alpha[(0, x)] \times \{\alpha(\mathbb{R}^+) - \alpha[(0, x)]\}}{\alpha^2(\mathbb{R}^+)[\alpha(\mathbb{R}^+) + 1]}. \end{aligned} \tag{6.13}$$

Next, the prior knowledge encapsulated in the GT fitted cdf $F(x|\Theta)$ (6.10) about the unknown component lifetime distribution $F(x)$ can be incorporated in the DP by setting in (6.12)

$$\alpha[(0, x)] = \alpha(\mathbb{R}^+) \times F(x|\Theta), \tag{6.14}$$

where $\Theta = (a, \eta_1, \eta_2, b, m, n, \alpha)$ with parameter setting (6.11).[1] This yields with (6.13)

$$\begin{aligned} E[F(x)|\alpha(\cdot)] &= F(x|\Theta), \\ V[F(x)|\alpha(\cdot)] &= \frac{F(x|\Theta) \times [1 - F(x|\Theta)]}{\alpha(\mathbb{R}^+) + 1}. \end{aligned} \tag{6.15}$$

[1] Instead of the GT Fit $F(x|\Theta)$ with parameters (6.11), the equi-weight mixture cdf could have been used directly.

Firstly, observe from (6.15) that the prior expectation of the DP equals now the GT fitted cdf $F(x|\Theta)$ (6.10). Secondly, observe from (6.15) that $\alpha(\mathbb{R}^+)$ is a positive constant that drives the variance in $F(x)$ with lesser values of $\alpha(\mathbb{R}^+)$ implying a larger variance. Denoting the expert elicited/fitted $GTSP$ cdfs to the data in Table 1 with $F_e(x|a, \eta, b, m, n)$, it is suggested to estimate/solve for $\alpha(\mathbb{R}^+)$ using the following procedure for determining the prior variance in (6.15).

Step 1: Evaluate x^* at which the sample variance of $F(x)$ over the elicited expert distributions is as large as possible, i.e., maximize:

$$\widehat{V}[F(x)] = \frac{1}{E-1}\sum_{e=1}^{E}[F_e(x|a, \eta, b, m, n) - F(x|\Theta)]^2. \tag{6.16}$$

It is important to note here that in (6.16) $F(x|\Theta)$ was obtained through the equi-weight mixture of the elicited expert distributions, and thus $\widehat{V}[F(x)]$ can indeed be interpreted as the estimate for the sample variance $F(x)$ at x, treating each value $F_e(x|a, \eta, b, m, n)$, $e = 1, \dots, E$ as data at x.

Step 2: Solve $\alpha(\mathbb{R}^+)$ from (6.15) by setting

$$V[F(x^*)|\alpha(\cdot)] = \widehat{V}[F(x^*)], \tag{6.17}$$

where $\widehat{V}[F(x)]$ follows from (6.16).

The procedure above is conservative in the sense that it (i) ensures the largest prior uncertainty in the DP and (ii) allows for the relative largest effect when updating the prior DP with failure data and maintenance data given the expert elicited fitted $GTSP$ pdfs.

Figure 6.5a plots $\widehat{V}[F(x)]$ (6.16) as a function of x on the left y-axis together with the expert elicited cdfs plotted on right y-axis and depicted in the same colors in Fig. 6.3b. Recall the expert judgment data provided in Table 6.1 served as input to obtain these expert elicited cdfs. The maximum of $\widehat{V}[F(x)]$ in Fig. 6.5a is attained at

$$x^* \approx 6.462 \text{ with } \widehat{V}[F(x^*)] \approx 0.0824 \text{ and } F(x^*|\Theta) \approx 0.435. \tag{6.18}$$

With (6.15), the value for $\alpha(\mathbb{R}^+)$ now follows from (6.18) as

$$\alpha(\mathbb{R}^+) \approx 1.983. \tag{6.19}$$

Hence, from (6.19) it follows that the elicited expert information (from three experts in this illustrative example) is roughly equivalent to the information provided by two failure data points.

While the support $[0, 30] \times [0, 1]$ of the DP naturally follows from the support $[0, 30]$ of the GT cdf $F(x|\Theta)$ (6.10) and that for any cdf $F(\cdot) \in [0, 1]$, Fig. 6.5b further depicts a summary of the resulting uncertainty in the prior DP for the cdf

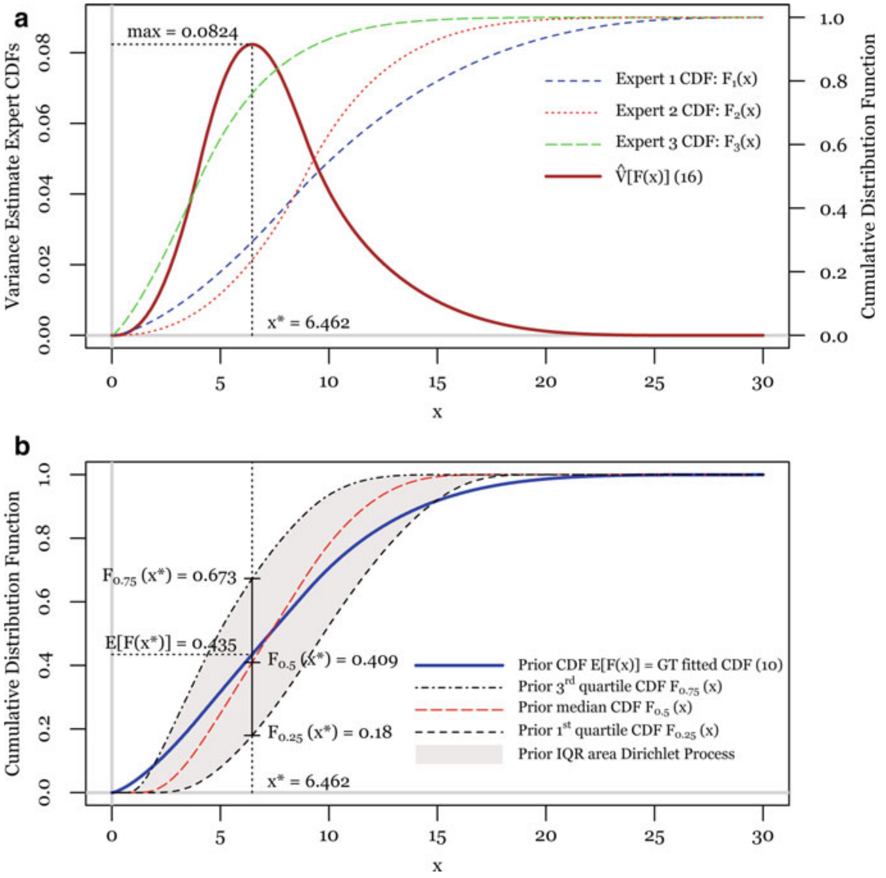


Fig. 6.5 Graphical depiction of analysis to solve for the parameters of the prior Dirichlet process for the cdf $F(x)$ from the expert judgment data in Table 6.1A and summary of resulting uncertainty in the Dirichlet process. **a**: Plot of $\widehat{V}[F(x)]$ given by (6.16) using left y-axis and ingredient expert elicited cdfs plotted on right y-axis depicted in the same colors in Fig. 6.3b; **b**: Prior Dirichlet process point estimate functions for the cdf $F(x)$. The error bar depicts their values at x^* from Fig. 5a. The gray area graphically depicts the IQRs for $F(x)$ as a function of x

$F(x)$. The blue line in Fig. 6.5b is the GT cdf $F(x|\Theta)$ (6.10) with parameters (6.11) that was set equal to $E[F(x)|\alpha(\cdot)]$ of the DP in (6.15). At $x^* = 6.462$, it follows next from (6.18) that $E[F(x^*)|\alpha(\cdot)] = F(x^*|\Theta) \approx 0.435$. The uncertainty in $F(\cdot)$ at $x^* \approx 6.462$ is depicted in Fig. 6.5b using error bars. The lower endpoint of the error bar equals the first quartile of the cdf random variable $F(x^*)$. Its value $F_{0.25}(x^*) \approx 0.180$ is obtained as the 25-th percentile of a beta distribution (6.2) with variance $\widehat{V}[F(x^*)] \approx 0.0824$ indicated in Fig. 6.5a. From (6.12), (6.14), (6.18), and (6.19). The parameters of this beta distribution equal $F(x^*|\Theta) \times \alpha(\mathbb{R}^+) \approx 0.862$ and $\{1 - F(x^*|\Theta)\} \times \alpha(\mathbb{R}^+) \approx 1.121$. In a similar manner, the upper endpoint of the error bar $F_{0.75}(x^*) \approx 0.673$ equals the third quartile of the cdf random variable $F(x^*)$, while the median of this cdf random variable $F(x^*)$ equals $F_{0.5}(x^*) \approx 0.409$.

Evaluating analogously for every point $x \in [0, 30]$ the median value for the cdf random variable $F(x)$ results in the red dashed line in Fig. 6.5b indicated by $F_{0.5}(x)$. The lower (upper) outer boundary of the gray area depicted in Fig. 6.5b equals the first quartile $F_{0.25}(x)$ (third quartile $F_{0.75}(x)$) of the cdf random variable $F(\cdot)$ at x, $x \in [0, 30]$. In other words, the gray area in Fig. 6.5b depicts the inter-quartile ranges (IQRs) for the cdf random variables $F(x)$, $x \in [0, 30]$ as a function of x defined by the DP with support $[0, 30] \times [0, 1]$ and that follow from the value $\alpha(\mathbb{R}^+) \approx 1.983$ solved for using (6.17) and by setting $E[F(x)|\alpha(\cdot)] = F(x|\Theta)$, where $F(x|\Theta)$ is the GT cdf (6.10) with parameters (6.11).

6.4 Bayesian Updating Using Failure and Maintenance Data

Let

$$(n_x, x) \equiv (x_{(1)}, \ldots x_{(n_{\underline{x}})}) \tag{6.20}$$

be a sample of ordered failure times x_j, $j = 1, \ldots, n_x$. The formal definition of a random sample from a DP is somewhat technical (see, Ferguson 1973), and it is difficult to verify that a realization of such a sample lives up to this definition. Hence, it shall be assumed that $(n_x, \underline{x})$ is a sample from a DP with parameter measure $\alpha(\cdot)$ defined by (6.14) and a value of $\alpha(\mathbb{R}^+)$ given by (6.19) determined using the procedure as described in Sect. 6.3.1. Ferguson's main theorem entails that the posterior distribution of $F(\cdot)$ given $(n_x, \underline{x})$ is again a DP, and thus conjugate, where

$$\alpha[(0, x)|(n_x, \underline{x})] = \alpha[(0, x)] + i \text{ for } x_{(i)} \leq x < x_{(i+1)}, i = 1, \ldots, n_x, \tag{6.21}$$

and $x_{(0)}$, $x_{(n+1)}$ are the boundaries of the support of $F(\cdot)$. Substitution of (6.14) in (6.21) yields with the properties of a DP that the random variable

$$[F(x)|(n_x, \underline{x})] \sim Beta(\alpha(\mathbb{R}^+) \times F(x|\Theta) + i, \alpha(\mathbb{R}^+) \times [1 - F(x|\Theta)] + n_x - i) \tag{6.22}$$

with posterior expectation

$$E[F(x)|\alpha(\cdot), (n_x, \underline{x})] = \lambda_{n_x} F(x|\Theta) + (1 - \lambda_{n_x})\widehat{F}_{n_x}[x|(n_x, \underline{x})], \tag{6.23}$$

where

$$\lambda_{n_x} = \frac{\alpha(\mathbb{R}^+)}{\alpha(\mathbb{R}^+) + n_x}, \tag{6.24}$$

$$\widehat{F}_{n_x}[x|(n_x, \underline{x})] = \frac{i}{n_x} \text{ for } x_{(i)} \leq x < x_{(i+1)}, i = 1, \ldots, n_x.$$

In other words, the posterior expectation (6.23) for the component lifetime distribution is a mixture of the prior expectation of the DP, i.e., the GT distribution $F(x|\Theta)$ (6.10) constructed in Sect. 6.3, and the empirical cumulative distribution function $\widehat{F}_{n_x}[x|(n_x, \underline{x})]$ in (6.24). Moreover, one observes from (6.24) that the mixture weight $\lambda_{n_x} \downarrow 0$ when $n_x \to \infty$. That is, as failure data accumulates a lesser weight is assigned to the prior distribution $F(x|\Theta)$. Because of the structure of λ_{n_x} in (6.24), the measure $\alpha(\mathbb{R}^+)$ may be referred to as the "prior" or "virtual" sample size for the prior information.

Given failure data

$$(n_x, \underline{x}) = (4, 10, 11, 13, 15) \Rightarrow n_x = 5, \tag{6.25}$$

Figure 6.6 compares, at $x^* = 6.462$, the prior and the posterior estimates for $F(x^*)$ (obtained under a squared error loss function) indicated in Fig. 6.6a, b with values $E[F(x^*)|\alpha(\cdot)] \approx 0.435$ and $E[F(x^*)|\alpha(\cdot), (n_x, \underline{x})] \approx 0.267$, respectively. At $x^* = 6.462$, prior and posterior median estimates for $F(x^*)$ (obtained under an absolute error loss function) are indicated at the error bars in Fig. 6.6a, b with values $F_{0.5}(x^*) \approx 0.409$ and $F_{0.5}[(x^*|(n_x, \underline{x})] \approx 0.244$, respectively. Thus, one observes a reduction of about 50% in the prior estimated values for $F(x^*)$ when updating this prior estimate with the failure data (6.25) as described above.[2]

In addition, Fig. 6.6 compares the prior cdf $F(\cdot|\Theta)$ (6.10) (in blue in Fig. 6.6a and also displayed in Fig. 6.5b) with parameters (6.11) with the empirical cdf $\widehat{F}_{n_x}[\cdot|(n_x, \underline{x})]$ (6.24) (in green in Fig. 6.6) for the failure data (6.25) and the posterior cdf $E[F(\cdot)|\alpha(\cdot), (n_x, \underline{x})]$ (6.23) (in red in Fig. 6.6b). Since $n_x = 5$ and from (6.19) $\alpha(\mathbb{R}^+) \approx 1.9832$, it follows with (6.24) that $\lambda_{n_x} \approx 0.2840$. Thus given (6.23), a larger weight is assigned to the empirical cdf, which is visually evident from Fig. 6.6 since the "distance" of the posterior cdf (6.23) to the empirical cdf (6.24) in Fig. 6.6b is smaller than the "distance" of the prior cdf $F(\cdot|\Theta)$ (6.10) to the empirical cdf (6.24) in Fig. 6.6a. Since the posterior cdf $E[F(\cdot)|\alpha(\cdot), (n_x, \underline{x})]$ (6.23) is a weighted mixture of the prior cdf $F(\cdot|\Theta)$ (6.10) and the empirical cdf $\widehat{F}_{n_x}[\cdot|(n_x, \underline{x})]$ (6.24), one observes from Fig. 6.6 that the posterior cdf (6.23) (red line in Fig. 6.6b) has the same discontinuities at the failure times $x_{(1)} = 4$, $x_{(2)} = 10$, $x_{(3)} = 11$, $x_{(4)} = 13$ and $x_{(5)} = 15$ as the empirical cdf (6.24) (green step function in Fig. 6.6).

Figure 6.6 also compares the posterior IQRs for the posterior random variables $[F(\cdot)|(n_x, \underline{x})]$ and the prior random variables $F(\cdot)$, indicated by the gray shaded areas in Fig. 6.6. In both cases, at the end points of the support of the lifetime random variable X, the expected value of $F(x)$ falls outside of these IQRs due to the skewness of the random variables $F(x)$ toward zero at the beginning of the support of X (approximately for $x \in (0, 2)$) and the skewness toward 1 of the random variables $F(x)$ toward the end of the support of X (for $x \in (15, 30)$). One observes from Fig. 6.6, the same discontinuity patterns in these IQRs at the failure times $x_{(1)} = 4$, $x_{(2)} = 10$, $x_{(3)} = 11$, $x_{(4)} = 13$ and $x_{(5)} = 15$. Overall, one observes from Fig. 6.6 a reduction in the width in the posterior IQRs for $[F(\cdot)|(n_x, \underline{x})]$ as compared to the prior IQRs for $F(\cdot)$. In fact, from the error bars in Fig. 6.6 at $x^* = 6.462$

[2] For convenience failure data $(n_x, \underline{x})$ given by (6.25) is denoted by D_f in Fig. 6.6.

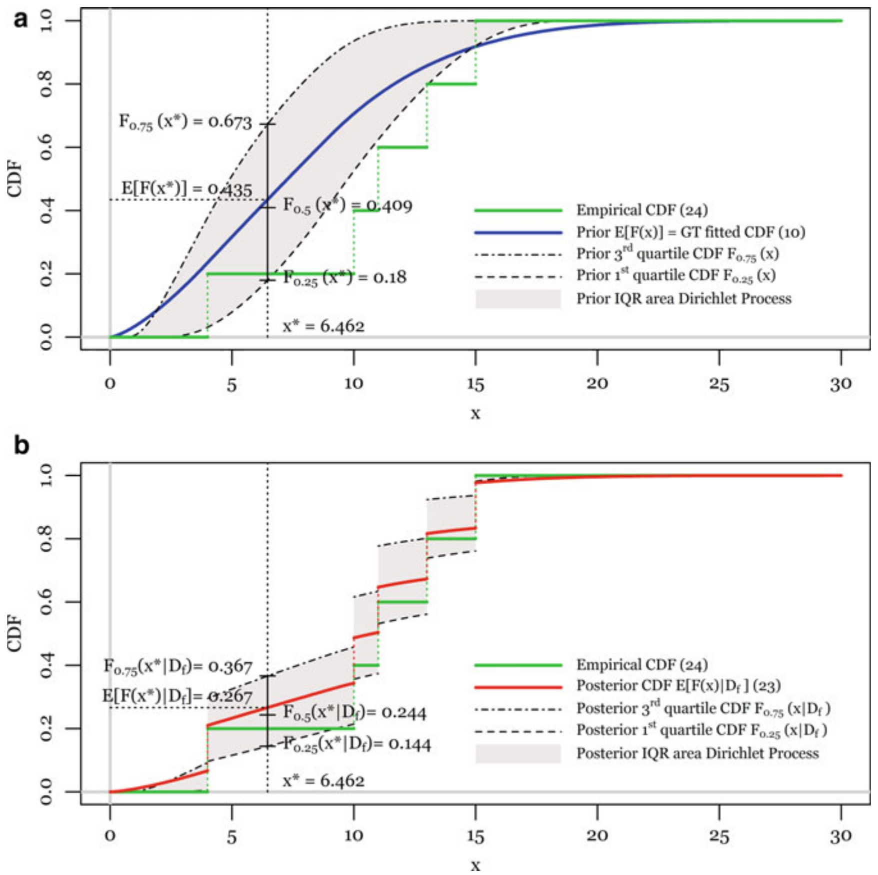


Fig. 6.6 Comparison of the empirical cdf $\widehat{F}_{n_x}[x|(n_x, \underline{x})]$ (6.24) with the prior and posterior estimates from the DP's for the cdf $F(\cdot)$ and their inter-quartile ranges. **a**: Prior Dirichlet process point estimate functions for the cdf $F(x)$. The error bar depicts their values at x^* from Fig. 6.5a. The gray area graphically depicts the IQR's for $F(x)$ as a function of x: **b**: Posterior Dirichlet process point estimate functions for the cdf $F(x|(n_x, \underline{x}))$ given failure data (6.25) denoted D_f in figure panel **b** above. The error bar depicts their values at x^* from Fig. 6.5a. The gray area graphically depicts the IQRs for $F(x|(n_x, \underline{x}))$ as a function of x

(where prior uncertainty as measured by $\widehat{V}[F(x^*)]$ in (6.18) was set at its largest value) posterior width of the IQR of $[F(x^*)|(n_x, \underline{x})]$ is evaluated in Fig. 6.6b as

$$F_{0.75}[(x^*|(n_x, \underline{x})] - F_{0.25}[(x^*|(n_x, \underline{x})] \approx 0.367 - 0.144 \approx 0.233,$$

whereas the prior width of the IQR of $F(x^*)$ is evaluated in Fig. 6.6a at

$$F_{0.75}(x^*) - F_{0.25}(x^*) \approx 0.673 - 0.18 \approx 0.493, \tag{6.26}$$

which amounts to a $(0.493 - 0.233)/0.493 \approx 54.8\%$ width reduction from the prior IQR width to the posterior IQR width at $x^* = 6.462$. Needless to say, other percent reductions in IQRs will be observed at different values of $x \in [0, 30]$.

6.4.1 Updating a Dirichlet Process with Failure and Maintenance Data

While Ferguson (1973) derived the posterior estimate (6.23) for the component lifetime distribution $F(x)$, a practical and efficient setting of preventive maintenance is not likely to result in an abundance of failure data. Instead, ordered censor time observations

$$[n_c, (\gamma, c)] \equiv [(\gamma_1, \underline{c}_{(1)}), \ldots (\gamma_{n_c}, c_{(n_c)})] \tag{6.27}$$

are more common, where $(\gamma_j, c_{(j)})$ indicates that a component was removed from service γ_j times at distinct censor time $c_{(j)}$ to be replaced or preventively maintained, $j = 1, \ldots, n_c$. Thus, γ_j can take on the values 1, 2, 3, etc. Joining the failure time data $(n_x, \underline{x})$ (6.20) with maintenance data $[n_c, (\underline{\gamma}, \underline{c})]$ (6.27) using the following notation:

$$[m_{z,}n_z, (\underline{\delta}, \underline{z})] = [(\delta_1, z_{(1)}), \ldots, (\delta_{m_z}, z_{(m_z)})]\,, m_z = n_x + n_c,\ n_z = n_x + \sum_{j=1}^{n_c} \gamma_i, \tag{6.28}$$

where

$$\delta_j = \begin{cases} 1, & z_{(j)} \in (x_{(1)}, \ldots x_{(n_x)}), \\ \gamma_j, & (\gamma_j, z_{(j)}) \in [(\gamma_1, c_{(1)}), \ldots, (\gamma_{n_c}, c_{(n_c)})], \end{cases}$$

and defining

$$n^+(x) = \sum_{\{i:z_{(i)}>x\}} \delta_i, \text{ and } n(x) = \sum_{\{i:z_{(i)}\geq x\}} \delta_i, \tag{6.29}$$

one concludes from (6.29) that $n^+(x)$ equals the number of z_i observations (i.e., censored or non-censored) strictly greater than x and $n(x)$ equals the number of z_i observations greater or equal to x. The following expression

$$\widehat{S}_{n_z}\{x|[m_z, n_z, (\underline{\delta}, \underline{z})]\} = \frac{n^+(x)}{n_z} \tag{6.30}$$

can next be interpreted as a lower bound for the empirical survival function since in the evaluation of (6.30) the censoring times $c_{(j)}$ and number of censor replacements γ_j are essentially interpreted as failure counts.

Susarla and Van Ryzin (1976) derived the following expression for the posterior moments of the survival function $S(x) = 1 - F(x)$ for $c_{(k)} \leq x < c_{(k+1)}$,

$k = 0, \dots, n_c$:

$$E[S^p(x)|\Psi] = \prod_{s=0}^{p-1}\left[\frac{\alpha\{(x,\infty)\} + s + n^+(x)}{\alpha(\mathbb{R}^+) + s + n_z}\right] \times \xi\{x, s|\alpha(\cdot), [n_c, (\underline{\gamma}, \underline{c})]\}, \quad (6.31)$$

$p = 1, 2, \dots,$ where $\Psi = \{\alpha(\cdot), [m_z, n_z, (\underline{\delta}, \underline{z})]\}$, by convention $c_{(0)} \equiv 0$, $c_{(n_c+1)} \equiv \infty$, and

$$\xi\{x, s|\alpha(\cdot), [n_c, (\underline{\gamma}, \underline{c})]\} = \prod_{j=1}^{k} \frac{\alpha(\mathbb{R}^+) \times S(c_{(j)}|\Theta) + s + n(c_{(j)})}{\alpha(\mathbb{R}^+) \times S(c_{(j)}|\Theta) + s + n(c_{(j)}) - \gamma_j}. \quad (6.32)$$

In deriving the above results, Susarla and Van Ryzin (1976) assumed that the joined failure and maintenance data $[m_z, n_z, (\underline{\delta}, \underline{z})]$ (6.28) is obtained from random observations $Z_i = \min(X_i, C_i)$, where the X_i random failure times are $i.i.d$, and the C_i's are random censoring times also independent from the X_i's. The C_i random variables are assumed to be mutually independent but do not have to be identically distributed and could be degenerate implying fixed maintenance times which is accommodated by the number of censored observations γ_j at the censor time $c_{(j)}$, $j = 1, \dots, n_c$ in the definition of the maintenance data (6.27). Substitution of $p = 1$ and $\alpha(\cdot)$ (6.14) into (6.31) yields for $c_{(k)} \le x < c_{(k+1)}, k = 0, \dots, n_c$, $c_{(0)} \equiv 0$, $c_{(n_c+1)} \equiv \infty$, the following alternative expression for $E[S(x)|\Psi]$ utilizing the $\widehat{S}_{n_z}\{x|[m_z, n_z, (\delta, z)]\}$ lower bound definition (6.30) for the empirical survival function:

$$\begin{aligned} E[S(x)|\Psi] =& \xi\{x, 0|\alpha(\cdot), [n_c, (\underline{\gamma}, \underline{c})]\} \times \quad (6.33)\\ & \left\{ \lambda_{n_z} S(x|\Theta) + (1 - \lambda_{n_z})\widehat{S}_{n_z}\{x|[m_z, n_z, (\underline{\delta}, \underline{z})]\} \right\}. \end{aligned}$$

where

$$\lambda_{n_z} = \frac{\alpha(\mathbb{R}^+)}{\alpha(\mathbb{R}^+) + n_z}, \quad S(x|\Theta) = 1 - F(x|\Theta), \quad (6.34)$$

and $F(x|\Theta)$ is the prior GT cdf (6.10). Comparing λ_{n_z} in (6.34) with λ_{n_x} in (6.24), it follows with $n_z = n_x + n_c$ that (6.34) assigns a lesser weight to the prior survival function $S(x|\Theta) = 1 - F(x|\Theta)$ with the inclusion of right-censored maintenance data $[n_c, (\underline{\gamma}, \underline{c})]$ in addition to the failure data $(n_x, \underline{x})$ in the posterior estimate evaluation of the DP. Moreover, from the structure of (6.32)–(6.34), one observes that in the case of no right-censoring, expression (6.33) is equivalent to the expression (6.23) derived by Ferguson (1973). Indeed, in case of no right-censoring $n_z = n_x$, the lower bound for the empirical survival function $\widehat{S}_{n_z}\{x|[m_z, n_z, (\underline{\delta}, \underline{z})]\}$ given by (6.30) reduces to the empirical survival function given only failure data $(n_x, \underline{x})$, and the running product term $\xi\{x, 0|\alpha(\cdot), [n_c, (\underline{\gamma}, \underline{c})]\}$ (6.32) reduces to the value 1 since $k \equiv 0$ in the no right-censoring case.

Given in addition to the failure data $(n_x, \underline{x})$ (6.25), the preventive maintenance data[3]

$$[n_c, (\underline{\gamma}, \underline{c})] \equiv \{(4, 3), (3, 6), (2, 9), (1, 12)\} \Rightarrow n_c = 4 \text{ and } \sum_{j=1}^{n_c} \gamma_i = 10 \,. \quad (6.35)$$

Figure 6.7 compares, at $x^* = 6.462$, posterior estimates for $F(x^*)$ (under a squared error loss function) with values $E[F(x^*)|\alpha(\cdot), (n_x, \underline{x})] \approx 0.267$ and $E[F(x^*)|\Psi] \approx 0.141$, respectively, where $\Psi = \{\alpha(\cdot), [m_z, n_z, (\underline{\delta}, \underline{z})]\}$. Posterior median estimates for $F(x^*)$ (obtained under an absolute loss function) are indicated in Fig. 6.7a, b at the error bars with values $F_{0.5}(x^*||\alpha(\cdot), (n_x, \underline{x})) \approx 0.244$ and $F_{0.5}[(x^*|\Psi] \approx 0.122$, respectively. Thus, one observes a further reduction of about 50% in the posterior estimated values for $F(x^*)$ by updating the DP using the maintenance data $[n_c, (\underline{\gamma}, \underline{c})]$ (6.35) in addition to the failure data $(n_x, \underline{x})$ (6.25).

In addition, Fig. 6.7 depicts the posterior cdf (6.21) (red line in Fig. 6.7a) and the posterior cdf $E[F(\cdot)|\Psi] = 1 - E[S(\cdot)|\Psi]$ (light blue line in Fig. 6.7b), where $E[S(\cdot)|\Psi]$ is given by (6.33). We have from (6.25), (6.35), and (6.28) that $n_z = 15$, from (6.19) $\alpha(\mathbb{R}^+) \approx 1.9832$, and from (6.34) now that $\lambda_{n_z} \approx 0.1167$. One observes (indirectly) from Fig. 6.7 that while the structure of $E[S(\cdot)|\Psi]$ (6.33) changes at every failure time and maintenance time observation $z_{(i)}$, $i = 1, \ldots, m_z$ (see (6.32) and (6.33)), only discontinuities are observed in the posterior cdf $E[F(\cdot)|\Psi] = 1 - E[S(\cdot)|\Psi]$ (6.33) at the failure observations $x_{(1)} = 4$, $x_{(2)} = 10$, $x_{(3)} = 11$, $x_{(4)} = 13$ and $x_{(5)} = 15$, similar to the posterior cdf (6.23). No discontinuities are observed at the censor times $c_{(1)} = 3$, $c_{(2)} = 6$, $c_{(3)} = 9$, $c_{(4)} = 12$ (indicated by the vertical red dashed lines in Fig. 6.7b) since the change in the second product term of (6.33) via $\widehat{S}_{n_z}\{c_{(k)}|[m_z, n_z, (\underline{\delta}, \underline{z})]\}$ is absorbed by a reciprocal change in the running product term $\xi\{x, 0|\alpha(\cdot), [n_c, (\underline{\gamma}, \underline{c})]\}$ defined by (6.32). Since this running product term $\xi\{x, 0|\alpha(\cdot), [n_c, (\underline{\gamma}, c)]\}$ is constant over the intervals $c_{(k)} \leq x < c_{(k+1)}$, $k = 0, \ldots, n_c$, a discontinuity is observed in (6.33) only when a failure data point falls within these intervals through the change in the value of the lower bound for the empirical survival function (6.30) as is the case for $[c_{(1)}, c_{(2)}) = [3, 6)$, $[c_{(3)}, c_{(4)}) = [6, 9)$ and $[c_{(4)}, \infty) = [12, \infty)$ in Fig. 6.7b.

Figure 6.7 also compares posterior IQRs, indicated by the gray shaded areas in Fig. 6.7, for the random variables $[F(\cdot)|(n_x, \underline{x})]$ and $\{F(\cdot)|[m_z, n_z, (\underline{\delta}, \underline{z})]\}$. The latter IQRs were obtained by fitting beta distributions to the first two posterior moments given by (6.31), while verifying that the third and fourth moments of these beta fitted distributions equal the third and fourth posterior moments, obtained using (6.31), upto three decimal places. Observe from Fig. 6.7 that when $x \in (15, 30)$, the expected value of $F(x)$ predominantly falls outside of these IQRs due to the skewness of the random variables $F(x)$ toward 1. From the error bars at $x^* = 6.462$ in Fig. 6.7a, b, one continues to observe a reduction in width of the posterior IQRs for $[F(x^*)|(n_x, \underline{x})]$ and $\{F(x^*)|[m_z, n_z, (\underline{\delta}, \underline{z})]\}$. In fact, at $x^* = 6.462$ (where prior

[3] For convenience failure data and maintenance data $[m_z, n_z, (\underline{\delta}, \underline{z})]$ given by (6.28) is denoted by $D_{f\&m}$ in Fig. 6.7.

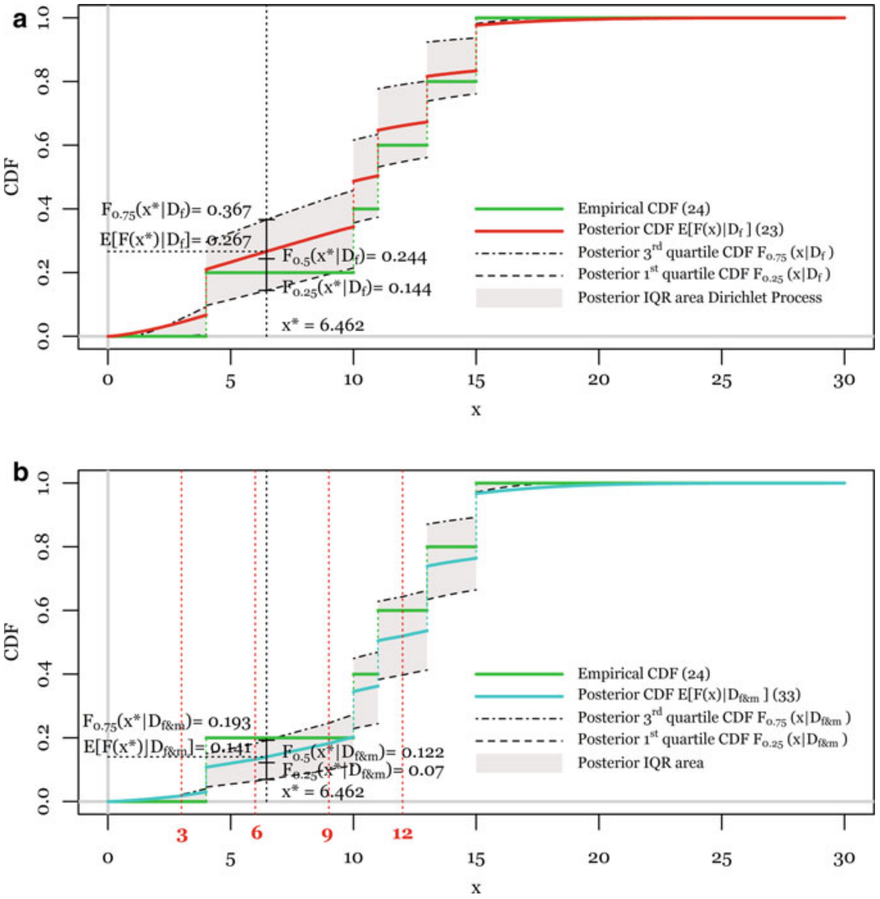


Fig. 6.7 Comparison of the empirical cdf $\widehat{F}_{n_x}[x|(n_x, \underline{x})]$ (6.24) with posterior estimates from the DP's for the cdf $F(\cdot)$ and their inter-quartile ranges. **a**: Posterior Dirichlet process point estimate functions for the cdf $F(x|(n_x, \underline{x}))$ given failure data (6.25) denoted D_f in figure panel a above. The error bar depict their values at x^* from Fig. 6.5a. The gray area graphically depicts the IQRs for $F(x|(n_x, \underline{x}))$ as a function of x; **b**: Posterior point estimate functions for the cdf $F(x|[m_z, n_z, (\underline{\delta}, \underline{z})])$ given failure data (6.25) and maintenance data (6.35) denoted $D_{f\&m}$ in figure panel B above. The error bar depicts their values at x^* from Fig. 6.5a. The gray area graphically depicts the IQRs for $F(x|[m_z, n_z, (\underline{\delta}, \underline{z})])$ as a function of x

uncertainty as measured by $\widehat{V}[F(x^*)]$ in (6.18) was set at its largest value), posterior width of the IQR for $[F(x^*)|(n_x, \underline{x})]$ is evaluated in Fig. 6.7a as

$$F_{0.75}[(x^*|(n_x, \underline{x})] - F_{0.25}[(x^*|(n_x, \underline{x})] \approx 0.367 - 0.144 \approx 0.233,$$

whereas the posterior width of the IQR for $\{F(x^*)|[m_z, n_z, (\delta, \underline{z})]\}$ is evaluated in Fig. 6.7b at

$$F_{0.75}(x^*|[m_z, n_z, (\underline{\delta}, \underline{z})]) - F_{0.25}(x^*|[m_z, n_z, (\underline{\delta}, \underline{z})]) \approx 0.193 - 0.069 \approx 0.124,$$

which amounts to a further width reduction of $(0.233 - 0.124)/0.223 \approx 44.4\%$. Overall from the prior IQR width (6.26), a reduction of $(0.493 - 0.124)/0.493 \approx 74.8\%$ is achieved by updating the prior cdf $F(x^*|\Theta)$ (6.10) with both failure data (6.25) and maintenance data (6.35).

6.5 Maintenance Optimization

A basic model within the context of maintenance optimization is the block replacement model. For an extensive discussion of this model, see Mazzuchi and Soyer (1996). In the block replacement model, a single maintenance activity is carried out at a pre-specified age x of the component. The long-term average costs (LTAC) per unit time follows from the expected costs during the preventive maintenance cycle with length x. Of course, during such a maintenance cycle a component can fail multiple times. The cost of a failure K_f during the maintenance cycles is assessed higher than the preventive maintenance cost K_p as the failure cost K_f is unplanned and may result in additional disruptions, thus $K_f > K_p$. In fact, when $K_f \leq K_p$ it is not worthwhile to preventively maintain the component. Denoting $\Lambda(x)$ to be the expected number of failures during the maintenance cycle with length x, one obtains for the LTAC per unit time $g(x)$ given $\Lambda(x)$

$$[g(x)|\Lambda(x)] = \frac{K_p + K_f \times \Lambda(x)}{x}. \tag{6.36}$$

Under a minimal repair assumption the failure process $N(x) \equiv \#$ Failures in the interval from $[0, x]$, can be described as non-homogenous Poisson process with mean value function $\Lambda(x)$. One obtains in that case:

$$Pr(N(x) = k) = \frac{\{\Lambda(x)\}^k}{k!} exp\{-\Lambda(x)\}, \text{ where } \Lambda(x) = \int_0^x \lambda(u)du,$$

and $\lambda(u)$ is the intensity function. Therefore, $E[N(x)] = \Lambda(x)$ and one obtains for the probability of zero failures in the interval $[0, x]$:

$$Pr(N(x) = 0) = exp[-\Lambda(x)].$$

Denoting $X_1 \sim F(\cdot)$ to be the failure time occurrence of the first failure with time-to-failure cdf $F(\cdot)$, one obtains the following equivalency:

$$Pr(N(x) = 0) = exp[-\Lambda(x)] \equiv Pr(X_1 > x) = 1 - F(x). \tag{6.37}$$

Utilizing (6.37) yields with (6.36) the following expression for the LTAC per unit time

$$[g(x)|F(x)] = \frac{K_p - K_f \times \ln\{1 - F(x)\}}{x}. \tag{6.38}$$

The optimal maintenance interval (given the cdf $F(\cdot)$) in the block replacement model (6.36) may be defined to be that time point $x^\bullet$ for which $[g(x^\bullet)|F(x^\bullet)]$ given by (6.38) is minimal.

Given that in a DP the cdf $F(x)$ is a random variable for every fixed value of x, the LTAC per unit time $[g(x)|F(x)]$ is, via the transformation (6.38), too a random variable. Figure 6.8 provides a comparison of the LTAC per unit time $[g(x)|F(x)]$ with $K_f = 20$ and $K_p = 2$, i.e., a failure is 10 times more costly than a preventive maintenance action, while setting $F(x)$ equal to (i) the prior GT cdf $F(x|\Theta)$ (6.10) (in blue in Fig. 6.8a), (ii) the posterior cdf $E[F(t)|\alpha(\cdot), (n_x, \underline{x})\,]$ (6.23) (in red in Fig. 6.8b), and (iii) the posterior cdf $E[F(x)|\Psi] = 1 - E[S(x)|\Psi]$ (6.33) (in green in Fig. 6.8c), where $\Psi = \{\alpha(\cdot), [m_z, n_z, (\underline{\delta}, \underline{z})]\}$. Furthermore, Fig. 6.8 plots the prior and posterior medians (in dark red dashed lines) of the random variables $[g(x)|F(x)]$ utilizing (6.38) and their IQRs (gray shaded areas).[4] One observes a more pronounced departure from the IQRs for $x \in (15, 30)$ for the blue, red and green lines in Fig. 6.8 than observed in Figs. 6.6 and 6.7 for this range. This follows from the earlier observations in Figs. 6.6 and 6.7 that (i) the expected value of $F(x)$ is skewed toward the value 1 for $x \in (15, 30)$, (ii) the expected value of $F(x)$ falls outside of the IQRs for $x \in (15, 30)$, and finally (iii) the steepness of the gradient of the function $\ln\{1 - F(x)\}$ in (6.38) when $F(x) \uparrow 1$.

The optimal maintenance interval that follows in Fig. 6.8a using the prior GT cdf $F(x|\Theta)$ (6.10) equals $x^\bullet \approx 3.18$ with an LTAC per unit time $g(x^\bullet) \approx 1.81$. Utilizing the prior median estimate for $F(x)$ in Fig. 6.8a, the optimal maintenance interval changes to $x^\bullet \approx 2.52$ with an LTAC per unit time $g(x^\bullet) \approx 1.03$. Finally, using the boundaries of the prior IQRs in Fig. 6.8a, a range of $(1.44, 3.78)$ is evaluated for the optimal maintenance interval $x^\bullet$. In other words, one observes uncertainty in both the optimal value for the LTAC and the length of the optimal maintenance interval in Fig. 6.8a.

When updating the DP for $F(x)$ with only the failure data (6.25) and utilizing the posterior cdf $E[F(t)|\alpha(\cdot), (n_x, \underline{x})\,]$ (6.23) for $F(x)$, the optimal maintenance interval increases to $x^\bullet = 4$ in Fig. 6.8b with an LTAC per unit time per unit time $g(x^\bullet) \approx 0.85$. An LTAC per unit time $g(x^\bullet) \approx 0.66$ is evaluated in Fig. 6.8b when utilizing the posterior median cdf estimate for $F(x)$ in (6.38) with an optimal maintenance interval also evaluated at $x^\bullet = 4$. In fact, using the boundaries of the posterior IQRs in Fig. 6.8b too results in optimal maintenance intervals evaluated at $x^\bullet = 4$. Summarizing, comparing Fig. 6.8a, b, one observes a large reduction in the uncertainty of the optimal value for the LTAC per unit time, but more importantly no remaining uncertainty is observed in the evaluation of the optimal maintenance intervals in Fig. 6.8b at $x^\bullet = 4$.

[4] For convenience failure data $(n_x, \underline{x})$ given by (6.25) is denoted by D_f in Fig. 6.8, and failure data and maintenance data $[m_z, n_z, (\underline{\delta}, \underline{z})]$ given by (6.28) is denoted by $D_{f\&m}$ in Fig. 6.8.

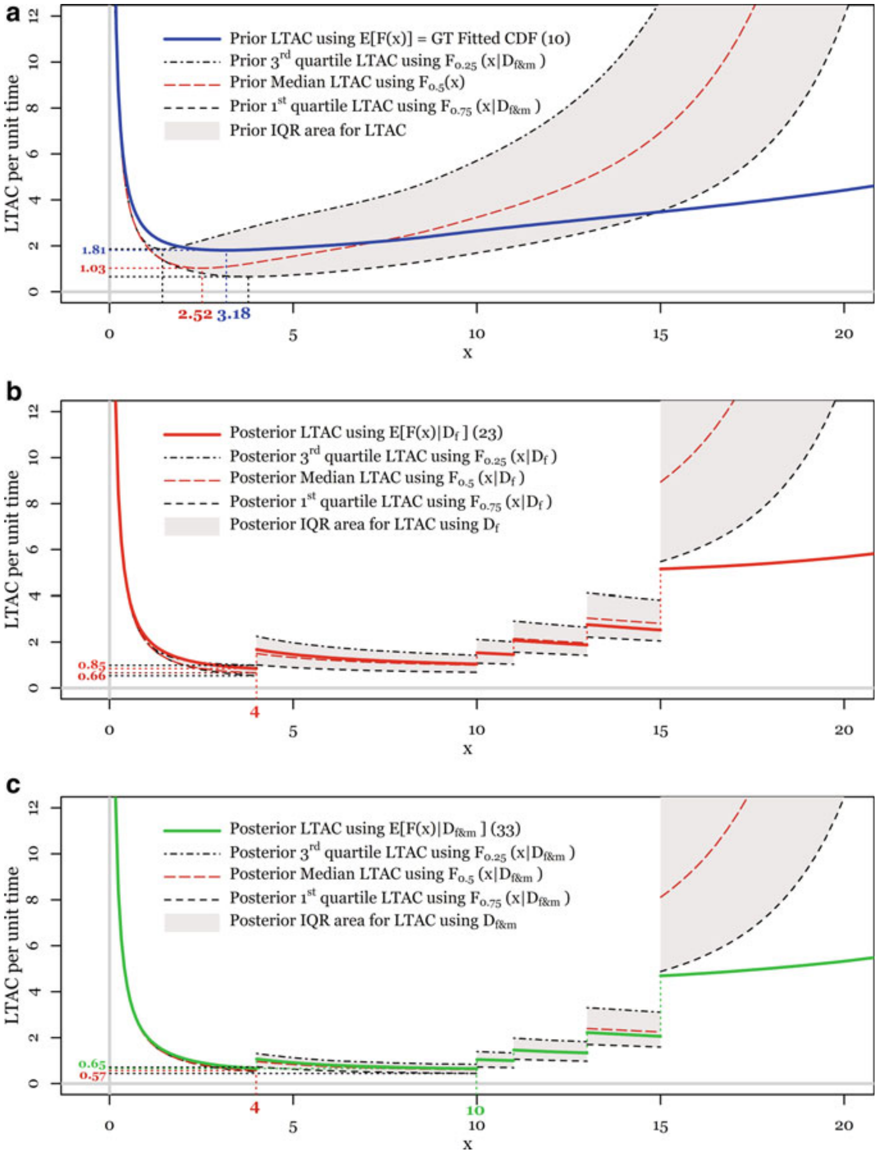


Fig. 6.8 Comparison of prior and posterior estimates for the Long-Term Average Cost (LTAC) per unit time $[g(x)|F(x)]$ (6.38) with $K_f = 20$ and $K_p = 2$ and their inter-quartile ranges. **a**: Prior LTAC estimate functions and LTAC IQR area; **b**: Posterior LTAC estimate functions and LTAC IQR area given failure data (6.25) denoted D_f in figure panel b above; **c**: Posterior LTAC estimate functions and LTAC IQR area given failure data (6.25) and maintenance data (6.35) denoted $D_{f\&m}$ in figure panel B above

When updating the DP for $F(x)$ with both failure data (6.23) and the right-censored data (6.31) and substituting the posterior cdf $E[F(x)|\Psi] = 1 - E[S(x)|\Psi]$ (6.33) in (6.38), the optimal maintenance interval further extends to $x^{\bullet} = 10$ in Fig. 6.8c with an LTAC per unit time $g(x^{\bullet}) \approx 0.65$. On the other hand, an LTAC per unit time $g(x^{\bullet}) \approx 0.57$ is evaluated in Fig. 6.8c when utilizing the posterior median cdf estimate for $F(x)$ in (6.38) with an optimal maintenance interval evaluated this time at $x^{\bullet} = 4$. In fact, using the boundaries of the posterior IQRs in Fig. 6.8c too results in optimal maintenance intervals evaluated at $x^{\bullet} = 4$. Summarizing, comparing Fig. 6.8b, c, one observes a further reduction in the uncertainty of the optimal value for the LTAC per unit time combined with a potential lengthening of the optimal maintenance interval in Fig. 6.8c from $x^{\bullet} = 4$ toward $x^{\bullet} = 10$.

In more general terms, one observes from Fig. 6.8 that the block replacement model (6.38) is most punitive when the frequency of preventive maintenance is higher than the optimal frequency (i.e., utilizing smaller than optimal preventative maintenance intervals) as compared to when the frequency of preventive maintenance is less. This follows from the steep decline of the block replacement cost curves in Fig. 6.8a–c up to optimality and a more slowly increasing cost curve thereafter. Thus, a reasonable recommendation that follows from the posterior analysis in Fig. 6.8 is that one ought not to preventively maintain using maintenance interval lengths less than $x^{\bullet} = 4$.

6.6 Conclusion

An easily implementable, fully Bayesian analysis for maintenance optimization has been presented, taking into account the real-life situation of the need for expert judgment and the heavy reliance on censored data given scarcely available failure data in a preventive maintenance context. The approach is readily extendable to other methods for determining expert weights such as those specified in Cooke (1991) as well as other distribution models fitted to the three-point expert elicited information. An R-implementation of the analysis procedure presented herein is available from the authors upon request.

Appendix

It will be proven in this appendix that under the condition that

$$0 < y_p < \theta < y_r < 1 \text{ and } 0 < p < r < 1, \tag{6.39}$$

a unique pair of power parameters (m^*, n^*) exist for the $GTSP$ pdf (6.3) such that the quantile constraint set (6.6) is met. Before this theorem can be proven, five lemmas to support the proof of this theorem have to be proven first.

Lemma 6.1 *Let $Y \sim GTSP(\theta, m, n)$ with pdf (6.3) and cdf (6.4). Under condition (6.39), the quantile constraint set (6.6) defines a continuous implicit function $\xi(n) > 0$ such that for all $n > 0$ the triplet $\{\theta, m^\bullet = \xi(n),\ n\}$ satisfies the first constraint in (6.6), i.e.,*

$$F(y_p|\theta, m^\bullet, n) = \pi(\theta, m^\bullet, n)\left(\frac{y_p}{\theta}\right)^{m^\bullet} = p. \tag{6.40}$$

In addition, under condition (6.39), the quantile constraint set (6.6) defines a continuous implicit function $\zeta(m) > 0$ such that for all $m > 0$ the triplet $\{\theta, m,\ n^\bullet = \zeta(m)\}$ satisfies the second constraint in (6.6), i.e.,

$$F(y_r|\theta, m, n^\bullet) = 1 - [1 - \pi(\theta, m, n^\bullet)]\left(\frac{1 - y_r}{1 - \theta}\right)^{n^\bullet} = r. \tag{6.41}$$

Proof From condition (6.39) and $\pi(\theta, m, n) = Pr(Y \le \theta)$, it immediately follows that:

$$\forall n, m > 0 : \pi(\theta, m, n) > p. \tag{6.42}$$

With (6.41) and (6.5), i.e.,

$$\pi(\theta, m, n) = \frac{\theta n}{(1 - \theta)m + \theta n},$$

one obtains the following upper bound for the power parameter m, given a fixed value for $n > 0$ and a specified quantile $y_p < \theta$ with quantile level p:

$$0 < m < M(n|p, \theta) = n \times \frac{\theta}{1 - \theta} \times \frac{1 - p}{p}. \tag{6.43}$$

With condition (6.39), one obtains $0 < (y_p/\theta) < 1$ and it next follows from (6.40) that

$$\forall n > 0 : F(y_p|\theta, M(n|p, \theta), n) < p. \tag{6.44}$$

Moreover, from (6.40) and $\pi(\theta, m, n)$ (6.5), it follows that when $m \downarrow 0$, while keeping $n > 0$ fixed:

$$F(y_p|\theta, m, n) \to 1. \tag{6.45}$$

Hence, from (6.44) and (6.45), it now follows with the continuity of $F(\cdot|\theta, m, n)$ and $F(\cdot|\theta, m, n)$ being a strictly increasing function that the first quantile constraint in (6.6) has a unique solution $0 < m^\bullet < M(n|p, \theta)$ for every fixed value of $n > 0$, and thus the first quantile constraint (6.6) defines a unique continuous implicit continuous function $\xi(n)$ such that the parameter combination $\{\theta, m^\bullet = \xi(n),\ n\}$ satisfies the first quantile constraint in (6.6) for all $n > 0$. Analogously, from $\theta < y_r$ and $\pi(\theta, m, n) < r$, it follows that

$$0 < n < N(m|r, \theta) = m \times \frac{1-\theta}{\theta} \times \frac{r}{1-r}, \tag{6.46}$$

and the second quantile constraint in (6.6) defines a unique continuous implicit function $\zeta(m)$ such that the parameter combination $(\theta, m, n^{\bullet} = \zeta(m))$ satisfies the second quantile constraint (6.6) for all $m > 0$. □

Lemma 6.2 *Let $Y \sim GTSP(\theta, m, n)$ with pdf (6.3) and cdf (6.4). Under condition (6.39), one obtains with Lemma 6.1 the following relationship for the derivative of the implicit function $\xi(n)$ defined in Lemma 6.1:*

$$\frac{d\xi(n)}{dn} = \frac{\theta}{1-\theta} \times \frac{1 - \pi(\theta, \xi(n), n)}{\pi(\theta, \xi(n), n) + n\frac{\theta}{1-\theta}\ln[\left(\frac{\theta}{y_p}\right)]}. \tag{6.47}$$

Furthermore, $\frac{d\xi(n)}{dn} > 0$ and the upper bound function $M(n|p, \theta)$ (6.43) is a tangent of the implicit function $\xi(n)$ as $n \downarrow 0$.

Proof From Lemma 6.1 and (6.40), one obtains

$$\pi(\theta, \xi(n), n)\left(\frac{y_p}{\theta}\right)^{\xi(n)} = p. \tag{6.48}$$

Taking natural logarithms on both sides of (6.48) yields with the definition of $\pi(\theta, m, n)$ given by (6.5) that

$$\ln[\theta n] - \ln[(1-\theta)\xi(n) + \theta n] + \xi(n)\ln[\left(\frac{y_p}{\theta}\right)] = \ln(p). \tag{6.49}$$

Taking the derivative with respect to n on both sides of (6.49) yields

$$\frac{\theta}{\theta n} - \frac{(1-\theta)\frac{d\xi(n)}{dn} + \theta}{(1-\theta)\xi(n) + \theta n} + \ln[\left(\frac{y_p}{\theta}\right)]\frac{d\xi(n)}{dn} = 0. \tag{6.50}$$

After some algebraic manipulations while using repeatedly that

$$\pi(\theta, \xi(n), n) = Pr(Y \le \theta | \theta, \xi(n), n) = \frac{\theta n}{(1-\theta)\xi(n) + \theta n} > p, \tag{6.51}$$

one obtains from (6.50) and (6.51) the relationship (6.47) for $\frac{d\xi(n)}{dn}$. With condition (6.39) and $n > 0$, it now follows from (6.47) that $\frac{d\xi(n)}{dn} > 0$. Furthermore, from Lemma 6.1 and (6.43) it follows that

$$0 < \xi(n) < M(n|p, \theta) = n \times \frac{\theta}{1-\theta} \times \frac{1-p}{p},$$

and thus

$$\xi(n) \downarrow 0 \text{ as } n \downarrow 0. \tag{6.52}$$

From Lemma 6.1 and given that the triplet $\{\theta, m^{\bullet} = \xi(n),\ n\}$ satisfies the constraint (6.40) for all $n > 0$, it next follows from (6.52) with the definition of the pdf (6.3) that the $GTSP(\theta, \xi(n),\ n)$ distribution converges to a $Bernoulli$ distribution with a point mass p at $y = 0$. As a result, it follows with (6.51) that

$$\pi(\theta, \xi(n), n) \downarrow p \text{ as } n \downarrow 0.$$

Subsequently with (6.47), it follows that $M(n|p, \theta)$ is a tangent line of the implicit function $\xi(n)$ at $n = 0$. □

Lemma 6.3 *Let $Y \sim GTSP(\theta, m, n)$ with pdf (6.3) and cdf (6.4). Under condition (6.39), one obtains with Lemma 6.1 the following relationship for the derivative of the implicit function $\zeta(m)$ defined in Lemma 6.1:*

$$\frac{d\zeta(m)}{dm} = \frac{1-\theta}{\theta} \times \frac{\pi[\theta, m, \zeta(m)]}{1 - \pi[\theta, m, \zeta(m)] + m\frac{1-\theta}{\theta}\ln[\frac{1-\theta}{1-y_r}]}. \tag{6.53}$$

Furthermore, $\frac{d\zeta(m)}{dm} > 0$ and the upper bound function $N(m|r, \theta)$ (6.46) is a tangent of the implicit function $\zeta(m)$ as $m \downarrow 0$.

Proof From Lemma 6.1 and (6.41), one obtains

$$\{1 - \pi[\theta, m, \zeta(m)]\}\Big(\frac{1-y_r}{1-\theta}\Big)^{\zeta(m)} = 1 - r\,. \tag{6.54}$$

Taking natural logarithms on both sides of (6.54) yields with the definition of $\pi(\theta, m, n)$ given by (6.5) that

$$\ln[(1-\theta)m] - \ln[(1-\theta)m + \theta\zeta(m)] + \zeta(m)\ln[\Big(\frac{1-y_r}{1-\theta}\Big)] = \ln(1-r). \tag{6.55}$$

Taking the derivative with respect to m on both sides of (6.55) yields

$$\frac{(1-\theta)}{(1-\theta)m} - \frac{(1-\theta) + \theta\frac{d\zeta(m)}{dm}}{(1-\theta)m + \theta\zeta(m)} + \ln[\Big(\frac{1-y_r}{1-\theta}\Big)]\frac{d\zeta(m)}{dm} = 0. \tag{6.56}$$

After some algebraic manipulations while using repeatedly that

$$1 - \pi(\theta, m, \zeta(m)) = Pr(Y > \theta|\theta, m, \zeta(m)) = \frac{(1-\theta)m}{(1-\theta)m + \theta\zeta(m)} > 1 - r, \tag{6.57}$$

one obtains from (6.56) and (6.57) the relationship (6.53) for $\frac{d\zeta(m)}{dm}$. With condition (6.39) and $m > 0$, it now follows that $\frac{d\zeta(m)}{dm} > 0$. Furthermore, from Lemma 6.1 and (6.53), it follows that

$$0 < \zeta(m) < N(m|r, \theta) = m \times \frac{1-\theta}{\theta} \times \frac{r}{1-r}$$

and thus

$$\zeta(m) \downarrow 0 \text{ as } m \downarrow 0. \tag{6.58}$$

From Lemma 6.1 and given that the triplet $\{\theta, m, n^\bullet = \zeta(m)\}$ satisfies the constraint (6.41) for all $m > 0$, it follows with the definition of the pdf (6.3) that the $GTSP(\theta, m, \zeta(m))$ converges to a $Bernoulli$ distribution with a point mass r at $y = 0$. From (6.57), it follows immediately that $\pi(\theta, m, \zeta(m)) = Pr(Y \leq \theta|\theta, m, \zeta(m)) < r$. As a result, one obtains with (6.58) that $\pi(\theta, m, \zeta(m)) \uparrow r$ as $m \downarrow 0$ and with (6.53) it now follows that $N(m|r, \theta)$ is a tangent line of the implicit function $\zeta(m)$ at $m = 0$. □

Lemma 6.4 *Let $Y \sim GTSP(\theta, m, n)$ with pdf (6.3) and cdf (6.4). Under condition (6.39), one obtains with Lemmas 6.1 and 6.2 that the implicit function $\xi(n)$ defined in Lemma 6.1 is a continuous strictly increasing concave function in n.*

Proof Since the $GTSP(\theta, \xi(n), n)$ converges to a $Bernoulli$ distribution with probability mass p at $y = 0$ as $n \downarrow 0$ and $\pi(\theta, \xi(n), n) = Pr(Y \leq \theta|\theta, \xi(n), n) > p$ it follows that $\pi(\theta, \xi(n), n)$ is a strictly increasing function in n (since $\pi(\theta, \xi(n), n)$ decreases as n decreases). From condition (6.39), one obtains $(\theta/y_p) > 1$ and thus the function

$$\pi(\theta, \xi(n), n) + n\frac{\theta}{1-\theta}\ln[\frac{\theta}{y_p}]$$

is also a strictly increasing function in n. Next, with $1 - \pi(\theta, \xi(n), n)$ being a strictly decreasing function in n, it follows from (6.47) that $\frac{d\xi(n)}{dn}$ is a strictly decreasing function in n and therefore the implicit function $\xi(n)$ is, with Lemmas 6.1 and 6.2, a continuous strictly increasing concave function in n. □

Lemma 6.5 *Let $Y \sim GTSP(\theta, m, n)$ with pdf (6.3) and cdf (6.4). Under condition (6.39), one obtains with Lemmas 6.1 and 6.3 that the implicit function $\zeta(m)$ defined in Lemma 6.1 is a continuous strictly increasing concave function in m.*

Proof Since the $GTSP(\theta, m, \zeta(m))$ converges to a $Bernoulli$ distribution with probability mass r at $y = 0$ as $m \downarrow 0$ and $\pi(\theta, m, \zeta(m)) = Pr(Y \leq \theta|\theta, m, \zeta(m)) < r$ it follows that $\pi(\theta, m, \zeta(m))$ is a strictly decreasing function in m (since $\pi(\theta, m, \zeta(m))$ increases as m decreases). From condition (6.39), one obtains $(1-\theta)/(1-y_r) > 1$ and thus the function

$$1 - \pi[\theta, m, \zeta(m)] + m\frac{1-\theta}{\theta}\ln[\frac{1-\theta}{1-y_r}]$$

is a strictly increasing function in m. With $\pi(\theta, m, \zeta(m))$ being a strictly decreasing function in m, it now follows that $\frac{d\zeta(m)}{dm}$ given by (6.53) is a strictly decreasing function in m and therefore the implicit function $\zeta(m)$ is, with Lemmas 6.1 and 6.3, a continuous strictly increasing concave function in m. □

Theorem 6.1 *Let $Y \sim GTSP(\theta, m, n)$ with pdf (6.3) and cdf (6.4). Under condition (6.39), a unique pair of power parameters (m^*, n^*) exist for the $GTSP$ pdf (6.3) such that the quantile constraint set (6.6) is met.*

Proof From Lemmas 6.1 through 6.5, it follows that

$$\begin{aligned}
&a)\ \xi(n) \text{ is a strictly increasing continuous concave function in } n,\\
&b)\ \forall n > 0 : \xi(n) < M(n|p, \theta) = n \times \frac{\theta}{1-\theta} \times \frac{1-p}{p},\\
&c)\ \xi(n) \downarrow 0 \ \text{ when } n \downarrow 0, \text{and}\\
&d)\ M(n|p, \theta) \text{ is a tangent line to } \xi(n) \text{ at } n = 0.
\end{aligned}$$

and

$$\begin{aligned}
&a)\ \zeta(m) \text{ is a strictly increasing continuous concave function in } m,\\
&b)\ \forall m > 0 : \zeta(m) < N(m|r, \theta) = m \times \frac{r}{1-r} \times \frac{1-\theta}{\theta},\\
&c)\ \zeta(m) \downarrow 0 \ \text{ when } m \downarrow 0,\\
&d)\ N(m|r, \theta) \text{ is a tangent line to } \zeta(m) \text{ at } m = 0.
\end{aligned} \tag{6.59}$$

Denoting the inverse function of $\zeta(m)$ with $\zeta^{-1}(n)$, the following properties follow from the continuity of $\zeta(m)$ and from (6.59) for the function $\zeta^{-1}(n)$:

$$\begin{aligned}
&a)\ \zeta^{-1}(n) \text{ is a strictly increasing continuous convex function in } n,\\
&b)\ \forall n > 0 : \zeta^{-1}(n) > N^{-1}(n|r, \theta) = n \times \frac{1-r}{r} \times \frac{\theta}{1-\theta},\\
&c)\ \zeta^{-1}(n) \downarrow 0 \ \text{ when } n \downarrow 0,\\
&d)\ N^{-1}(n|r, \theta) \text{ is a tangent line to } \zeta^{-1}(n) \text{ at } n = 0.
\end{aligned}$$

Next, from condition (6.39), one obtains that

$$\frac{d}{dn}N^{-1}(n|r, \theta) = \frac{1-r}{r} \times \frac{\theta}{1-\theta} < \frac{1-p}{p} \times \frac{\theta}{1-\theta} = \frac{d}{dn}M(n|p, \theta).$$

From $N^{-1}(n|r, \theta)$ and $M(n|p, \theta)$ being tangent lines to $\zeta^{-1}(n)$ and $\xi(n)$, respectively, $\zeta^{-1}(n)$ being a strictly increasing continuous convex function in n, and $\xi(n)$ being a strictly increasing concave function in n, it now follows that $\zeta^{-1}(n)$ and $\xi(n)$ have a unique intersection point

$$[n^*, m^* = \zeta^{-1}(n^*)] = [n^*, m^* = \xi(n^*)],$$

where $n^* > 0$ such that at this point (n^*, m^*) the quantile constraint set (6.6) is met. □

References

AbouRizk, S. M., Halpin, D. W., & Wilson, J. R. (1991). Visual interactive fitting of beta distributions. *Journal of Construction Engineering and Management*, *117*(4), 589–605.

Cooke, R. (1991). *Experts in uncertainty: Opinion and subjective probability in science*. Oxford University Press on Demand.

DeBrota, D. J., Dittus, R. S., Roberts, S. D., & Wilson, J. R. (1989). Visual interactive fitting of bounded Johnson distributions. *Simulation*, *52*(5), 199–205.

Dekker, R. (1996). Applications of maintenance optimization models: A review and analysis. *Reliability Engineering & System Safety*, *51*(3), 229–240.

Dewispelare, A. R., Herren, L. T., & Clemen, R. T. (1995). The use of probability elicitation in the high-level nuclear waste regulation program. *International Journal of Forecasting*, *11*(1), 5–24.

van Dorp JR (1989) Expert opinion and maintenance data to determine lifetime distributions. Master's thesis, Delft University of Technology.

van Dorp, J. R., & Kotz, S. (2002). A novel extension of the triangular distribution and its parameter estimation. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *51*(1), 63–79.

van Dorp, J. R., & Kotz, S. (2003). Generalized trapezoidal distributions. *Metrika*, *58*(1), 85–97.

van Dorp, J. R., & Mazzuchi, T. A. (2000). Solving for the parameters of a beta distribution under two quantile constraints. *Journal of Statistical Computation and Simulation*, *67*(2), 189–201.

Ferguson, T. S. (1973) A Bayesian analysis of some nonparametric problems. *The annals of statistics* (pp. 209–230).

Garthwaite, P. H., Kadane, J. B., & O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, *100*(470), 680–701.

Herrerías-Velasco, J. M., Herrerías-Pleguezuelo, R., & van Dorp, J. R. (2009). The generalized two-sided power distribution. *Journal of Applied Statistics*, *36*(5), 573–587.

Mazzuchi, T. A., & Soyer, R. (1996). A Bayesian perspective on some replacement strategies. *Reliability Engineering & System Safety*, *51*(3), 295–303.

Mazzuchi, T. A., Noortwijk, J. M., & Kallen, M. J. (2007). *Maintenance optimization*. Wiley StatsRef: Statistics Reference Online.

Morris, D. E., Oakley, J. E., & Crowe, J. A. (2014). A web-based tool for eliciting probability distributions from experts. *Environmental Modelling & Software*, *52*, 1–4.

van Noortwijk, J. M., Dekker, A., Cooke, R. M., & Mazzuchi, T. A. (1992). Expert judgment in maintenance optimization. *IEEE Transactions on Reliability*, *41*(3), 427–432.

Oakley, J. E., O'Hagan, A. (2018). The SHeffield ELicitation Framework (SHELF). School of Mathematics and Statistics, University of Sheffield. Retrieved January 2019, from http://tonyohagan.co.uk/shelf/edn.

Pulkkinen, U., Simola, K. (2000). An expert panel approach to support risk-informed decision making. Technical report STUK-YTO-TR 129, Radiation and Nuclear Safety Authority of Finland STUK, Helsinki, Finland.

Shih, N. (2015). The model identification of beta distribution based on quantiles. *Journal of Statistical Computation and Simulation*, *85*(10), 2022–2032.

Susarla, V., & Van Ryzin, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *Journal of the American Statistical Association*, *71*(356), 897–902.

Wagner, M. A. F., & Wilson, J. R. (1996). Using univariate Bézier distributions to model simulation input processes. *IIE Transactions*, *28*(9), 699–711.

Chapter 7
Adversarial Risk Analysis as a Decomposition Method for Structured Expert Judgement Modelling

David Ríos Insua, David Banks, Jesús Ríos, and Jorge González-Ortega

Abstract We argue that adversarial risk analysis may be incorporated into the structured expert judgement modelling toolkit for cases in which we need to forecast the actions of competitors based on expert knowledge. This is relevant in areas such as cybersecurity, security, defence and business competition. As a consequence, we present a structured approach to facilitate the elicitation of probabilities over the actions of other intelligent agents by decomposing them into multiple, but simpler, assessments later combined together using a rationality model of the adversary to produce a final probabilistic forecast. We then illustrate key concepts and modelling strategies of this approach to support its implementation.

Keywords Structured expert judgement · Adversarial risk analysis · Decomposition · Security · Cybersecurity

7.1 Introduction

Structured Expert Judgement (SEJ) elicitation has a long history of successes, both in methodology and applications, many of them stemming from Roger Cooke's work, e.g. Cooke (1991) and Goossens et al. (1998). Hence, it has become a major ingredient within risk and decision analysis (Bedford and Cooke 2011). A significant feature in the practice of these disciplines, as already acknowledged in the classic book by Raiffa (1968), is the emphasis in decomposing complex problems into smaller pieces that are easier to understand and recombining the piecewise solutions to tackle the global problem.

D. R. Insua (✉) · J. González-Ortega
Instituto de Ciencias Matemáticas (CSIC-UAM-UC3M-UCM), Madrid, Spain
e-mail: david.rios@icmat.es

D. Banks
Department of Statistical Science (Duke University), Durham, NC, USA

J. Ríos
IBM Research Division (IBM), Yorktown Heights, NY, USA

A. M. Hanea et al. (eds.), *Expert Judgement in Risk and Decision Analysis*,
International Series in Operations Research & Management Science 293,
https://doi.org/10.1007/978-3-030-46474-5_7

In particular, belief assessment benefits from decomposition, typically through the argument of *extending the conversation*. Direct elicitation of probabilities can be a very difficult cognitive task. For example, there may be many factors influencing the occurrence of an outcome of interest whose effects experts would have to identify and balance in their heads to produce a probability judgement. Thus, rather than directly assessing this probability (with a standard SEJ technique), one could find a conditioning partition and estimate the probabilities of the outcome given the corresponding events. From these, and the probabilities of the conditioning events, the law of total probability enables calculation of the unconditional probability of the outcome. Ravinder et al. (1988) and Andradottir and Bier (1997, 1998) provide a methodological framework to validate the advantages of this approach, empirically tested in e.g. MacGregor and Kleinmuntz (1994) and MacGregor (2001). Tetlock and Gardner (2015) call this approach *Fermitisation* and present it as a key strategy for the success of their super-forecasters, and SEJ at large. Decompositions uncover the complexity underlying a direct probability assessment, eliminating the burden on experts to perform sophisticated modelling in their heads. This simplifies complex cognitive tasks, reveals assumptions experts make in their judgements and mitigate their reliance on heuristics that can introduce bias, ensuring that they actually analyse the relevant problem (Montibeller and von Winterfeldt 2015). Decompositions typically entail more assessments, though these tend to be simpler and more meaningful, leading to improved judgements and decisions. In turn, this would allow for better harnessing expert knowledge e.g. by assigning the proper expertise to the different sub-tasks of an assessment.

In many settings, especially in contexts such as security, counterterrorism or cybersecurity, experts will have to face adversarial problems in the sense that they need to deal with probabilities referring to actions carried out by opponents. As an example, in Chen et al. (2016), nearly 30% of the questions posed to experts somehow involved adversaries (e.g. *Will Syria use chemical or biological weapons before January 2013?*). Though we could think of using the standard SEJ tools as illustrated in other chapters in this volume, we present Adversarial Risk Analysis (ARA) as a decomposition strategy to support SEJ when forecasting adversarial actions. Regardless of the many issues associated with how an expert can translate domain knowledge into a probability, there is always the problem of how to best structure the elicitation process to get to a probability. When this is too difficult to assess but can be expressed as a combination of other simpler probabilities, decomposition becomes a critical part of the SEJ procedure. Our focus is on how ARA, as a structured SEJ technique, determines what the right questions to ask are and how experts' answers to these questions are combined to produce an adversarial probabilistic forecast.

After sketching the ARA approach to decomposition (Sect. 7.2), we show how this can actually improve expert assessment of opponent actions (Sect. 7.3). We then propose several ways to implement ARA in practice (Sect. 7.4), include a numerical example (Sect. 7.5) and end with a discussion (Sect. 7.6).

7.2 ARA as a SEJ Decomposition Method

ARA was originally introduced to deal with game-theoretic problems studied from a Bayesian perspective, (Ríos Insua et al. 2009; Banks et al. 2015). It stems from the observation that common knowledge assumptions in standard game-theoretic approaches based on Nash equilibria and their refinements do not hold in many applications, such as counterterrorism or cybersecurity, as competitors try to conceal information. Games are formulated in a Bayesian manner, as in Kadane and Larkey (1982) and Raiffa (2003), and operationalised by providing procedures to forecast the actions of the adversary.

To simplify the discussion, we consider the basic ARA approach through a sequential Defend–Attack game: agent D (she, defender) first makes her decision $d \in \mathcal{D}$, then agent A (he, attacker) observes d and chooses his alternative $a \in \mathcal{A}$. The outcome s of their interaction is a random variable S whose distribution depends upon d and a. As an example, imagine that a company deploys cybersecurity controls and then, having observed them, a cybercriminal decides whether to launch a cyberattack. The cost to the company would be a random variable that is conditioned upon both decisions (the controls deployed and the attack launched). The problem that agent D faces is depicted in the influence diagram in Fig. 7.1.

To solve it, she requires $p_D(s \mid d, a)$, which reflects her beliefs on the outcome given both agents' actions, and her utility function $u_D(d, s)$, modelling her preferences and risk attitudes over the consequences, which we assume depends on the outcome and the defence implemented. Besides, she needs the distribution $p_D(a \mid d)$, which is her assessment of the probability that A will choose action a after having observed her choice d. Once D has completed these judgements, she can compute the expected utility of decision d as

$$\psi_D(d) = \int \left[\int u_D(d, s)\, p_D(s \mid d, a)\, \mathrm{d}s \right] p_D(a \mid d)\, \mathrm{d}a,$$

and seek for the optimal decision $d^* = \arg\max_{d \in \mathcal{D}} \psi_D(d)$.

This is a standard risk or decision analysis exercise except for the elicitation of $p_D(a \mid d)$, which entails strategic aspects. D could try to assess it from a standard belief elicitation perspective, as in Cooke (1991) or O'Hagan et al. (2006), but ARA usefully suggests a decomposition approach to such assessment that requires her

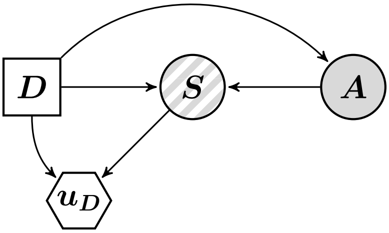


Fig. 7.1 The decision problem as seen by D

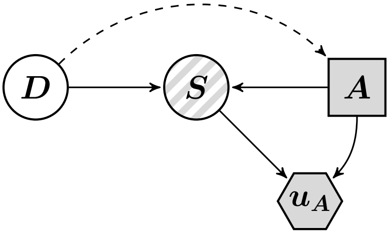


Fig. 7.2 *D*'s analysis of the decision problem as seen by *A*

to analyse the problem from *A*'s perspective, as shown in the influence diagram in Fig. 7.2.

Thus, *D* puts herself in *A*'s shoes. She would use all the information she can obtain about *A*'s probabilities $p_A(s \mid d, a)$ and utilities $u_D(d, s)$, assuming he is an expected utility maximiser. Then, instead of using point estimates for p_A and u_A to find *A*'s optimal response for a given *d*, her uncertainty about *A*'s decision would derive from her uncertainty about (p_A, u_A), through a distribution *F* on the space of probabilities and utilities. This induces a distribution over *A*'s expected utility, which for each *d* and *a* is

$$\Psi_A(d, a) = \int U_A(a, s)\; P_A(s \mid d, a)\, \mathrm{d}s,$$

where (P_A, U_A) follow the distribution of *F*. Then, *D* finds the required $p_D(a \mid d)$ as $\mathbb{P}_F\left[a = \arg\max_{x \in \mathcal{A}}\; \Psi_A(d, x)\right]$, in the discrete case and, analogously, in the continuous one. She could use Monte–Carlo simulation to approximate $p_D(a \mid d)$, as shown in Sects. 7.3 and 7.5.

Observe that the ARA approach weakens the standard, but unrealistic, common knowledge assumptions in game-theoretic approaches (Hargreaves-Heap and Varoufakis 2004), according to which the agents share information about their probabilities and utilities. In our case, not having common knowledge means that *D* does not know (p_A, u_A), and thus we model such uncertainty through *F*. The approach extends to simultaneous decision making problems, general interactions between both agents, multiple agents, agents who employ principles different than maximum expected utility, as well as to other contexts presented in Banks et al. (2015). Here, we exclusively explore the relevance of ARA as part of the SEJ toolkit.

7.3 Assessing ARA Decompositions

We hereafter study ARA as a decomposition approach through the sequential Defend–Attack model described above, comparing direct SEJ and the ARA decomposition.

7.3.1 Framework

As mentioned, there are two possible ways to assess the distribution $p_D(a \mid d)$:

- One could do it directly with standard SEJ procedures (Cooke 1991). Denote such assessment by $p_D^{SEJ}(a \mid d)$.
- Otherwise, one could determine it indirectly through ARA as in Sect. 7.2. D would model her uncertainty about A's beliefs and preferences, represented by $(P_A, U_A) \sim F$, and then solve A's decision making problem using these random probabilities and utilities to estimate

$$p_D^{ARA}(a \mid d) = \mathbb{P}_F \left[a = \text{arg max}_{x \in \mathcal{A}} \int U_A(x, s)\, P_A(s \mid d, x)\, \mathrm{d}s \right].$$

To compare both approaches, we make three simplifying assumptions:

(i) D has only two options, defend (d_1) or not (d_0);
(ii) A can solely choose between attacking (a_1) or not (a_0) and
(iii) if A decides to attack, the only two outcomes are success (s_1) or failure (s_0).

For A, the problem can be viewed as the decision tree in Fig. 7.3, with $d \in \{d_0, d_1\}$, which parallels the influence diagram in Fig. 7.2. The ARA approach obtains the required conditional probabilities $p_D^{ARA}(a \mid d)$ by solving the decision tree using D's (random) assessments over A's inputs.

Suppose D thinks A bases his decision on a cost–benefit analysis. In that case, the consequences for A are described in Table 7.1. For this, D might use a multi-attribute value model to decompose her judgement about A's valuation of consequences into simpler assessments regarding such costs and benefits. Later, she can aggregate these estimates as shown in the row *Profit* in Table 7.1, reflected in Fig. 7.3.

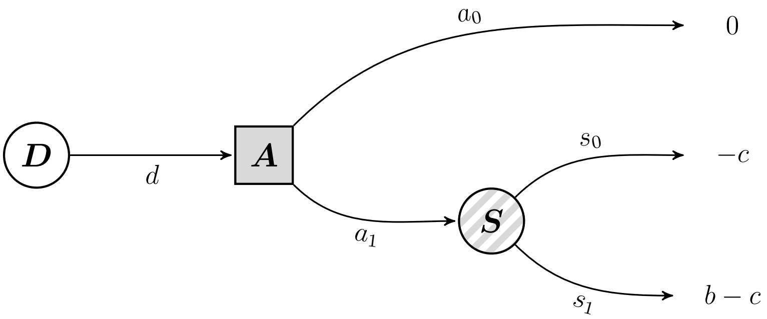


Fig. 7.3 Decision tree representing A's problem. c represents the cost of implementing an attack; b, the benefit of a successful attack

Table 7.1 Cost–benefit analysis of A's consequences

	(Attack, Outcome)—(a, s)		
	(a_0, s_0)	(a_1, s_0)	(a_1, s_1)
Cost	0	c	c
Benefit	0	0	b
Profit	0	$-c$	$b - c$

This requires D to assess two quantities: c and b, A's cost of undertaking an attack and his benefit if successful, respectively. We assume that $0 < c < b$, implying that attacking is more costly for A than not attacking, but potentially more beneficial, and that a successful attack is better for A than an unsuccessful one. Since D is generally uncertain about these quantities, she will provide probability distributions to model her beliefs about them. Suppose her self-elicitations lead to the uniform distributions

- A's cost of an attack: $c \sim \mathcal{U}(c_{\min}, c_{\max})$.
- A's benefit from a successful attack: $b \sim \mathcal{U}(b_{\min}, b_{\max})$.

These allow D to estimate the random values related to A's consequences in Table 7.1. We have assumed that D believes that A's costs and benefits are uniformly distributed and, very importantly, independent. However, in many cases, there is dependence; e.g. a more costly attack is most likely correlated with larger benefits for A. In that case, one needs to model c and b jointly. For simplicity, this discussion assumes independence.

If D believes that A is risk neutral (i.e. seeking to maximise expected profit), she would now elicit her beliefs about A's impression on his probability of success. Otherwise, beforehand, she would have to model A's risk attitudes. She could do that by eliciting a utility function over profits for him and model his risk attitude as shown in Sect. 7.4.2 and exemplified in Sect. 7.5, where her uncertainty about the attacker risk attitude is captured through a probability distribution over the risk aversion coefficient of a parametric utility function. Alternatively, because there are just three possible outcomes for A (no attack, failed attack, successful attack), D may directly assess her belief about his utility for each of them. Without loss of generality, utilities of 0 and 1 can be, respectively, assigned to the worst and best consequences for A. Since D believes that $-c < 0 < b - c$, $u_A(-c) = 0$ and $u_A(b - c) = 1$, even if she does not know the exact values of b and c. Thus, she just needs to elicit her distribution for $u_A(0) = u$, knowing that $0 < u < 1$, though being uncertain of A's exact value of u. Recall that this could be elicited as the probability at which A is indifferent between getting profit 0 for sure and a lottery ticket in which he wins $b - c$ with probability u and looses c with probability $1 - u$. This way, D could elicit a distribution for the random variable U_A that represents her full uncertainty over A's utility u.

Having done this, D would also need to assess A's beliefs about his chance of success, determined by $p_A(s_1 \,|\, d_0, a_1) = \pi_{d_0}$ and $p_A(s_1 \,|\, d_1, a_1) = \pi_{d_1}$. She should model her uncertainty about these with random probabilities $\pi_{d_0} \sim P_A^{d_0}$ and $\pi_{d_1} \sim P_A^{d_1}$, with $\pi_{d_1} < \pi_{d_0}$ to ensure that defending (d_1) reduces the chance of a successful attack. Then, based on the above assessments, for each $d \in \{d_0, d_1\}$, D can compute A's random expected utilities as

$$\Psi(d, a_0) = u_A(0) = u \sim U_A,$$

$$\Psi(d, a_1) = u_A(b - c) \times p_A(s_1 \,|\, d, a_1) + u_A(0) \times p_A(s_0 \,|\, d, a_1) = \pi_d \sim P_A^d,$$

and the ARA probabilities of attack, given the implemented defence, through

$$p_D^{ARA}(a_1 \,|\, d) = \mathbb{P}_{(U_A, P_A^d)}\left(u < \pi_d\right). \tag{7.1}$$

These probabilities represent the defender's ARA probabilistic predictions of how A will respond to each of her possible choices. As an example, suppose that we assess these distributions as $U_A \sim \mathcal{B}e(1, 2)$ (beta) and $P_A^{d_0} \sim \mathcal{U}(0.5, 1)$ and $P_A^{d_1} \sim \mathcal{U}(0.1, 0.4)$. Then, using Monte–Carlo (MC) simulation, we estimate the attack probabilities as $\hat{p}_D^{ARA}(a_1 \,|\, d_0) \approx 0.92$ and $\hat{p}_D^{ARA}(a_1 \,|\, d_1) \approx 0.43$ (based on an MC sample size of 10^6). In this case, choosing to defend (d_1) acts as a deterrent for A to attack (a_1).

7.3.2 Comparison

We now address whether this ARA decomposition approach leads to improved attack probability estimates over those obtained by direct SEJ methods. Adopting a normative viewpoint, we show through simulation that under certain conditions, the variance of the ARA estimates is smaller than those of the SEJ estimates.

In our case, due to the assumptions behind expression (7.1), we have no reason to believe that D finds one attack distribution more (or less) likely than another, except that an attack is more likely when no defence is attempted. That is, $p_{d_0}^{SEJ} \geq p_{d_1}^{SEJ}$ where $p_{d_i}^{SEJ} = p_D^{SEJ}(a_1 \,|\, d_i), i = 1, 2$. Thus, as a high-entropy benchmark, we assume that $p_{d_0}^{SEJ}, p_{d_1}^{SEJ}$ are uniformly distributed over the set $\{0 \leq p_{d_1}^{SEJ} \leq p_{d_0}^{SEJ} \leq 1\}$, whose variance–covariance matrix is analytically computed as

$$\begin{pmatrix} \frac{1}{18} & \frac{1}{36} \\ \frac{1}{36} & \frac{1}{18} \end{pmatrix} \approx \begin{pmatrix} 5.56 & 2.78 \\ 2.78 & 5.56 \end{pmatrix} \cdot 10^{-2}. \tag{7.2}$$

In turn, D's assessment of the ARA attack probabilities involves eliciting distributions $(U_A, P_A^{d_0}, P_A^{d_1})$. It is reasonable to assume that u is independent of π_{d_0} and π_{d_1}. Since the support of all three random variables is [0, 1], an equitable framework for the benchmark may assume that $U_A \sim \mathcal{U}(0, 1)$ and $(P_A^{d_0}, P_A^{d_1})$ are uniformly distributed over the set $\{0 \leq \pi_{d_1} \leq \pi_{d_0} \leq 1\}$. We computed 10^4 MC estimates of the attack probabilities using these distributions, each based on an MC sample size of 10^4, leading to a variance–covariance matrix for $p_{d_0}^{ARA}$ and $p_{d_1}^{ARA}$ of

$$\begin{pmatrix} 2.24 & 1.10 \\ 1.10 & 2.22 \end{pmatrix} \cdot 10^{-5}. \tag{7.3}$$

Thus, as a result of the decomposition approach inherent to the ARA methodology, both variance and the covariance in the ARA approach (7.3) are significantly smaller than those in the benchmark (7.2), providing a more precise assessment.

Yet, typically, one would have more information about $(U_A, P_A^{d_0}, P_A^{d_1})$. For example, suppose D believes that the mean values of the three random variables are $E[U_A] = \frac{2}{5}$, $E[P_A^{d_0}] = \frac{2}{3}$ and $E[P_A^{d_1}] = \frac{1}{3}$. If she assumes they all are uniformly distributed with maximum variance, then $U_A \sim \mathcal{U}(0, \frac{4}{5})$, $P_A^{d_0} \sim \mathcal{U}(\frac{1}{3}, 1)$ and $P_A^{d_1} \sim \mathcal{U}(0, \frac{2}{3})$ (with $\pi_{d_1} \leq \pi_{d_0}$). In this case, the estimated variance–covariance matrix for $p_{d_0}^{ARA}$ and $p_{d_1}^{ARA}$ is

$$\begin{pmatrix} 1.42 & 0.65 \\ 0.65 & 2.35 \end{pmatrix} \cdot 10^{-5}.$$

Compared to (7.3), these assumptions reduce the variance for $p_{d_0}^{ARA}$ and the covariance, although slightly increase the variance of $p_{d_1}^{ARA}$. Finally, if the random variables followed beta distributions with common variance $\frac{1}{10}$, then $U_A \sim \mathcal{B}e(0.56, 0.84)$, $P_A^{d_0} \sim \mathcal{B}e(0.81, 0.41)$ and $P_A^{d_1} \sim \mathcal{B}e(0.41, 0.81)$ (and $\pi_{d_1} \leq \pi_{d_0}$), and the variance–covariance matrix for $p_{d_0}^{ARA}$ and $p_{d_1}^{ARA}$ is

$$\begin{pmatrix} 1.52 & 0.64 \\ 0.64 & 2.25 \end{pmatrix} \cdot 10^{-5}.$$

Again, the covariance matrix is significantly more precise than the benchmark.

For further insights, assume that the direct elicitation process incorporates additional information, so that $p_{d_0}^{SEJ}$ and $p_{d_1}^{SEJ}$ are now uniformly distributed over the set $\{\varepsilon \leq p_{d_1}^{SEJ} \leq p_{d_0}^{SEJ} \leq 1 - \varepsilon\}$, requiring $0 \leq \varepsilon \leq \frac{1}{2}$ to be defined. Then, the variance–covariance matrix for $p_{d_0}^{SEJ}$ and $p_{d_1}^{SEJ}$ is

$$\begin{pmatrix} \frac{(1-2\varepsilon)^2}{18} & \frac{(1-2\varepsilon)^2}{36} \\ \frac{(1-2\varepsilon)^2}{36} & \frac{(1-2\varepsilon)^2}{18} \end{pmatrix}. \tag{7.4}$$

From (7.3) and (7.4), we see that one must take $\varepsilon > 0.49$, a very precise assessment, so that the corresponding variance–covariance matrix of $p_{d_0}^{SEJ}$ and $p_{d_1}^{SEJ}$ becomes less variable than $p_{d_0}^{ARA}$ and $p_{d_1}^{ARA}$.

All these comparisons indicate that although the ARA approach requires more assessments to obtain the relevant probabilities of the adversarial actions, ARA tends to provide more precise estimates. However, if the direct information is very precise, then direct elicitation can outperform ARA in terms of reduced variance for the relevant probabilities.

7.4 ARA Modelling Strategies

We have shown that the ARA decomposition can have advantages over the plain SEJ approach. Consequently, it is worth describing how to implement it. We thus present a catalogue of strategies to model the random probabilities and utilities necessary to put ARA into practice.

7.4.1 Random Probabilities

We focus first on D's assessments over A's perspective of the different random events involved in the problem, that is, the random probabilities. To fix ideas, assume we have a single chance node S which depends on both D's and A's choices. Our task is to develop a (random) distribution $P_A(s \mid d, a)$ that reflects D's uncertainty about A's prospect of S. We distinguish three cases. In all of them, as shown in Sect. 7.4.1.1, Bayesian updating could be used to dynamically adjust the assessed priors as data accumulates, thus attaining subsequent random posterior distributions that better reflect D's information and perspective over A's uncertainty.

7.4.1.1 Probability of a Single Event

Suppose first that the chance node S consists of a single event which may $(s = 1)$ or not $(s = 0)$ happen. Then, $p_A(s \mid d, a)$ is completely determined by $p_A(s = 1 \mid d, a)$, for each pair (d, a), as $p_A(s = 0 \mid d, a) = 1 - p_A(s = 1 \mid d, a)$.

One possibility would be to base $P_A(s = 1 \mid d, a)$ on an estimate π_D of $p_A(s = 1 \mid d, a)$, with some uncertainty around it. This may be accomplished in several ways. We could do it through a uniform distribution $\mathcal{U}(\pi_D - \mu, \pi_D + \mu)$ centred around π_D in which the parameter μ would have to be assessed also. For example, if we get that the expected variance of the distribution is ν, we get $\mu = \sqrt{3\nu}$. Another option would be to use a beta distribution $\mathcal{B}e(\alpha, \beta)$ in which π_D may be regarded as the mean (or the median or the mode) of the distribution and we would have to assess the parameters α and β to shape the distribution, e.g. based on a further assessment of the variance ν. This would lead, when π_D is the mean, to

$$\alpha = \frac{\pi_D}{\nu}\left(\pi_D\left(1-\pi_D\right)-\nu\right), \qquad \beta = \frac{1-\pi_D}{\nu}\left(\pi_D\left(1-\pi_D\right)-\nu\right)$$

Note that when D thinks that A has information similar to hers, an adequate best guess for π_D could be based on her own assessment $p_D(s = 1 \mid d, a)$.

If the possible occurrence of event s were to be repeated over time, random prior distributions could be reassessed by means of Bayesian updating. Consider, for example, the second case in which a beta distribution $\mathcal{B}e(\alpha, \beta)$ is used. If event s has had y opportunities to happen and materialises only z of them, our random posterior would be $\mathcal{B}e(\alpha + z, \beta + y - z)$.

7.4.1.2 Probabilities of Multiple Events

We assume now that the chance node S includes N events $\{s_1, \ldots, s_N\}$. In this case, probabilities $p_A(s = s_1 \mid d, a), \ldots, p_A(s = s_{N-1} \mid d, a)$ determine $p_A(s \mid d, a)$ completely, for each pair (d, a), as $p_A(s = s_N \mid d, a) = 1 - \sum_{n=1}^{N-1} p_A(s = s_n \mid d, a)$. Therefore, we only need to model $P_A(s = s_1 \mid d, a), \ldots, P_A(s = s_{N-1} \mid d, a)$, which we jointly designate $P_A(s \mid d, a)$.

In line with the previous case, we could base $P_A(s \mid d, a)$ on a best guess $\pi_D(s)$, for example $p_D(s \mid d, a)$ when D believes that A has similar information, with some uncertainty around it. We could use a parametric probability distribution, randomising each of its parameters much as we have done in the preceding subsection. In this manner, for each pair d and a, we could estimate $\pi_{D,n}$ of $p_A(s = s_n \mid d, a)\, \forall n \in \{1, \ldots, N-1\}$ and, then, incorporate the uncertainty through a uniform $\mathcal{U}(\pi_{D,n} - \mu_n, \pi_{D,n} + \mu_n)$ or a beta distribution $\mathcal{B}e(\alpha_n, \beta_n)$ centred around $\pi_{D,n}$, making sure that their sum does not exceed 1.

A more effective way would model $P_A(s \mid d, a)$ as a Dirichlet distribution with mean $\pi_D(s)$ and parameters assessed based on one further judgement concerning, e.g. the variance of one of the probabilities. To do this, for each pair (d, a), we would obtain from D an estimate $\pi_{D,n}$ of $p_A(s = s_n \mid d, a)\ \forall n \in \{1, \ldots, N\}$ and associate random variables S_n such that $E\,[S_n] = \pi_{D,n}$. Their joint distribution could then be described as Dirichlet, $(S_1, \ldots, S_N) \sim \mathcal{D}ir(\alpha)$, with parameters $\alpha = (\alpha_1, \ldots, \alpha_N)$. If $\hat{\alpha} = \sum_{n=1}^{N} \alpha_n$, it follows that

$$E\,[S_n] = \frac{\alpha_n}{\hat{\alpha}}, \qquad Var\,[S_n] = \frac{\alpha_n(\hat{\alpha} - \alpha_n)}{\hat{\alpha}^2(\hat{\alpha} + 1)};$$

and it suffices to elicit one value, e.g. $Var\,[S_1]$, to calculate the required α_n parameters.

7.4.1.3 The Continuous Case

We consider now the case in which the chance node S involves a continuous set of events. Techniques are similar to those described to assess the probabilities of multiple events. We could base $P_A(s \mid d, a)$ on a guess $\pi_D(s)$, say $p_D(s \mid d, a)$, with some uncertainty around it. For example, this may be achieved by means of a Dirichlet process, with base distribution $\pi_D(s)$ and concentration parameter ρ as perceived by D, which allows to sample approximate distributions of $P_A(s \mid d, a)$. Other non-parametric approaches such as hierarchical Pitman–Yor processes (Teh and Jordan 2010) could be used with reference to the above guess.

7.4.2 Random Utilities

We draw now attention over D's beliefs on A's preference assessments over the consequences of the decisions, that is, the random utilities. We shall usually have some information about A's multiple interests. For example, when dealing with terrorism cases, Keeney (2007) and Keeney and von Winterfeldt (2010) present extensive classifications of criteria amongst which to choose. Keeney (2007) then advocates that standard utility methods may be adopted by interviewing experts in the problem at hand, therefore developing utility functions modelling A's preferences. However, note that such preferences are not directly elicited from A, but rather through a surrogate. Thus, intrinsically, there is uncertainty about A's preferences.

An alternative approach, illustrated in Banks et al. (2015), is to aggregate the objectives with a weighted measurable value function, as in Dyer and Sarin (1979). As an example, we could consider an additive value function for A in which his objectives $v_1, \ldots, v_R$ are aggregated using weights $w_1, \ldots, w_R \geq 0$, $\sum_{r=1}^{R} w_r = 1$ as $v_A = \sum_{r=1}^{R} w_r \, v_r$. The uncertainty about the weights could be modelled using a Dirichlet distribution, as in Sect. 7.4.1.2, so that we may estimate their value and then associate random variables W_r such that $E\,[W_r] = w_r$, their joint distribution being Dirichlet, $(W_1, \ldots, W_R) \sim \mathcal{D}ir(\alpha)$, with parameters $\alpha = (\alpha_1, \ldots, \alpha_R)$ with one further judgement, e.g. fixing the variance of one of the parameters. Finally, using the relative risk aversion concept (Dyer and Sarin 1982), we could assume different risk attitudes when modelling A's utility function. Continuing the example and assuming an exponential utility function, we may transform the (random) value function $V_A = \sum_{r=1}^{R} W_r \, v_r$ into one of the three following utilities depending on A's risk attitude: *risk aversion*, $U_A = 1 - \exp(-\lambda \, V_A + c), \quad \lambda > 0$; *risk neutrality*, $U_A = V_A + c$; or *risk proneness*, $U_A = \exp(\lambda \, V_A + c), \quad \lambda > 0$. Further uncertainty about the risk coefficient λ and the adjusting constant c may be modelled, e.g. through uniform distributions $\Lambda \sim \mathcal{U}(\lambda_1, \lambda_2)$ and $C \sim \mathcal{U}(c_1, c_2)$. In any case, to determine all the required distributions, we may ask experts to directly elaborate such distributions or request them to provide point estimates of the weights and coefficients and build the distributions from these.

An alternative to building a distribution over A's preferences is described in Wang and Bier (2013). As before, suppose that they are represented through a multi-attribute utility function, which involves the above attributes $v_1, \ldots, v_R$ as well as an unobserved one v_0. For simplicity, consider A's utility to be linear in the attributes. Then we ask several experts to provide rank orders of A's action valuations and derive probability distributions that can match those orderings to obtain the (random) weights $(W_0, W_1, \ldots, W_R)$ for his utility function. For this, we consider as input such rankings and as output a distribution over A's preferences (expected utilities) for which two methods are suggested. One is an adaptation of probabilistic inversion (Neslo et al. 2008); essentially, it identifies a probability distribution Q over the space of all possible attribute weights $(W_0, W_1, \ldots, W_R)$ that can match the empirical distribution matrix of expert rankings with minimum Kullback–Leibler divergence to a predetermined (e.g. non-informative, Dirichlet) starting probability measure Q_0. The other one uses Bayesian density estimation (Müller et al. 2015) based on a prior distribution Q_p (e.g. chosen in accordance to a Dirichlet process with base distribution Q_0) over the space of attribute weights $(W_0, W_1, \ldots, W_R)$ and treating the expert rankings as observations to update that prior leading to a posterior distribution Q, obtained through the Gibbs sampling.

7.5 A Numerical Example

As an illustration, consider a sequential defend–attack cybersecurity problem. A user (D, defender) needs to make a connection to a site, either through a safe, but costly, route (d_0) or through a cheaper, but more dangerous protocol. In the latter case, she may use a security key, rendering the protocol less dangerous. While using the dangerous protocol, whether unprotected (d_1) or protected by a security key (d_2), the defender may be the target of a cybercriminal (A, attacker) who may decide to attack (a_1) or not (a_0). The case may be viewed through the game tree in Fig. 7.4.
The following parameters are used:

(i) h is the cost of using the expensive protocol;
(ii) θ_1 is the fraction of assets lost by the defender when attacked and unprotected;
(iii) θ_2 is the fraction of assets lost by the defender when attacked but protected;
(iv) k is the security key's cost;
(v) c is the defender's scaling cost relative to the fraction of assets lost;
(vi) L is the uncertain cost of an attack and
(vii) G is the uncertain cybercriminal's scaling gain relative to the fraction of assets lost by the defender.

Table 7.2 (respectively, Table 7.3) displays the defender's (respectively, attacker's) consequences, expressed as costs, for the various defend and attack possibilities, reflected in the tree

Table 7.2 Defender's loss function

		Attack	
		a_0	a_1
Defence	$\boldsymbol{d_0}$	h	–
	$\boldsymbol{d_1}$	0	$c\,\theta_1$
	$\boldsymbol{d_2}$	k	$k + c\,\theta_2$

Table 7.3 Attacker's loss function

		Attack	
		a_0	a_1
Defence	$\boldsymbol{d_0}$	0	–
	$\boldsymbol{d_1}$	0	$L - G\,\theta_1$
	$\boldsymbol{d_2}$	0	$L - G\,\theta_2$

The defender believes that the asset fractions θ_i follow distributions $p_D(\theta_i \mid d_i, a_1)$ with $\theta_i \sim \mathcal{B}e(\alpha_i^D, \beta_i^D)$, $i = 1, 2$. She is risk averse and her utility function is strategically equivalent to $1 - e^{\lambda_D\,x}$, where x is her cost and $\lambda_D > 0$ her risk aversion coefficient. She expects θ_1 to be greater than θ_2 (but not necessarily), reflected in the choice of the beta parameters, with $E\,[\theta_1] = \frac{\alpha_1^D}{\alpha_1^D + \beta_1^D} > \frac{\alpha_2^D}{\alpha_2^D + \beta_2^D} = E\,[\theta_2]$. Table 7.4 provides the defender's expected utilities u_D under the various interaction scenarios.

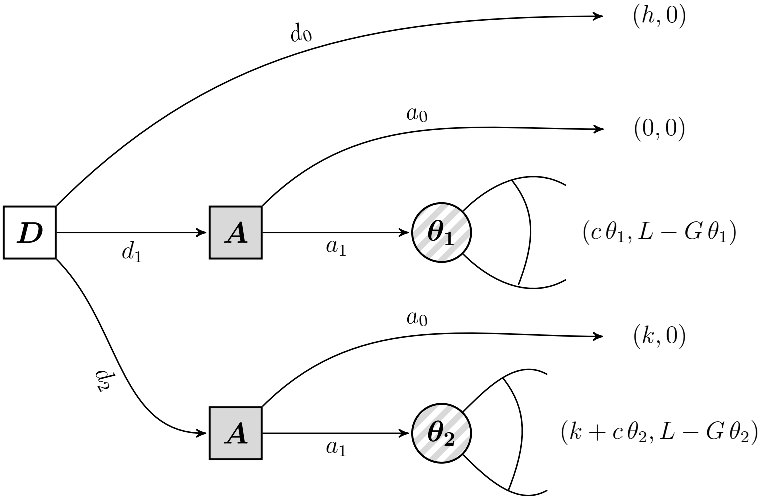


Fig. 7.4 Game tree for the cybersecurity routing problem (losses). Outcomes after θ_i, $i = 1, 2$ are continuous

Table 7.4 Defender's expected utility

		Attack	
		a_0	a_1
Defence	$\boldsymbol{d_0}$	$1 - e^{\lambda_D h}$	–
	$\boldsymbol{d_1}$	0	$1 - \int e^{\lambda_D c\, \theta_1}\, p_D(\theta_1)\, \mathrm{d}\theta_1$
	$\boldsymbol{d_2}$	$1 - e^{\lambda_D k}$	$1 - \int e^{\lambda_D (k + c\, \theta_2)}\, p_D(\theta_2)\, \mathrm{d}\theta_2$

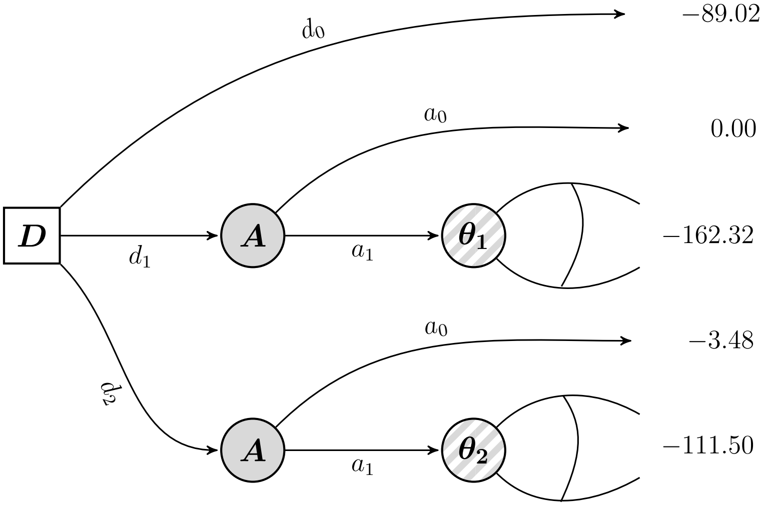


Fig. 7.5 Decision tree representing the defender's problem (expected utilities)

Suppose we assess from the defender the following parameter values (with standard elicitation techniques):

(i) a protocol cost $h = 150,000$ €;
(ii) a security key cost $k = 50,000$ €;
(iii) a scaling cost $c = 200,000$ €;
(iv) a risk aversion coefficient $\lambda_D = 3 \cdot 10^{-5}$;
(v) the distribution $\theta_1 \sim \mathcal{B}e(\alpha_1^D, \beta_1^D)$ with expected fraction (mean) of 0.6 of the assets lost and standard deviation 0.15 when attacked and unprotected, leading to $\alpha_1^D = 0.36$ and $\beta_1^D = 0.24$
(vi) and the distribution $\theta_2 \sim \mathcal{B}e(\alpha_2^D, \beta_2^D)$ with expected fraction (mean) of 0.3 of the assets lost and standard deviation 0.07 when attacked but protected, leading to $\alpha_2^D = 0.6$ and $\beta_2^D = 1.4$.

These are standard decision analytic assessments and the resulting problem faced by her is described in the decision tree in Fig. 7.5.

The expected utility of the first alternative (d_0, use the expensive protocol) may be directly estimated as

Table 7.5 Attacker's random expected utility

		Attack	
		a_0	a_1
Defence	$\boldsymbol{d_0}$	0	–
	$\boldsymbol{d_1}$	0	$\int e^{\Lambda_A (G\,\theta_1 - L)}\, P_A(\theta_1)\, \mathrm{d}\theta_1 - 1$
	$\boldsymbol{d_2}$	0	$\int e^{\Lambda_A (G\,\theta_2 - L)}\, P_A(\theta_2)\, \mathrm{d}\theta_2 - 1$

$$\psi_D(d_0) = 1 - e^{\lambda_D h} \approx -89.02,$$

since there is no chance of attack in this scenario. However, those of the other two alternatives have the form

$$\psi_D(d_i) = \sum_{j=0}^{1} p_D(a_j \,|\, d_i)\, u_D(d_i, a_j), \quad i = 1, 2;$$

where $u_D(d_i, a_j)$ may be obtained from Table 7.4 with the specific values indicated in Fig. 7.5. Thus, we need to assess the attack probabilities $p_D(a_1 \,|\, d_i)$ (and $p_D(a_0 \,|\, d_i) = 1 - p_D(a_1 \,|\, d_i), i = 1, 2$) and we adopt an ARA approach to assess them.

The attacker has different beliefs about θ_i, $p_A(\theta_i \,|\, d_i, a_1)$, with $\theta_i \sim \mathcal{B}e(\alpha_i^A, \beta_i^A)$, $i = 1, 2$; the defender's uncertainty about α_i^A and β_i^A inducing its randomness. He is risk prone and his utility function is strategically equivalent to $e^{-\Lambda_A x} - 1$, where x is his cost and $\Lambda_A > 0$ his uncertain risk proneness coefficient. Table 7.5 provides the attacker's random expected utilities, respectively, U_A under the various interaction scenarios.

Suppose that, in line with Sect. 7.4, we assess that

(i) $L \sim \mathcal{U}(10^4, 2 \cdot 10^4)$ with an expected cost of $15,000$ €;
(ii) $G \sim \mathcal{U}(10^4, 5 \cdot 10^4)$ with an expected scaling gain of $30,000$ €;
(iii) $\Lambda_A \sim \mathcal{U}(10^{-4}, 2 \cdot 10^{-4})$ with an expectation of $1.5 \cdot 10^{-4}$;
(iv) the distribution $\theta_1 \sim \mathcal{B}e(\alpha_1^A, \beta_1^A)$ has a expected fraction (mean) of 0.6 assets lost when the defender is attacked but protected, with $\alpha_1^A \sim \mathcal{U}(5, 7)$ and $\beta_1^A \sim \mathcal{U}(3, 5)$ and
(v) the distribution $\theta_2 \sim \mathcal{B}e(\alpha_2^A, \beta_2^A)$ has a expected fraction (mean) of 0.3 assets lost when the defender is attacked but protected, with $\alpha_2^A \sim \mathcal{U}(2, 4)$ and $\beta_2^A \sim \mathcal{U}(6, 8)$.

We may then use Algorithm 1 to estimate the required probabilities $\hat{p}_D(a_1 \,|\, d)$, where $\Psi_A^n(d_i, a)$ designates the expected utility that the cybercriminal obtains when the defender implements d, he chooses action a and the sampled parameters are $l^n, g^n, \lambda_A^n, \alpha_i^{A,n}$ and $\beta_i^{A,n}$.

Algorithm 1 Numerical example: Simulation of $\hat{p}_D(a_1 \mid d)$

Data: Number of iterations N.

```
Set $p_1, p_2 = 0$.
For $n = 1$ to $N$ do
    Draw $l^n$ from $\mathcal{U}(10^4, 2 \cdot 10^4)$, $g^n$ from $\mathcal{U}(10^4, 5 \cdot 10^4)$.
    Draw $\lambda_A^n$ from $\mathcal{U}(10^{-4}, 2 \cdot 10^{-4})$.
    Draw $\alpha_1^{A,n}$ from $\mathcal{U}(2, 7)$, $\beta_1^{A,n}$ from $\mathcal{U}(1, 5)$.
    Draw $\alpha_2^{A,n}$ from $\mathcal{U}(0, 3)$, $\beta_2^{A,n}$ from $\mathcal{U}(1, 6)$.
    For $i = 1$ to 2 do
        $\Psi_A^n(d_i, a_0) = 0$.
        $\Psi_A^n(d_i, a_1) = \int e^{\lambda_A^n (g^n \theta_i - l^n)} \frac{\theta_i^{\alpha_i^{A,n}-1} (1-\theta_i)^{\beta_i^{A,n}-1}}{\mathrm{Beta}(\alpha_i^{A,n}, \beta_i^{A,n})} \, \mathrm{d}\theta_i - 1$.
        If $\Psi_A^n(d_i, a_1) \geq \Psi_A^n(d_i, a_0)$ then
            $p_i = p_i + 1$.
        End If
    End For
End For
For $i = 1$ to 2 do
    $\hat{p}(a_1 \mid d_i) = p_i / N$.
End For
```

In our case, with $N = 10^6$, we obtain $\hat{p}(a_1 \mid d_1) = 0.66$ (and, consequently, $\hat{p}(a_0 \mid d_1) = 0.34$). Similarly, $\hat{p}(a_1 \mid d_2) = 0.23$ (and $\hat{p}(a_0 \mid d_2) = 0.77$). Then, we have $\psi_D(d_0) = -89.02$, $\psi_D(d_1) = -107.13$ and $\psi_D(d_2) = -28.32$. Thus, the optimal cyberdefense is $d^*_{ARA} = d_2$, that is, employing the dangerous protocol protected by the security key.

7.6 Discussion

ARA is an emergent paradigm when supporting a decision maker who faces adversaries so that the attained consequences are random and depend on the actions of all participating agents. We have illustrated the relevance of such an approach as a decomposition method to forecast adversarial actions in competitive contexts, therefore being of relevance to the SEJ toolkit. We have also presented key implementation strategies. We have limited the analysis to the simpler sequential case, but ideas extend to simultaneous problems, albeit with technical difficulties, due to the belief recursions typical of level-k thinking.

As usual, in applications, this tool could be combined with other SEJ strategies. For example, when assessing $p_D(a|d)$, we could use extending the conversation through $\sum_i p_D(a|b_i, d)p_D(b_i)$ and then assess the $p_D(a|b_i, d)$ probabilities through ARA. Similarly, throughout the discussion, we have assumed just one single expert available to provide the $p(a|d)$ probabilities through ARA. In practice, several experts might be available and we could aggregate their ARA probabilities through e.g. Cooke's classical method (Cooke 1991). Diverse adversarial rationalities, such as non-strategic or prospect-maximising players, could be handled by means of mixtures.

The ARA decomposition strategy breaks down an attack probability assessment into (random) multi-attribute utility and probability assessments for the adversary. This approach may lead to more precise probabilities than the ones that would have been directly obtained and, also, that the corresponding increased number of necessary judgements are cognitively easier. Behavioural experiments will be conducted to validate these ideas.

Acknowledgments The work of DRI is supported by the Spanish Ministry of Economy and Innovation programs MTM2014-56949-C3-1-R and MTM2017-86875-C3-1-R AEI/FEDER, UE, the ESF-COST Action IS1304 on Expert Judgement and the AXA-ICMAT Chair on Adversarial Risk Analysis. JGO's research is financed by the Spanish Ministry of Economy and Competitiveness under FPI SO grant agreement BES-2015-072892. This work has also been partially supported by the Spanish Ministry of Economy and Competitiveness, through the "Severo Ochoa" Program for Centers of Excellence in R&D (SEV-2015-0554) and project MTM2015-72907-EXP received funding from the European Union's H2020 Program for Research, Technological Development and Demonstration, under grant agreement no. 740920 (CYBECO) and based upon work partially supported by the US National Science Foundation through grant DMS-163851.

References

Andradottir, S., & Bier, V. M. (1997). Choosing the number of conditioning events in judgemental forecasting. *Journal of Forecasting*, *16*(4), 255–286.

Andradottir, S., & Bier, V. M. (1998). An analysis of decomposition for subjective estimation in decision analysis. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, *28*(4), 443–453.

Banks, D., Ríos, J., & Ríos Insua, D. (2015). *Adversarial risk analysis*. Boca Raton, FL: CRC Press.

Bedford, T., & Cooke, R. M. (2011). *Probabilistic risk analysis: Foundations and methods (first published 2001)*. Cambridge, UK: Cambridge University Press.

Chen, E., Budescu, D. V., Lakshmikanth, S. K., Mellers, B. A., & Tetlock, P. E. (2016). Validating the contribution-weighted model: Robustness and cost-benefit analyses. *Decision Analysis*, *13*(3), 128–152.

Clemen, R. T., & Reilly, T. (2013). *Making hard decisions with decisiontools*. Mason, OH: Cengage Learning.

Cooke, R. M. (1991). *Experts in uncertainty: Opinion and subjective probability in science*. New York, NY: Oxford University Press.

Dyer, J. S., & Sarin, R. K. (1979). Group preference aggregation rules based on strength of preference. *Management Science*, *25*(9), 822–832.

Dyer, J. S., & Sarin, R. K. (1982). Relative risk aversion. *Management Science*, *28*(8), 875–886.

French, S., & Ríos Insua, D. (2000). *Statistical decision theory*. New York, NY: Wiley.

González-Ortega, J., Radovic, V., & Ríos Insua, D. (2018). Utility elicitation. In *Elicitation: The science and art of structuring judgement* (pp. 241–264), Springer, New York, NY.

Goossens, L. H. J., Cooke, R. M., & Kraan, B. C. P. (1998). Evaluation of weighting schemes for expert judgement studies (Vol. 1937–1942).

Hargreaves-Heap, S., & Varoufakis, Y. (2004). *Game theory: A critical introduction (first published 1995)*. New York, NY: Routledge.

Kadane, J. B., & Larkey, P. D. (1982). Subjective probability and the theory of games. *Management Science*, *28*(2), 113–120.

Keeney, G. L., & von Winterfeldt, D. (2010). Identifying and structuring the objectives of terrorists. *Risk Analysis*, *30*(12), 1803–1816.

Keeney, R. L. (2007). Modeling values for anti-terrorism analysis. *Risk Analysis*, *27*(3), 585–596.

MacGregor, D. G. (2001). Decomposition for judgmental forecasting and estimation. In *Principles of Forecasting* (pp. 107–123). Boston, MA: Springer.

MacGregor, D. G., & Armstrong, J. S. (1994). Judgmental decomposition: When does it work? *International Journal of Forecasting*, *10*(4), 495–506.

Montibeller, G., & von Winterfeldt, D. (2015). Biases and debiasing in multi-criteria decision analysis. *IEEE 2015 48th Hawaii International Conference on System Sciences* (pp. 1218–1226).

Müller, P., Quintana, F. A., Jara, A., & Hanson, T. (2015). *Bayesian nonparametric data analysis*. Cham, Switzerland: Springer.

Neslo, R., Micheli, F., Kappel, C. V., Selkoe, K. A., Halpern, B. S., & Cooke, R. M. (2008). Modeling stakeholder preferences with probabilistic inversion: Application to prioritizing marine ecosystem vulnerabilities. In *Real-Time and deliberative decision making* (pp. 265–284). Dordrecht, Netherlands: Springer.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., et al. (2006). *Uncertain judgements: Eliciting experts' probabilities*. Chichester, UK: Wiley.

Raiffa, H. (1968). *Decision analysis: Introductory lectures on choices under uncertainty*. Menlo Park, CA: Addison-Wesley.

Raiffa, H. (2003). *The art and science of negotiation* (first published 1982). Cambridge, MA: Harvard University Press. arth

Ravinder, H. V., Kleinmuntz, D. N., & Dyer, J. S. (1988). The reliability of subjective probabilities obtained through decomposition. *Management Science*, *34*(2), 186–199.

Ríos Insua, D., Ríos, J., & Banks, D. (2009). Adversarial risk analysis. *Journal of the American Statistical Association, 104*(486), 841–854.

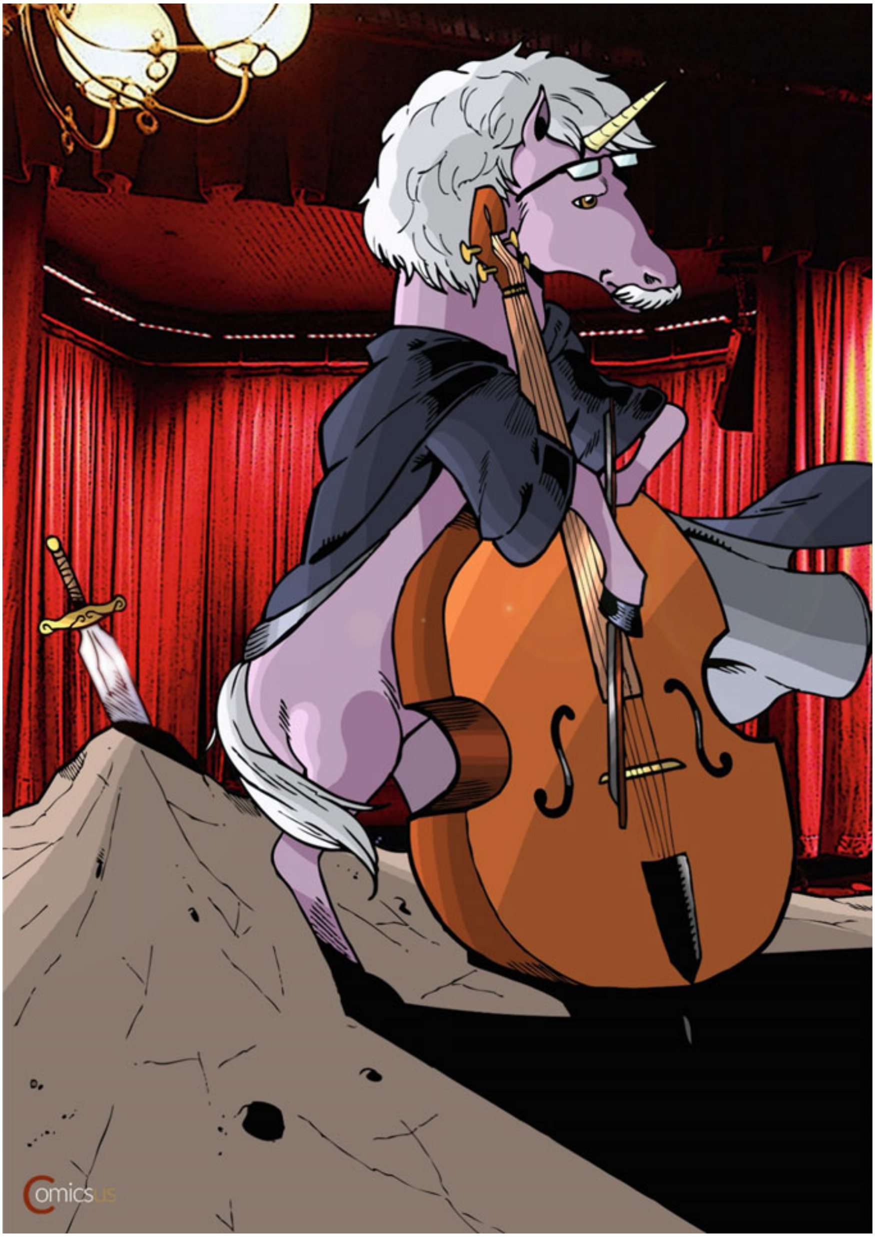
Teh, Y. W., & Jordan, M. I. (2010). Hierarchical Bayesian nonparametric models with applications. In *Cambridge Series in Statistical and Probabilistic Mathematics: Bayesian Nonparametrics* (pp. 158–207). New York, NY: Cambridge University Press.

Tetlock, P. E., & Gardner, D. (2015). *Superforecasting: The art and science of prediction* (2015th ed.). New York, NY: Broadway Books.

Wang, C., & Bier, V. M. (2013). Expert elicitation of adversary preferences using ordinal judgments. *Operations Research*, *61*(2), 372–385.

Part II
Cooke and the Classical Model

Roger Cooke has been a leader in the field of structured expert judgement since the mid-1980s. He has inspired the field and championed its application. The chapters here chart the development of and reflect on that leadership and his Classical Model. He has also led the development of two software packages: Excalibur, an implementation of the Classical Model and Unicorn, a Monte Carlo risk analysis package. He is also an outstanding jazz bass player.


Comicsus

Chapter 8
A Number of Things

Roger M. Cooke

Oration delivered in accepting the position of Professor in Applications of Decision Theory at the Faculty of Technical Mathematics and Informatics at the Delft University of Technology, Delft on November 8, 1995 by Dr. R. M. Cooke (Translated from Dutch by the author).

Esteemed Rector, and other members of the University Directorate, worthy colleagues and other members of the university community, honored guests, ladies and gentlemen:

8.1 Introduction

I hail from the United States, a country where the custom of delivering an oration is unknown. From the all no-nonsense Yankee business jargon currently emanating from Dutch university administrations, I inferred that the oration had come to resemble a Medieval morality play, somehow out of step with the times. I am a

R. M. Cooke (✉)
Resources for the Future, Washington, DC, USA
e-mail: cooke@rff.org

A. M. Hanea et al. (eds.), *Expert Judgement in Risk and Decision Analysis*,
International Series in Operations Research & Management Science 293,
https://doi.org/10.1007/978-3-030-46474-5_8

parvenu Dutchman, and people whom I judge much wiser than myself have convinced me, not without a certain impish pleasure, that I too must give an oration.

But how? A study of the genre reveals that the ideal oration opens with a quote; a quote which surprisingly explains a seemingly nondescript title by linking the Aspirant Professor's field to large themes from, preferably, Dutch history; and this all with a bombastic intellectual swagger which somehow stays entertaining. I shall try to perpetuate this tradition. A nondescript title was easily found. Long did I search for the quote. It appears that history's key figures seldom refer to mathematics. Has mathematics had nothing to say to them?

Some hope could be gleaned from Max Weber's *Die protestantische Ethik und der Geist d es Kapitalismus.* Weber uncovered a strong link between the origins of capitalism, the industrial revolution, and the emergence of Dutch Calvinism. The Calvinist doctrine of predestination, as you know, had the effect of devaluing the most important asset of the Catholic Church, real estate in the Here After. The decision who would go to Heaven and who would not had already been taken, and the Church could not intercede. Moreover, those who had been elected for salvation could not be identified by any outer or inner property. What is then the point of this short life on Earth? Our only earthly goal must be to nurture hope for an undeserved salvation. In The Netherlands, that translated to earning as much money as possible without enjoying it. There was no alternative but to apply the unspent gain to garner yet more gain, and capitalism was born, according to Weber. The hallmark of the spirit of capitalism, says Weber, is that everything, but then really *everything,* should be calculated in terms of capital. As spokesman par excellence of this new spirit, Weber cites an erstwhile compatriot. "*The good paymaster*" says Benjamin Franklin.

> is lord of another man' s purse. He that is known to pay punctually and exactly to the time he promises, may at any time, and on any occasion, raise all the money his friends can spare.[1]

Perhaps Prince William of Orange recognized in this spirit of capitalism the possibility of an alliance between Calvinist ministers and the Dutch sea pirates, from which the State of The Netherlands eventually emerged.

If such calculation does lie at the basis of the Dutch nation state and the creation of modern credit worthy man, then the large themes emerge: religion, the origin of nations, and numbers. What hidden relations bind these concepts? How is the earth divided into "We's" and "They's"? Why are there nations and Gods, why so many, and for how long? Though Franklin's quote conjures all these questions, in no way does it cover the activities of the Chair of Applications of Decision Theory. We must dig deeper.

It appears from the first European national anthem[2] that the founding of The Netherlands is intimately bound up with the gift of God to David of "a kingdom in Israel, most great."

[1] Weber (Weber 1958).

[2] The Wilhelmus van Nassouwe; see Schama, S. (1987) The Embarrassment of Riches: an Interpretation of Dutch Culture in the Golden Age, Fontana Press, London, p. 103. The Dutch often emphasized the analogy between the Israelites and their own quest for nationhood, as reflected in the eighth stanza of the Dutch National Anthem, the first European national anthem.

How was that exactly? The founding of Israel is symbolized in the founding of the Temple of King David in Jerusalem. The story is told in the Bible, First Chronicles, Chapters 20–22. In his last battle, David defeated several Ammonite cities. He led the inhabitants out and "cut them with saws and with harrows of iron and with axes," in accordance with the wishes of the Lord. Shortly thereafter, however, he listened to Satan and ordered the Israelites to be *counted.* The wrath of the Lord was immediate. David was given a choice, "either three years famine, or three months to be destroyed before thy foes, … or else three days the sword of the Lord." David chose the latter, and seventy thousand Israelites were laid low by God before David repented (he was allowed to count the dead). The angel of the Lord showed David the spot where he should build an altar to the Lord, and on that spot, the Temple of Jerusalem was built.

The roots are laid bare. Imagine, ladies and gentlemen; the pictures are familiar from the daily news. Naked children torn from their mothers' breasts, children scream, mothers plead; but the Lord is implacable and the saw teeth of the Lord chew on. For indeed, those children would have grown up worshipping a different God. David need show no remorse for this ethnic cleansing. He is unfaithful to the Lord only when he counts the number of his own people. David counted the Israelites because, like any commander, he wanted to know his military strength, but he should have known that his strength came solely from the Lord. The Lord would deliver him if he put his faith in the Lord. Trying to take his fate in his own hands was high blasphemy. Sawing the children of the enemy to pieces did not incur the Lord's displeasure.

At a technical university, we count, calculate, and measure to gain control over our fate. In my field of risk analysis, we attempt daily to frustrate the "acts of God." Is that too high blasphemy?

The founders of nations renounce existing earthly law, and appeal to incontrovertible supernatural authority. That is the way it has always been, and that is the way it is today. How does science *ultimately* relate to the fruits of such labor? This is the old question of the relation between reason and authority, between science and faith. During the Enlightenment, the ethical basis of modern constitutional democracy was negotiated by, among others, Immanuel Kant. Kant's answer came down to an armed truce between reason and faith. Each was assigned its own territory and instructed not to pester the other. Can this compromise hold its own in the face of the continual reallocation of the earth? If I believed that, I should have chosen a different subject for this oration. The problem is that the various incontrovertible authorities cannot leave *each other* alone, and if reason is kept out, then only the saws, iron harrows, and axes remain.

The question of the relation of reason and authority receives a much more radical answer in a casual aside of the Danish physicist Niels Bohr. His comment also perfectly describes what we in Applied Decision Theory try to do.

One day, Bohr visited the Danish Parliament as guest of an eminent politician. A heated discussion was under way, and his host remarked "…this is certainly quite different from the discussions at your institute, is it not Professor Bohr?" Bohr answered that discussions at his institute also became quite heated. He paused for a moment and added "…but there is one difference, at our institute we *try* to agree."[3]

"Is that all?" I hear you ask. Yes, that is *all.* Gods do not try to agree. Allah and Jehovah will never agree which incontrovertible authority is the true one. Politicians make compromises, that is, they find power equilibria. Scientists, on the other hand, *agree.* If the founding of nations is bound up with appeals to incontrovertible supernatural authority, then science is building a sort of anti-nation. Science creates a "we" which is not based on mutual recognition via a commonly recognized authority, but it is based on something else. And what is that ladies and gentlemen? Numbers. Numbers are the things on which Homo sapiens can agree. We in decision theory try to replace discussions about power and authority with discussions about numbers.

8.2 Applications of Decision Theory

Let me explain. When my daughter studied at the Royal Conservatory of Ballet, we once took a vacation in the mountains. We chanced upon a deep ravine over which a large tree had fallen. Dear daughter jumps on the tree and starts across. "If you fall off you will never dance again" advise I. "But I won't fall off" she answers indignantly. I could have appealed to my parental authority, but then I would always remain the father who forbade the tree. Instead, I applied decision theory. "Okay, go ahead if you must, but first estimate the chance that you will fall, is it one in a hundred, one in five hundred? tell me". Daughter reflects for a moment and climbs off the tree.

Once we start counting people, we do not stop. I have here a graph showing the world population from 10,000 years ago up to the present.[4] The graph begins with a population of 10 million in 8,000 BC and creeps slowly upward until the year 1650, then suddenly it shoots up. Before 1650, the world population grew at the rate of 50% per thousand years; every 1000 years, it increased by 50%. After 1650, it increases at the rate of 2000% per 1000 years (Fig. 8.1).

What explains this kink around 1650? Dutch Calvinism perhaps? Alas, I must disappoint you. On a scale of 10,000 years, there have been hundreds of Hollands, hundreds of Calvins, and hundreds of people who returned from the Dead. Yet there is only one kink. We are dealing here with the anni mirabiles between the publication of Copernicus' De Revolutionibus Orbium Caelestium in 1543 and the Philosophiae Naturalis Principia Mathematica of Newton in 1687. These are the years in which modern science and the industrial revolution were born.

What is going on? During the anni mirabiles, a unique event occurred in the West. Everywhere there was technology, the fabrication of tools, and many cultures possessed some form of science. At this time in the west, however, the two came together. The marriage between science and technology meant in the first place that scientists acquired better instruments with which they could discover natural laws.

[3] Personal anecdote of Prof. H.B.G. Casimir.

[4] Hauser (1975).

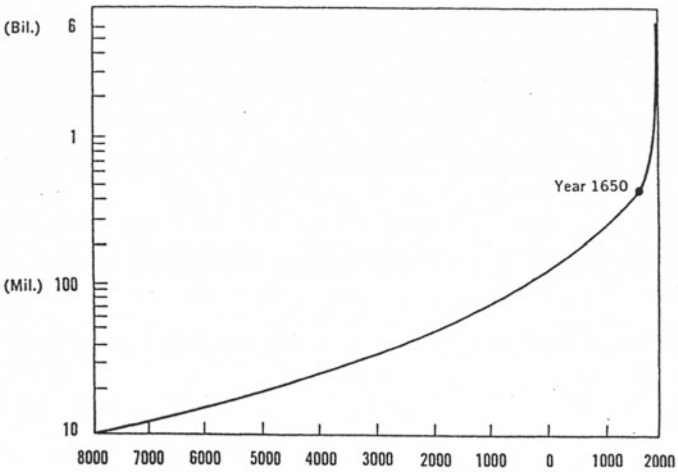


Fig. 8.1 World population from 8000 BC

Knowledge of these laws enabled them to make more accurate instruments, with which they could discover still more laws, make better measurements, etc. Better instruments served not only for better measurements. They also provided better navigation, better methods of production, better agriculture; more people could be fed with less labor. There was more free time for still more improvements, and thus 2000% per 1000 years.

The "scientization" of technology is an event, which is visible on a time scale of 10,000 years. The activities of applied decision theory are not visible on this scale, but they are visible on a scale of 30 km.

The figure below shows the lateral spread of a plume of airborne radioactive material after a hypothetical accident at a nuclear power station under stable atmospheric conditions in northern Europe. Despite intensive efforts of large research laboratories like Kernforschungzentrum Karlsruhe[5] en de National Radiological Protection Board[6]; the prediction of such a plume spread still requires a raft of uncertain parameters (Fig. 8.2).

In the 1980s, the research labs performed various "uncertainty analyses" of consequence models. The uncertainty in the input parameters was quantified, usually informally, and propagated through the models. The resulting uncertainty in model predictions can be summarized in 90% uncertainty bands. The next figure illustrates the 90% uncertainty bands for lateral plume spread under stable conditions. According to these analyses, we may be 90% certain that in a real accident under these conditions, the lateral plume spread will lie between the upper and the lower plumes (Fig. 8.3).

It will be noted that these uncertainty bands are rather narrow. The scientists are quite certain of the degree to which they can predict the plume spread. Is this degree

[5] Fischer et al. (1990).

[6] Crick et al. (1988).

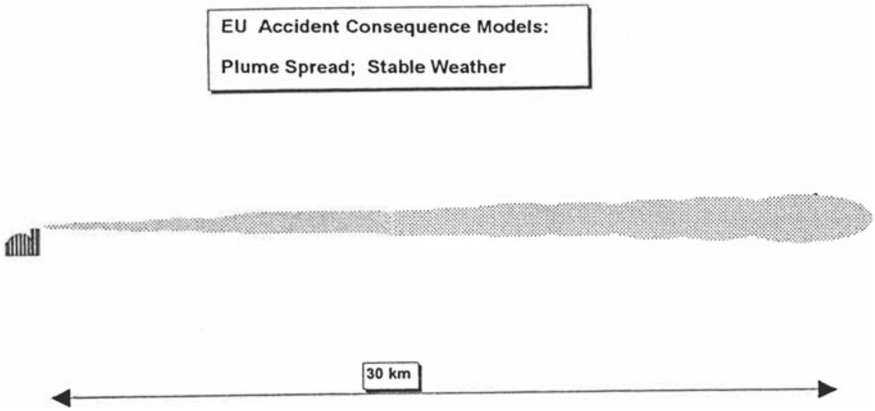


Fig. 8.2 Lateral plume spread under stable atmospheric conditions

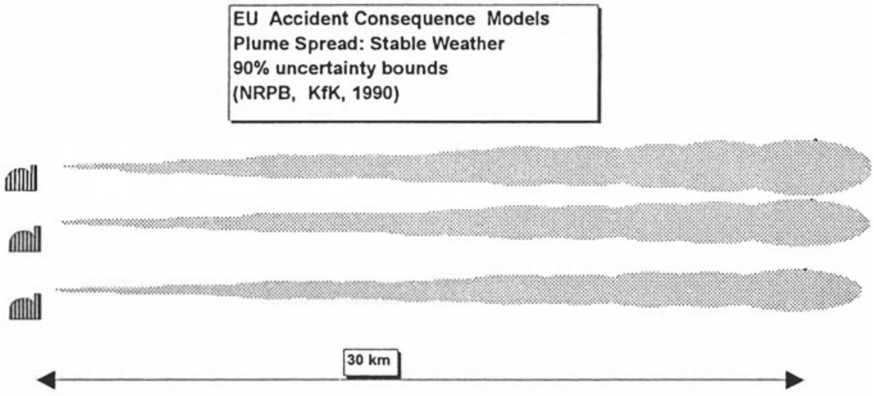


Fig. 8.3 90% uncertainty bands for lateral plume spread under stable atmospheric conditions

of certainty justified? Such questions can easily degenerate into discussions of power and authority.

In 1990, a joint research program was initiated between the European Union and the American Nuclear Regulatory Commission (USNRC). The goal was to redefine the state of the art regarding the uncertainty analysis of large-scale consequence models. In the course of this project, uncertainties for input and output variables for European and American models are being determined. A large number of European research labs participate, and overall coordination of the European effort rests with the Safety Science group in Delft. The chair of Applications of Decision Theory provides mathematical support.

The analysis of uncertainties in large risk models involves many interesting mathematical questions. One of these lends itself for illustration this afternoon. By way

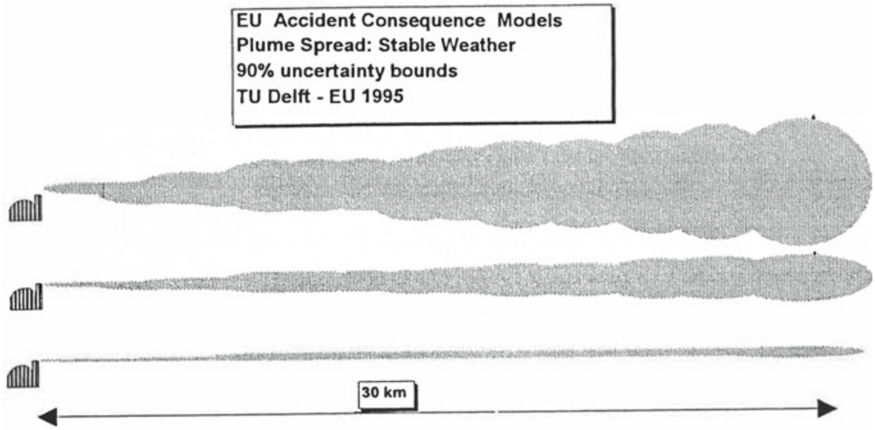


Fig. 8.4 90% uncertainty bands for lateral plume spread under stable conditions; TU Delft method

of introduction, I show our results for the uncertainty in lateral plume spread under stable conditions (Fig. 8.4).[7]

Comparing Figs. 8.3 and 8.4, it is evident that a new picture of the uncertainties has emerged. If you reflect that the seriousness of an accident is determined in large measure by the degree to which the plume does *not* spread, then you can imagine the consequences of this new picture for emergency planning.

How has this new picture emerged? Our first problem was to clarify what exactly the accident consequence models were supposed to predict. It soon became clear that the model builders themselves did not all share the same views. Should the models predict the consequences of an accident, or the consequences of an "average" or "typical" accident? A clear picture of the uncertainty in model predictions could never be attained so long *that* remained unclear—"untypical" accidents are more likely than "untypical averages."

Why was the community of model builders unclear as to what exactly their models should predict? Simply because this question had never been clearly posed. For an uncertainty analyst, this may seem incomprehensible; but I dare to assert that, for most applied mathematical models, the question "what exactly does the model predict?" never gets posed. Let this argue for a greater use of uncertainty analysis in applied mathematical modeling.

One of our first tasks was then to obtain a clear statement from the responsible authorities in Brussels what the accident consequence models should predict, an accident or an average accident. If they predict an average accident, then, we asked, over what should the average to be taken? The answer was that the models should predict the consequences of an accident and not an average accident.

[7] Cooke (1994).

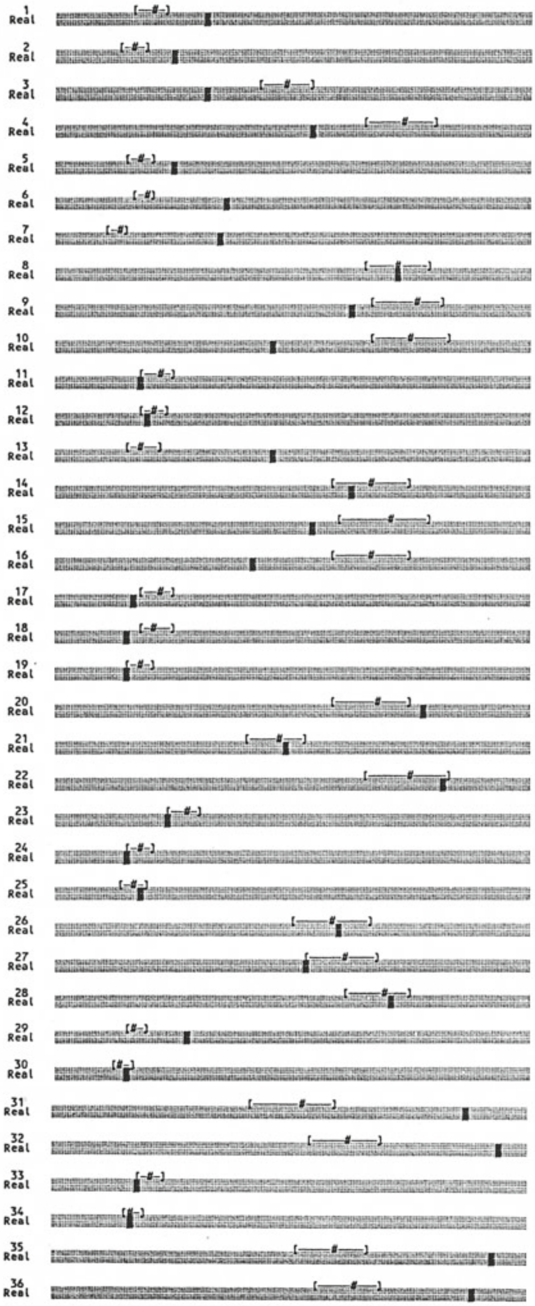


Fig. 8.5 Predictions of plume spread with uncertainty and realizations

Having that cleared that up, the following picture (Fig. 8.5) could be composed. You see here the model predictions from previous studies (as in Fig. 8.3) indicated with "#" for the lateral and vertical plume spread under various atmospheric conditions. The 90% uncertainty bands for these predictions are also shown as "[———]." A realization is given beneath each prediction; these are results of measured plume spreads in tracer experiments performed under the relevant atmospheric conditions. In this exercise, there were 36 probabilistic predictions for which realizations were available; 20 of the 36 realizations fall outside the respective uncertainty bands.

We went to work applying the "performance-based" combination of expert judgments developed in Delft.[8] The distinctive feature of this method is that uncertainty, in this case the experts' uncertainty, is treated as a scientifically measurable quantity. Different experts are asked to quantify their uncertainty with regard to results of physical measurements. The questions must be chosen so that some of the measurements are actually performed. This enables us to measure the performance of experts as probabilistic assessors and subsequently to combine their judgments so that the performance of the "combined expert," i.e., the decision maker, is optimal. This optimization involves many interesting mathematical issues, some of which I indicate in a non-technical fashion. The measurement of performance must

i. reward experts' statistical accuracy (e.g., 90% of the realizations fall within the 90% bands, in the long run).
ii. reward experts' informativeness (e.g., the 90% bands are narrow).
iii. not encourage experts to state judgments at variance with their true opinions.

The last point is of special interest for this afternoon. High measured performance entails large influence on the optimized decision maker, power if you will. The last point says that an expert maximizes his/her expected influence only by saying what he/she really thinks. He who wants power must be honest (Fig. 8.5).

The following (Fig. 8.6) shows the results of a number of probabilistic predictions of lateral and vertical plume spread. Eight international experts participated in this research, and their median estimates and 90% uncertainty bands are pictured, together with those of the optimized decision maker.

It is also interesting to compare the optimized decision maker with the "equal weight decision maker," that is, with the result of simply averaging all the experts' uncertainty distributions. The following figure (Fig. 8.7) shows the probabilistic predictions for these two decision makers for all variables for which a realization was available.

The optimized decision maker is more informative (i.e., has narrower uncertainty bands) and also has greater statistical accuracy (this last is not apparent to the naked eye, but emerges from the calculations). Of course, one data set by itself says little. Confidence in the value of this method grows as it proves itself in many different problems. This method has been applied in many problems in risk analysis, optimal maintenance, and environmental modeling. The value of performance measurement

[8] Cooke (1991).

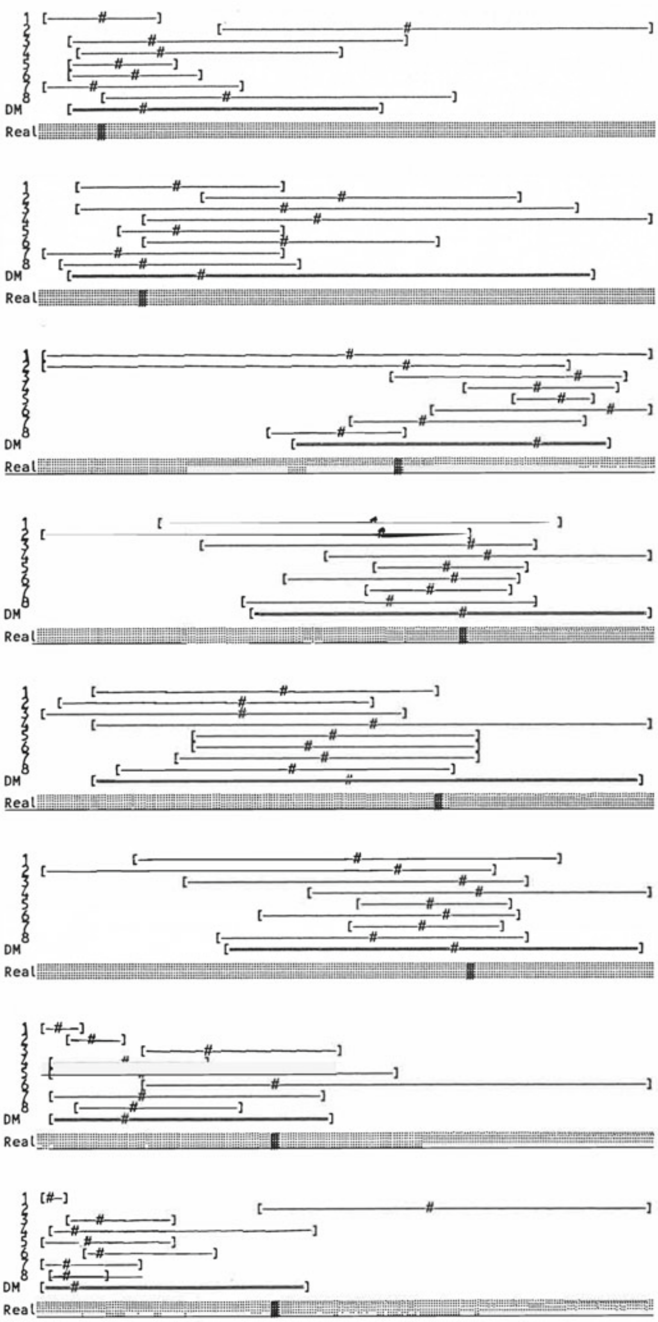
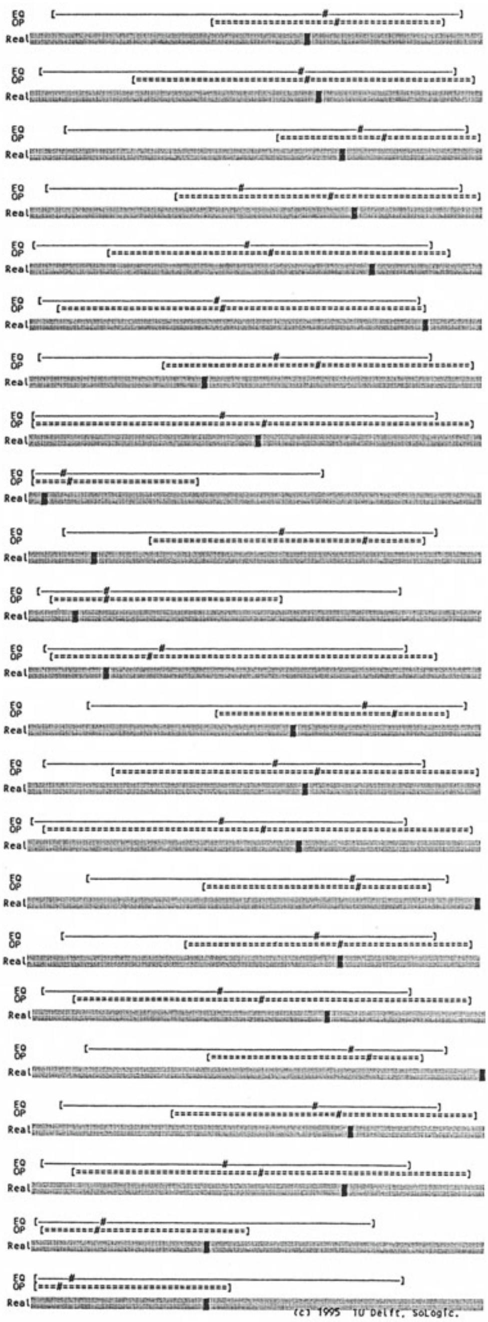


Fig. 8.6 Eight experts and performance-based decision maker of lateral and vertical plume spread; EU–TU Delft study

Fig. 8.7 Dispersion predictions for optimized and equal weight decision makers



and optimization has been proved in each case; sometimes the improvement relative to the equal weight decision maker is marginal, sometimes it is dramatic.

An example of such a dramatic improvement relative to the equal weight decision maker emerged in this research with the USNRC in regard to the dry deposition velocities of aerosols. It concerns the speed with which airborne radioactive particles deposit onto various surfaces. Of the eight international experts, the optimized decision maker opted to neglect seven of them and to go completely with one single expert. The difference in performance between this one expert and the equal weight decision maker is shown in the following figure. The median assessments of the equal weight decision maker all lie below the realizations. This would lead to significantly more optimistic predictions of the consequences of a possible accident (Fig. 8.8).

It is no exaggeration to say that our American friends had some difficulty with this outcome. As a result, they had difficulty with the phenomenon of performance-based weighting. Our European sponsors stood firm, however; and authorized us to proceed with performance-based weighting in the uncertainty analyses. They also decided to award us a contract to write a European procedures guide for uncertainty analysis of accident consequence models with expert judgment. We are hard at work on this. Our American friends have since recovered from the shock and are now fully back in the game.

Permit me one last remark on this example before I conclude. Colleagues, especially colleagues in the social sciences often wonder how world-renowned experts can be scored on performance as if they were school children. People without a background in the empirical sciences are surprised to hear that the experts actually *enjoy* this. The overwhelming majority of experts appreciate any attempt to replace discussions of power and authority with discussions of numbers, even if it concerns their own power and authority. They would all feel very much at home in Bohr's institute.

8.3 In Conclusion

Our culture still needs symbols of incontrovertible authority. A striking example of this is closer than you may realize. During a recent "professors dinner," I learned that when a professor dons his/her cap, then he/she exercises his/her official function and cannot be contradicted. By delivering this oration with my cap, I am an accomplice in this symbolism. Is that entirely consistent with the aim of replacing discussions of authority with discussions of numbers? After extended internal debate, I concluded that I could consistently wear this cap, for the following reason. Challenging symbols of incontrovertible authority do not reduce the need for such symbols. If this need emanates from fear, then such a challenge only amplifies the fear and thus intensifies the need. What is the antidote for fear? Socrates prescribed irony. After all what is more ironic than a scientist with a cap posing as incontrovertible authority? Socrates

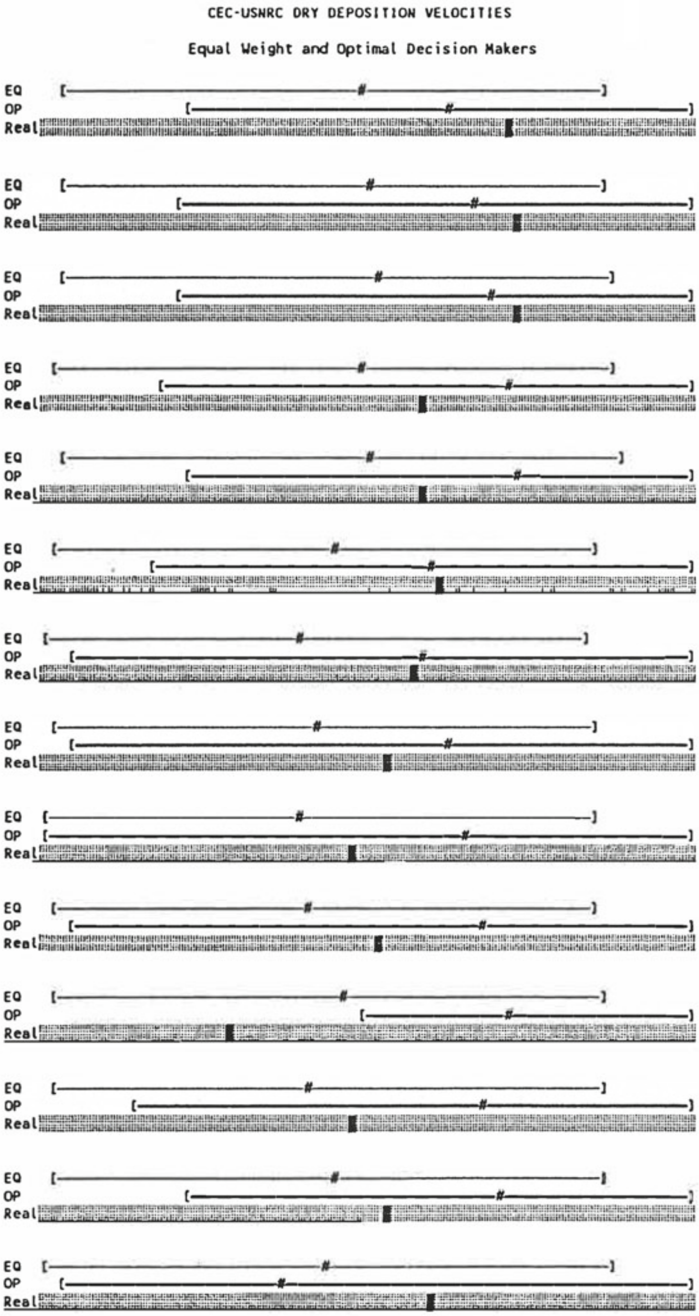
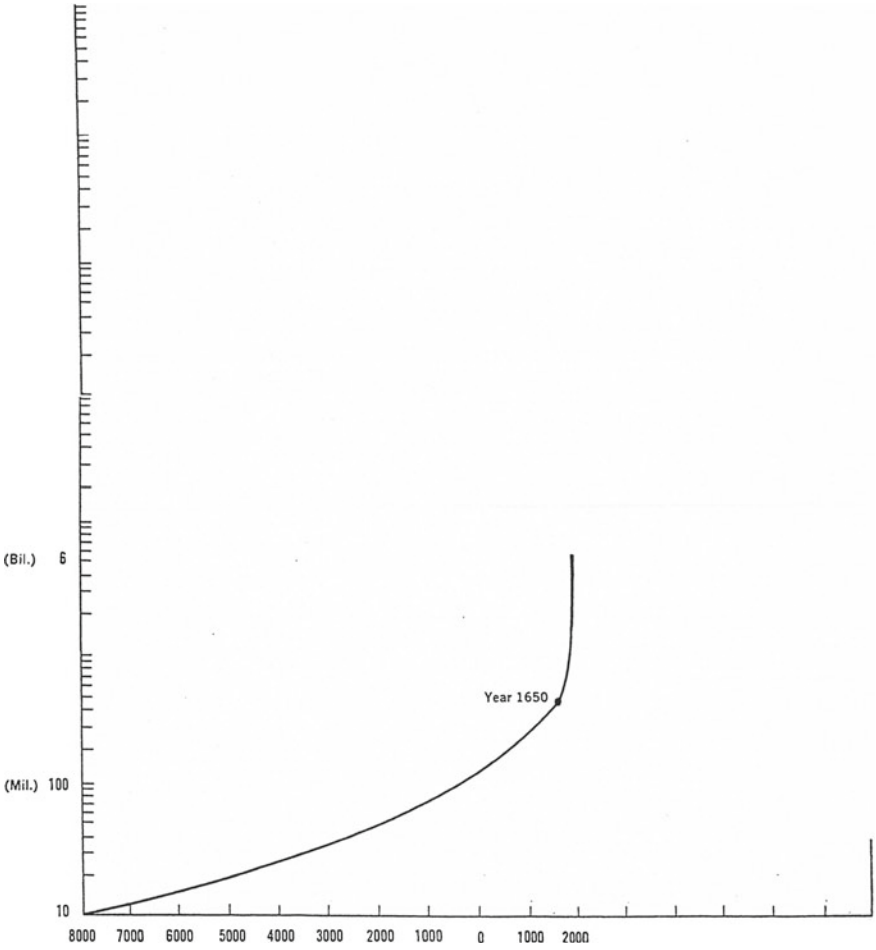


Fig. 8.8 Dry deposition predictions of the optimal and equal weight decision maker

made a distinction between irony and hypocrisy… by drinking the hemlock.[9] In the long run, however, there is only one cure for fear, and that is knowledge.

But how long is the long run? I return to the picture of the world population from 8000 BC. I have here the same picture, but now the time axis is extended out to 8000 AD. Mathematicians like to extrapolate; how should we extrapolate the world population line out to the year 8000 AD. When the population line reaches the top of the graph, then there will be one square meter of the earth's surface for each person. A little while ago, I said that the marriage of science and technology was visible on at time scale of 10,000 years. I can predict that there will be another event visible on a scale of 10,000 years. No one can say what event that will be, but I can tell you, it depends on a number of things.



[9] The Dutch expression "ergens gif op innemen" (to take poison on it) means roughly "to bet your life on it". The play on words in this context is untranslatable. Socrates' irony was in deadly earnest. After being found guilty of corrupting the youth by teaching them to question authority, he surprised his followers by refusing escape and drinking the Hemlock poison.

References

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

Cooke, R. M. (1994). Uncertainty in dispersion and deposition in accident consequence modeling assessed with performance-based expert judgment. Reliability Engineering and System Safety no. 45 35–46. (Cooke, R. M., Goossens, L. J. H., & Kraan, B. C. P. (1995). Methods for CEC/USNRC accident consequence uncertainty analysis of dispersion and deposition. EUR-15856 EN).

Crick, J. J., Hofer, E. Jones, J. A. & Haywood, S. M. (1988). Uncertainty analysis of the foodchain and atmospheric dispersion modules of MARC. National Radiological Protection Board, Report 184.

Fischer, F., Ehrhardt, J., & Hasemann, I. (1990). Uncertainty and sensitivity analyses of the complete program system UFOMOD and of selected submodels. Kernforschungzentrum Karlsruhe, Report 4627.

Hauser, P. M. (1975). World population problems. Headline Series Foreign Policy Association no. 174.

Weber, M. (1958). *The protestant ethic and the spirit of capitalism* (p. 49). New York: Scribner's Sons.

Chapter 9
The Classical Model: The Early Years

Simon French

Abstract Roger Cooke and his colleagues at the Delft University of Technology laid the foundations of the Classical Model for aggregating expert judgement in the 1980s. During 1985–1989, a research project funded by the Dutch Government saw the Classical Model developed and embedded in expert judgement procedures along with a Bayesian and a paired comparison method. That project and a subsequent working group report from the European Safety and Reliability Research and Development Association were instrumental in moving structured expert judgement procedures into the toolbox of risk analysts, particularly within Europe. As the number of applications grew, the Classical Model and its associated procedures came to dominate in applications. This chapter reflects on this early work and notes that almost all the principles and practices that underpin today's applications were established in those early years.

9.1 A Visit from the Christiaan Huygens Society

During the summer of 1986, while I was at the University of Manchester, I was contacted with a request to host a visit from the Christiaan Huygens Society of the Delft University of Technology (TU Delft). The society was and is for students of Applied Mathematics and Computer Science at the Delft University of Technology. I have no recollection of the visit itself, but for one meeting. I met Roger Cooke for the first time. To be frank, I had not heard of him. I certainly did not know of his recent work on calibration and structured expert judgement (SEJ) presented the previous year: see Cooke et al. (1988). He knew, however, that I had Bayesian interests in the topic (see, e.g. French 1985). During our brief meeting, we discussed the use of expert judgement in risk and decision analysis, a topic that at that time had a growing theoretical literature, but limited practical applications. Roger also told me

S. French (✉)
University of Warwick, Coventry CV4 7AL, UK
e-mail: simon.french.50@gmail.com

A. M. Hanea et al. (eds.), *Expert Judgement in Risk and Decision Analysis*,
International Series in Operations Research & Management Science 293,
https://doi.org/10.1007/978-3-030-46474-5_9

about the *Expert Opinions in Safety Studies* Project, which had just begun and was led by himself, Louis Goossens and Jacques van Steen.

9.2 The 'Expert Opinions in Safety Studies' Project

On July 10 1976, a catastrophic accident at a chemical plant in the Italian town of Seveso 20 km north of Milan contaminated a large area with dioxin. Within a short period, some 450 people were diagnosed with skin lesions and many animals had to be slaughtered. The accident and the issues it raised led to a revision of European approaches to technological disasters: the so-called Seveso Directives. The first of these was passed into law in June 1982, placing requirements on member governments' regulation of industrial plant in which substantial quantities of dangerous substances were used or stored. The Dutch Government recognised that careful risk assessments were needed for such industrial plants if the risks were to be controlled and that for many risks there were insufficient data for full quantitative analysis. This was the motivation for the *Expert Opinions in Safety Studies* Project, which ran for 2 years from the summer of 1986. The project was undertaken jointly by TU Delft's departments of Safety Science and Mathematics and the Netherlands Organization for Applied Scientific Research (TNO). Its reports, published in 1989, provide:

- an overview report (Goossens et al. 1989);
- a substantial literature review (van Steen and Oortman Gerlings 1989);
- A Model Description Report giving details of the Classical Model, a Bayesian model and a paired comparison model (Cooke et al. 1989, henceforth MDR);
- the Mathematical development of the Classical Model (Cooke 1989), subsequently published as Cooke (1990);
- four case studies (Cooke et al. 1989; Oortman Gerlings 1989; Stobbelaar et al. 1989; Stobbelaar and van Steen 1989).

Looking at these outputs in more detail, the literature review covered a vast body of knowledge ranging from behavioural studies of judgement through to statistical analyses of judgemental data. Methods of eliciting and quantifying uncertainty were discussed, and the review concluded that subjective probability provided the best mechanism for quantifying uncertainty, though it was mindful of the existence of heuristics and biases that would need to be addressed during elicitation. Behavioural and mathematical aggregation of judgements across a pool of experts was surveyed. Applications of SEJ in many safety studies were noted. It is also notable from the other reports though not discussed explicitly in the literature review itself that a strong philosophical perspective flowed through the project emphasising that reproducibility and accountability should be present in all SEJ studies and that there should be a balance between normative principles and empirical feasibility.

The literature review identified three approaches for aggregating the experts' judgements suitable for further investigation during the project: a Bayesian approach (Mendel and Sheridan 1989), several paired comparisons scaling methods found in the psychological literature and the model being developed at TU Delft based on weighting experts by their calibration and the informativeness of their judgements, namely the Classical Model. The Bayesian and Classical models essentially required the same data, namely that the experts should give several quantiles (e.g. 5, 50 and 95%) on a series of so-called seed and target variables. The pair comparisons methods required that each expert compared several target variables in pairs saying which was more or less likely. The methods then turned these paired comparisons into a single ranking and if the probability of one or two of these were known quantitatively, so-called reference values, then the methods could scale the target variables providing quantitative probabilities.

The three models were developed in detail in the MDR, with the mathematical development of the Classical Model in a separate report. The MDR begins with a discussion of rational consensus and then states five methodological principles which should be embodied in any SEJ study.

- *Reproducibility*: "It must be possible for scientific peers to review and if necessary reproduce all calculations. This entails that the calculational models must be fully specified and the ingredient data must be made available." (MDR, p. 8)
- *Accountability*: "The source of expert subjective probabilities must be identified." (MDR, p. 9)
- *Empirical Control*: "Expert probability assessments must in principle be susceptible to empirical control." (MDR, pp. 9–10)
- *Neutrality*: "The method for combining/evaluating expert opinion should encourage experts to state their true opinions." (MDR, pp.10–11)
- *Fairness*: "All experts must be treated equally, prior to processing the results of observations." (MDR, pp. 11–12).

There are arguments in favour of these principles in the MDR and these are developed further in Cooke (1991). The principles of empirical control, neutrality and fairness are key in developing the Classical Model. They lead to the idea of a proper group-scoring rule, and from that—after some uncomfortable mathematics—follows the weighting structure balancing calibration and informativeness.

An interesting observation in the MDR was that the team thought at the outset that the Bayesian model was "unquestionably more powerful" than the Classical Model, but that it would require many more seed variables to "warm up" and overcome the very strong ignorance assumptions built into the prior distribution.

Excalibur, the software workhorse of the Classical Model, was developed during the project, though only as software for the very cognoscente user. Another couple of years were to elapse before it was deemed safe to be released into the public domain (Cooke and Solomatine 1992). The effectiveness of the software is shown by the fact that, despite its aged interface, it is still used today (but see Leontaris and Morales-Nápoles (2018) for a recently released package written to modern standards).

The four case studies were used to evaluate both the general feasibility and acceptability of using SEJ studies in safety analyses and to evaluate the three models. The first related to gas pressure regulators at Gasunie. The company had used several different types of regulator and wanted to make a decision on future purchases.

The decision would need to take account of failure rates of the membranes and seals within the regulators. It was decided to investigate the use of several paired comparison models to assess the relative failure rates between pairs of regulators. This resulted in the experts being asked to respond to ten questions about relative rates for each pair. The study found that the experts liked the format of the elicitations: the questions were simple to understand and easy to answer. However, inconsistencies, i.e. intransitivities, were present in their responses. Moreover, paired comparison models only give a ranking of uncertainties. To provide quantitative probabilities, it was necessary to introduce known reference values. However, it was found that different paired comparison models gave different probabilities for the membrane and seal failure rates, when these reference values were introduced, raising the question of which model to use—a question that the project was unable to answer.

The second case study took place at the European Space Research and Technology Centre (ESTEC) of the European Space Agency. The problem concerned about the safety of a propulsion system. It was decided to use both the Classical Model and the Bayesian Model of Mendel and Sheridan since both use the same data structures. Although it had been expected that the identification of seed variables would be difficult, in the event some 13 were identified fairly easily. The four experts, all mathematically very competent, found the elicitation straightforward. The study made a number of conclusions relating to the elicitation, particularly in relation to the importance of having the analyst present to ensure that the experts clearly and similarly understand the questions. The Classical Model seemed to perform well with good calibration scores for the aggregated values. The Bayesian model's performance in terms of calibration and entropy was, however, "downright poor" confirming concerns raised in the MDR: in practice, it seemed to need very many seed variables to warm up and give sound probabilities on the target variables. The ESTEC member of the team endorsed the use of SEJ after the study and accepted its results.

The third case study took place at an anonymised company. It related to the microbiological reliability of a pilot plant. Failures of a fermentation process were being caused by contamination. It was believed that there were seven possible sources of this, and the problem was to assess the likelihood of each of these. It was decided to use paired comparison methods and the Classical Model to aggregate the judgements of eleven experts. In the event, the analysis was confounded with problems in the elicitation. The time allowed for this did not permit much training or discussion with the analysts. The experts objected that their responses were not anonymised. Once it was agreed to anonymise their responses, the paired comparison elicitation seemed to go without problems, but that for the Classical Model ran into extreme difficulties. The model and the requirement to give 5, 50 and 95% quantiles for the variables was explained to the expert group rather quickly because of time constraints, before they completed the forms over the following days without the analyst being present. Only two experts returned the forms and one of these was unusable. Moreover, the paired comparison results for the expert who did manage to complete the form did not correlate with his Classical Model responses. The analysis using paired responses also threw up some issues. Although the problem owner felt the output useful, internal analysis of the data gave the project team pause for thought. Among the 11 experts, 5 worked in the pilot plant and 5 were laboratory personnel. The project team felt that the pilot plant people were more expert from their closer experience of the occurrences of contamination in the plant, yet their responses showed many more inconsistencies than the laboratory staff. The main conclusion from this study was that SEJ elicitations should not be rushed. The experts need to be prepared for their task, perhaps with training in articulating probabilities and certainly with careful explanations of the elicitation questions. The analyst needs to be present and active in the elicitations to ensure that the experts understand fully what is required of them. Anonymity can also be an issue.

The fourth study took place at DSM, a chemical company, and related to irregularities with flanged connections at a particular plant. It was decided to investigate the use of paired comparisons and the Classical Model. Fourteen experts took part, with all 14 participating in the paired comparisons elicitation, but only 10 more highly educated and—so it was assumed—more numerate undertaking the Classical Model elicitation. The elicitations were conducted much more thoroughly than in the third study, using one-to-one interviews between the analyst and each expert. The interviews included an introduction to the problem and the elicitation process, careful collection of the paired comparison responses without time pressure. Then, if the expert was participating in the Classical Model elicitation, training was given in uncertainty quantification and the principles behind the model, after which the elicitation took place. The elicitations went smoothly for both paired comparisons and the Classical Model. Both the qualitative ranking of uncertainty produced by paired comparisons and the quantitative results of the Classical Model seemed sensible and meaningful. But the scaled quantitative results of the pair comparisons did not agree well with the Classical Model's results, even though the references values were taken from the output of the Classical Model. What was clear though was that discussion of the paired comparisons qualitative results did much to stimulate a consensus on the underlying issues.

I was fortunate to be able to watch the *Expert Opinions in Safety Studies* Project as a 'critical friend'. In September 1987, I was invited to TU Delft for 2 weeks to learn about progress, to comment on progress and suggest potential directions for development. Then at the end of June 1988, together with George Apostolakis, Stan Kaplan, Max Mendel and Miley Merkhofer, I attended a workshop to review the project and its conclusions. We were impressed by both the specific results and outputs and its comprehensiveness. It covered both theory and practice, building both on a clear set of principles.

It is remarkable how many aspects of SEJ practice today have their foundations in the project's outputs. While warning that expert opinions should never serve as a substitute for empirical data, if enough are available, the project clearly demonstrated that SEJ was a viable approach for obtaining probabilities for risk and decision analyses, *provided that the elicitation was conducted carefully with sufficient time for the experts to understand clearly what was required of them.* There was a recognition that SEJ studies were not cheap and that they required substantial effort and careful planning. The process needed to be structured, and this term was used regularly in discussions.

Questions were raised about the value of pair comparisons in providing a mechanism for producing quantitative probabilities that appropriately aggregated several expert's judgements. However, it was noted that discussion among the experts of the qualitative ranking was very effective in building understanding. The Bayesian method performed poorly on realistically sized sets of seed variables. The Classical Model has roots in work that began before the project (see, e.g. Cooke et al. 1988), but the method rounded off that development and presented fully for the first time. It has changed little since then. The Classical Model was used in two studies. One demonstrated that a too hurried application could lead to very poor results, but the other showed considerable promise. Over the next few years, many other studies conducted by Roger, his colleagues and his doctoral students would confirm that promise.

9.3 ESRRDA Report on Expert Judgement

As one project finished another started. In September 1988, the European Safety and Reliability Research and Development Association (ESSRDA) formed a project group to:

- "improve communication and encourage information exchange between researchers in the field of expert judgement in risk and reliability analysis, and
- "promote demonstration activities concerning expert judgement in risk and reliability analysis".

The project team brought together representatives of many research and industrial organisations: TNO, TU Delft, DSM, Shell, Leeds University,[1] KEMA, Gesellschaft Fűr Reaktorsicherheit (GRS), AEA Technology, the European joint research centre at Ispra and Université Libre de Bruxelles. The project was funded by a grant from the Commission of the European Communities and reported in 1990 (ESRRDA Project Group 1990).

The project effectively, though certainly not explicitly, took the findings from the *Expert Opinions in Safety Studies* Project and broadened them, drawing in further perspectives and experiences from other projects and organisations, particularly experiences involving the use of behavioural aggregation to enable experts to articulate an agreed consensus view after a structured and usually facilitated discussion. There was no resource for new theoretical developments or applications, but surveys were conducted of both the current literature and applications and critiques offered of both.

The project's conclusions were that there were a number of unresolved issues that needed research. On the technical side, there was a need to explore

[1] I was the representative from Leeds University.

- how to recognise which approaches were more suited to a particular problem's characteristics
- how to decompose a wider risk or decision model so that it was more suited to using probability distributions derived from expert judgement
- how to choose an appropriate group of experts
- better approaches to elicitation as more was understood about behavioural biases and calibration
- how training might help
- balancing subjective data with empirical data
- how to deal with dependencies between expert judgements in subsequent sensitivity and uncertainty analyses.

The possibility of some sort of benchmark exercises to address these was considered.

The report also discussed the issue of how to communicate the results of an SEJ study to management and problem owners. There needed to be an emphasis on providing clear explanations of:

- the problem and how it was modelled;
- why SEJ had been needed (i.e. where data were lacking);
- how the model was validated and where there were empirical controls.

There needed to be a full audit trail for all parts of the study.

9.4 The Publication of *Experts in Uncertainty*

In 1991, Roger Cooke's text *Experts in Uncertainty* was published. Early drafts of the text had been instrumental in shaping the *Expert Opinions in Safety Studies* Project and, conversely, experiences and discussions in that project honed the text. Continual references to it throughout the 28 years since then show what a milestone in SEJ this publication was; currently, it has been cited over 2000 times.[2] It established the leadership of Roger Cooke and the TU Delft group in SEJ.

The breadth and multidisciplinarity of the book are substantial. A historical survey of the use of expert opinions leads into discussions of uncertainty, its representation and the psychological understanding of it. There are chapters on subjective probability from its theoretical development by Savage through to scoring rules and calibration tools to evaluate the quality of stated subjective probabilities. Then comes the development of approaches to combining expert opinions including, of course, the development of the Classical Model. Roger's background in Mathematics *and Philosophy* shines through the book. His arguments are carefully made and there is a continual awareness of the balance between what one *should* do in making judgements with what *one* can do in terms of one's psychology.

[2] As recorded by Google scholar on the 2 June 2019.

9.5 Further Applications and Validation Studies

The publication of Roger's *Experts in Uncertainty*, in a sense, marks the end of the early development of the Classical Model. The coming years saw few developments of the model itself. The public domain version of Excalibur was released about the same time as the book appeared (Cooke and Solomatine 1992). The broad principles that guided the design of any expert judgement study had been laid down in the *Expert Opinions in Safety Studies* Project, although there were developments in the procedures for eliciting the experts' judgements. During the 1990s, the emphasis of the TU Delft group moved to applications of the method and training of students and analysts in its use. Many of the group's doctoral students moved on into academia, research centres and consultancies, spreading skills in SEJ worldwide. The group became recognised internationally as a centre—perhaps, *the* centre—for the use of SEJ studies in risk analyses. Goossens et al. (2008) describe these developments.

It is worth mentioning the emphasis which has been placed on the validation of the Classical Model over the years and which continues to the present day. On the theoretical side, there have been explorations of the model itself. For instance, theoretically, its weights involve *p*-values taken from χ^2 distributions, instabilities can arise with small samples and the number of seed variables in a calibration set is not large. So experiments were made with using alternative Kolmogorov–Smirnov alternatives. These showed no advantage (Wiper et al. 1994). All data from the groups' and many of their alumni's studies are published in a growing database (Cooke and Goossens 2007) and more recent ones are available via http://rogermcooke.net/. These studies have provided data for many validation studies. A joint EC/USNRC project (Goossens and Harper 1998, Goossens and Kelly 2000) provided a large-scale comparative evaluation of SEJ within probability safety assessment of the nuclear plant, and in a general sense picked up on many of the issues raised in the ESRRDA report. This work led to a careful summary exposition of the procedures needed to apply SEJ in practice (Cooke and Goossens 2000).

9.6 Reflections and Conclusions

Looking back on those days in the late 1980s and early 1990s, one thing stands out: so much that we do today in SEJ studies was laid down in those years. Indeed, it was probably the *Expert Opinions in Safety Studies* project that brought the term *structured expert judgement* into regular use. It was recognised then that SEJ studies were much more about the process—discussion, careful definition of terms and variables, elicitation—than about the application of a mathematical model to judgemental data.

It is true that both the project and that run subsequently by ESRRDA recognised the need for further research and indeed itemised some topics; but that research agenda was much more about honing up SEJ methods by application than the development of entirely new theory, models or processes. Broadly, all the elements that we recognise today were present then. The intervening years have seen some theoretical and mathematical developments, but mainly an increasing number of applications. Within these, the Classical Model has come to dominate whenever mathematical aggregation is required.

If I may make a brief aside: as a Bayesian, I was and still am disconcerted that the Classical Model is so successful while being non-Bayesian! At the outset, it clearly beat Mendel and Sheridan (1989)'s model. Mike Wiper and I tried to produce a Bayesian model to rival its success (see, e.g. Wiper and French 1995), but to no avail. The results were no better and the computational effort orders of magnitude greater. Maybe our latest results are more promising, see Chap. 5; but the computational effort is still far greater than that required by the Classical Model.

But that aside is a personal niggle. Much more important is that during those early years, SEJ was shown to be a practical and sensible way of proceeding when data were scarce. The Classical Model has grown to become *the* tool to aggregate experts judgements mathematically. Nowadays, SEJ is very much a part of the mainstream decision and risk analyses. The European Food Safety Authority has adopted SEJ as a standard approach (EFSA 2014) with the Classical Model as its preferred means of mathematical aggregation. It has been a privilege and pleasure to watch, and to some extent, participate in how those early efforts have borne so much fruit today.

References

Cooke, R. M. (1989). Expert opinions in safety studies (Vol. 4). A theory of weight for combining expert opinion. TU Delft: Delft, the Netherlands.

Cooke, R. M. (1990). Statistics in expert resolution: A theory of weights for combining expert opinion. In *Statistics in science* (pp. 41–72). Berlin: Springer.

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

Cooke, R., & Solomatine, D. (1992). *EXCALIBUR-integrated system for processing expert judgements*. Delft, The Netherlands: Delft University of Technology and SoLogic Delft.

Cooke, R. M., & Goossens, L. H. J. (2000). Procedures guide for structured expert judgement in accident consequence modelling. *Radiation Protection Dosimetry, 90*(3), 303–309.

Cooke, R. M., & Goossens, L. H. J. (2007). TU Delft expert judgement database. *Reliability Engineering and System Safety, 93*(5), 657–674.

Cooke, R., Mendel, M., & Thijs, W. (1988). Calibration and information in expert resolution; a classical approach. *Automatica, 24*(1), 87–93.

Cooke, R. M. van Steen, J.F.J., Stobbelaar, M. F., & Mendel M. (1989). Expert opinions in safety studies (Vol. 3). Model description report. TU Delft: Delft, The Netherlands.

EFSA. (2014). Guidance on expert knowledge elicitation in food and feed safety risk assessment. *EFSA Journal*.

ESRRDA Project Group. (1990). Expert judgment in risk and reliability analysis: experiences and perspective. Ispra, Italy: European Safety and Reliability Research and Development Association.

French, S. (1985). Group consensus probability distributions: a critical survey (with discussion). In J. M. Bernardo, M. H. DeGroot, D. V. Lindley & A. F. M. Smith (Eds.), *Bayesian statistics 2* (pp. 183–201). North-Holland.

Goossens, L., & Harper, F. (1998). Joint EC/USNRC expert judgement driven radiological protection uncertainty analysis. *Journal of Radiological Protection, 18*(4), 249.

Goossens, L. H. J., & Kelly, G. N. (2000). Special: Issue: Expert judgement and accident consequence uncertainty analysis. *Radiation Protection Dosimetry, 90*(3), 293–381.

Goossens, L., Cooke, R., & van Steen, J. (1989). Expert opinions in safety studies (Vol. 1). Final Report. TU Delft: Delft, The Netherlands.

Goossens, L. H., Cooke, R., Hale, A. R., & Rodić-Wiersma, L. (2008). Fifteen years of expert judgement at TUDelft. *Safety Science, 46*(2), 234–244.

Leontaris, G., & Morales-Nápoles, O. (2018). ANDURIL—A MATLAB toolbox for ANalysis and decisions with UnceRtaInty: Learning from expert judgments. *SoftwareX, 7,* 313–317.

Mendel, M. B., & Sheridan, T. B. (1989). Filtering information from human experts. *IEEE Transactions on Systems, Man, and Cybernetics, 19*(1), 6–16.

Oortman Gerlings, P. D. (1989). Expert opinions in safety studies (Vol. 5). Case Report 3: XYZ Case Study. TU Delft: Delft, the Netherlands.

Stobbelaar, M. F., & van Steen J. (1989). Expert opinions in safety studies (Vol. 5). Case Report 1: Gasunie Case Study. TU Delft: Delft, The Netherlands.

Stobbelaar, M. F., Cooke, R. M., & van Steen, J. (1989). Expert opinions in safety studies (Vol. 5). Case Report 2: DSM Case Study. TU Delft: Delft, The Netherlands.

van Steen, J., & Oortman Gerlings P. D. (1989). Expert opinions in safety studies (Vol. 2). Literature Survey Report. TU Delft: Delft, The Netherlands.

Wiper, M. W., & French, S. (1995). Combining experts' opinions using a normal-Wishart model. *Journal of Forecasting, 14,* 25–34.

Wiper, M. W., French, S., & Cooke, R. M. (1994). Hypothesis test based calibration scores. *The Statistician, 43,* 231–236.

Chapter 10
An In-Depth Perspective on the Classical Model

Anca M. Hanea and Gabriela F. Nane

Abstract The Classical Model (CM) or Cooke's method for performing Structured Expert Judgement (SEJ) is the best-known method that promotes expert performance evaluation when aggregating experts' assessments of uncertain quantities. Assessing experts' performance in quantifying uncertainty involves two scores in CM, the calibration score (or statistical accuracy) and the information score. The two scores combine into overall scores, which, in turn, yield weights for a performance-based aggregation of experts' opinions. The method is fairly demanding, and therefore carrying out a SEJ elicitation with CM requires careful consideration. This chapter aims to address the methodological and practical aspects of CM into a comprehensive overview of the CM elicitation process. It complements the chapter "Elicitation in the Classical Model" in the book *Elicitation* (Quigley et al. 2018). Nonetheless, we regard this chapter as a stand-alone material, hence some concepts and definitions will be repeated, for the sake of completeness.

10.1 The Classical Model: Overview and Background

Structured expert elicitation protocols have been deployed in many different areas of applications (e.g. Aspinall 2010; Cooke and Goossens 2008; Hemming et al. 2018; O'Hagan et al. 2006) and Part 4 of this book. Even though most are guided by similar methodological rules, they differ in several aspects, e.g. the way interaction between experts is handled and the way an aggregated opinion is obtained from individual experts.

A. M. Hanea (✉)
Centre of Excellence for Biosecurity Risk Analysis (CEBRA), University of Melbourne, Parkville, VIC 3010, Australia
e-mail: anca.hanea@unimelb.edu.au

Centre for Environmental and Economic Research, University of Melbourne, Melbourne, Australia

G. F. Nane
Delft Institute of Applied Mathematics, Delft University of Technology, Delft, Netherlands

A. M. Hanea et al. (eds.), *Expert Judgement in Risk and Decision Analysis*,
International Series in Operations Research & Management Science 293,
https://doi.org/10.1007/978-3-030-46474-5_10

As mentioned in the introductory chapter of this book, the two main ways in which experts' judgements are aggregated are behaviourally (by striving for consensus via facilitated discussion) and mathematically (by using a mathematical rule to combine independent individual expert estimates). Mathematical rules provide a more transparent and objective approach. A weighted linear combination of opinions is one example of such a rule. While evidence shows that equal weighting frequently performs well relative to unequal, performance-based weighting methods for reliably estimating central tendencies (e.g. Clemen and Winkler 1999), when uncertainty quantification is sought, differential weighting provides superior performance (Colson and Cooke 2017).

A widely used version of a differential weighting scheme is the Classical Model (CM) for Structured Expert Judgement (SEJ) (Cooke 1991). CM was developed and used in numerous professional applications[1] involving the quantification of various uncertainties required to aid rational decision-making. These uncertain quantities usually refer to unknown variables measured on a continuous scale. Point/"best" estimates are not sufficient when the quantification of uncertainty is the main aim, since they do not give any indication of how much the actual (unknown) values may plausibly differ from such point estimates. Expert uncertainties are thus quantified as subjective probability distributions. Experts are, however, not asked about full distributions, or parameters of distributions, but rather about a fixed and finite number of percentiles (usually three) of a distribution. From these percentiles, a minimally informative non-parametric distribution is constructed. Parametric distributions may be fitted instead, but these will add extra information to the three percentiles provided by the experts, when compared to the minimally informative non-parametric distribution. This extra information may or may not be in accordance with experts' views.

Experts are elicited individually, and face-to-face interviews were recommended in the CM's original formulation. Variants of the CM's elicitation protocol involve workshops (ranging from half a day to three days), remote elicitations or a combination of these. Each method has its advantages and disadvantages. Having all experts in one (potentially virtual) room may permit facilitated discussion prior to the actual elicitation with the aim of reducing ambiguity, providing feedback on practice questions and a better understanding of the heuristics to be avoided in order to reduce biases. However, these may come to the price of group biases, halo-effects, dominating or recalcitrant personalities, etc.

Rather than consensus, CM advances the idea of rational consensus, in which the parties (experts and facilitators) pre-commit to a scientific method for aggregating experts' assessments. CM operationalises four principles which formulate necessary conditions for achieving rational consensus (the aim of rational decision-making). These principles are detailed in the introductory chapter of this book and repeated here for convenience: *scrutability/accountability*, *empirical control*, *neutrality* and *fairness*. Cooke argues that a rational subject could accept these principles, but not necessarily accept a method implementing them. If this were the case, such a rational

[1] We call a professional application one for which the problem owner is distinct from the analyst.

subject "incurs a burden of proof to formulate additional conditions for rational consensus which the method putatively violates" (Cooke et al. 1999). Even though part of the expert judgement community does not regard CM as an appropriate method for expert judgement (Bolger and Rowe 2015a, b), to the best of our knowledge, no additional conditions for rational consensus, as proposed by Cooke, were formulated or identified as being violated. We note that there are numerous other sets of axioms proposed within the literature, see, e.g. French (1985).

The empirical control requirement is essential to the CM and, some would argue, e.g. Hanea et al. (2018), to any elicitation protocol which calls itself *structured*. It is this requirement that justifies the use of seed (calibration) variables to derive performance-based weights, providing an empirical basis for validating experts' judgements that is absent in other approaches. We note, however, that other methods, lacking empirical control, but eliciting expert judgements in a structured manner, following a rigorous protocol, are also considered SEJ protocols (ESFA 2014). "Seed" (or calibration) variables are variables taken from the problem domain for which ideally, true values become known post hoc (Aspinall 2010). However, this is rarely feasible in practice, hence variables with known realisations (values) are used instead. The questions about the seed variables that the experts need to answer are called seed questions. Experts are not expected to know the answers to these questions precisely, but they are expected to be able to capture them within informative ranges, defined by ascribing suitable values to the chosen percentiles (usually the 5th, 50th and 95th).

The theoretical background and mathematical motivation for many of the modelling choices which define the CM are detailed in Cooke (1991). However interesting and technically complete this book is, many CM neophytes find it difficult to decipher or navigate. For excellent short descriptions of the CM, written for practitioners and less technically inclined audiences, we recommend (Aspinall 2008; Quigley et al. 2018).

CM is implemented in the software Excalibur, freely available from https://lighttwist-software.com/excalibur/. Excalibur is a fully functioning application (if somewhat old) which was originally developed at Delft University of Technology and it is now maintained by Lighttwist Software.

This chapter aims to complement the existing CM descriptions, draw attention to methodological and practical aspects which were not covered in the aforementioned descriptions, update recommendations made when the CM protocol was originally designed and clarify assumptions and misconceptions. As we will emphasise throughout the chapter, some issues arise from necessary theoretical requirements, while others are reasonable pragmatic assumptions. We stress that theoretical requirements define the rigorous setting of the Classical Model, while the pragmatic assumptions allow for model flexibility that can be explored by a more experienced user.

The remainder of this chapter is organised as follows: Section 10.2 discusses several elements that need to be organised prior to the elicitation and dwells on aspects which may be problematic or are critical for a successful elicitation. Section 10.3 details some steps of the elicitation protocol, from constructing an expert's distribution from elicited percentiles to evaluating experts performance using a calibration score, an information score and a combined score. These performance measures are discussed from a theoretical, practical and intuitive view point. Section 10.4 discusses different mathematical aggregations of experts' distributions and ways to evaluate them. Section 10.5 concludes the chapter with a few remarks.

10.2 Pre-elicitation for the Classical Model

If decision-making is supported by quantitative models and the modelling is associated with uncertainties, then assessing uncertainty over the model inputs is essential. Assume a model is chosen appropriately (i.e. in accordance with needs and resources) and the sources of uncertainty are identified. Next, the modellers and analysts should collate and evaluate the available resources (e.g. data, prior studies, related literature). After completing this step, the data gaps will become apparent and the requirements for expert input can be formulated. With this, we are entering what is often called the pre-elicitation stage. Many elicitation guidelines cover this stage (e.g. ESFA 2014; Cooke et al. 1999), so in this section, we will merely complement the existing guidelines by addressing only a few, less discussed, aspects.

10.2.1 Formal Documents

Sometimes research that involves collecting subjective data from human participants needs a *human ethics approval*. Moreover, some journals require such approval to publish research informed by subjective data. Although less common in Europe and the United States,[2] this is very often a requirement in New Zealand and Australia.

A *project description* is another useful document. This will be outlining the purpose of the project, the relevant time-frames, the required expert input and potential payments. A *consent form* sometimes accompanies the project description, and it is sent to participants to formalise their agreement to take part in the elicitation and to disclose any conflict of interests.

A *briefing document* guides participants through the elicitation, including the specific way to answer questions, the reasons behind asking the questions in a particular format and the ways in which the answers are evaluated. An example of such document is Aspinall (2008).

The project description and briefing document are sometimes combined into one single document as recommended in ESFA (2014). Alternatively, a much larger document can be compiled and made available prior to the elicitation, as done in the ample SEJ study described in the Chap. 16, this volume. This document is an

[2] In some instances, it has been ruled that experts in an elicitation are not experimental subjects. If needed, human ethics only applies if the number of subjects is larger than nine, and only if the elicitation is conducted by the Federal Government (R. M. Cooke, personal communication 2018).

extended version of the briefing document, augmented with background information and available literature, especially useful to inform assessments about the target variables. However, the available literature should not contain the answers to the seed variables, as this would invalidate the calibration exercise.

10.2.2 Framing the Questions

The most common format of asking experts to quantify their uncertainty about a continuous variable is eliciting three percentiles, normally the 5th, 50th and the 95th percentiles. Eliciting five percentiles has also been used in practice (e.g. Van Elst 1997), where the 25th and 75th percentiles are elicited additionally to the three percentiles mentioned beforehand. Eliciting other percentiles or other number of percentiles (i.e. four percentiles) is nonetheless possible, posing no theoretical or practical problems. Excalibur supports formats with three, four or five elicited percentiles, which can be specified by the analyst.

However, for certain types of questions, this is easier said than done. The difficulties can arise from several reasons, and we will touch upon three of these: (1) the underlying elicited variables are not continuous, (2) the questions are not about variables that experts are familiar with, but rather they address the transformation of these variables and (3) the experts are not statistically trained. The following discussion applies to both seed and target variables. Specific seed variables issues are discussed in a dedicated subsection.

10.2.2.1 Modelling Discrete Data with Continuous Variables

Modelling discrete data with continuous random variables is not an unfamiliar practice in statistics, i.e. age of patients or months since surgery. Similarly, when eliciting bounded variables measured on a countable scale, most practitioners assume a continuous approximation of these variables and use the percentile elicitation procedure. This can be challenging for the experts. For example, assume a population of 10 healthy coral reefs. The experts are then asked about the number of future diseased coral reefs. Assume an expert's best estimate (corresponding to their median, the 50th percentile) is one. The only value strictly less than one that they can estimate as their 5th percentile is zero. However, that means that there is a one in 20 chance for the number of diseased coral reefs to be negative, which is physically impossible.

Situations like the one in the above example may lead experts to assign equal values for two or even all three percentiles, or to assign physical bounds instead of the extreme percentiles, even though they understand that in theory the percentiles of a continuous variable have to be distinct, and different than the bounds.

10.2.2.2 Unfamiliar Framing

Framing the question in a way that is different from the context experts are familiar with dramatically increases the cognitive load and should be avoided whenever possible.

For example, asking for three percentiles of variable X in relation to something normally expressed as a ratio, say 1/X, can be awkward. It is even worse if the expert thinks in terms of something which is naturally expressed as a different ratio, say, X = Y/Z.

10.2.2.3 Statistical Proficiency

The assumption of an underlying continuous distribution comes with very clear theoretical constraints, among which: the extreme (upper and lower) elicited percentiles should not equal the physical bounds of the support of the variable, and the three percentile values should be strictly increasing. Above, we touched upon a situation where these constraints may be violated because the modelled variable is not in fact continuous (but rather approximated with a continuous variable). We now want to draw attention to situations where these constraints are violated because of the difficulty of the questions, coupled with an inadequate probabilistic and statistical training of the experts.

Let us consider the example of eliciting percentages which are thought to be extreme. When experts need to estimate a very small or a very large percentage, they may assess the 5% percentiles to be 0% or the 95% percentile to be 100%. It is the analyst's job to emphasise that the elicited quantity is uncertain and to try to guide the expert through probabilistic thinking. Advising experts to reason in terms of relative frequencies may sometimes be a solution. However, if it does not help, the experts' assessments are usually slightly modified (i.e. by adding or subtracting a very small number such as 10^{-8}) to comply with the theoretical restrictions.

In certain situations, experts will assign equal values for two (or all three) percentiles even after a brief probabilistic training. If time allows, we advice that during training, an example should be used to emphasise why equal percentiles are problematic for modelling distributions of continuous random variables. To exemplify this, consider expert's assessments for an unknown variable X to be 3 for the 5th percentile, 3 for the 50th percentile and 10 for the 95th percentile. Then, the probability that the true percentage is 3 is 0.45, that is $P(X = 3) = 0.45$. Nonetheless, X is assumed to be a continuous random variable and the probability that X attains any specific value is zero, hence $P(X = 3)$ should be zero. Obviously, the expert does not acknowledge that her assessments do not correspond to a continuous random variable. And it is the analyst's job to clarify the setting. Finally, the requirement of strictly increasing percentiles has also been implemented in Excalibur.

The facilitators and analysts need to be aware of these issues when framing the questions. Sometimes, certain, possibly problematic formats cannot be avoided.

Then, the experts need to be made aware of these difficulties and, if needed, be contacted after the elicitation for re-assessment.

10.2.3 Seed Variables

The seed questions/variables are an essential element of CM, since one of the main assumptions of CM is that prior performance on seed questions is a good predictor of future performance on the target variable/questions of interest.[3] When building the differential weighted aggregated distributions, these aggregations are basically fitted to seed questions and the entire model is calibrated on them. Their importance is paramount. A strong recommendation for analysts and facilitators is to consult a couple of domain experts when looking for and formulating seed variables (see also the dry-run section below). Given their involvement with the seed questions, these experts' judgements cannot be formally elicited during the elicitation.

Seed variables and the purposes they serve are also discussed in detail in Sect. 2.3 of Quigley et al. (2018). We reiterate below the main four types of seed variables (domain-prediction, domain-retrodiction, adjacent-prediction and adjacent-retrodiction), as categorised in Cooke and Goossens (2000), and qualify their desirability.

As mentioned beforehand, the answers to seed questions should not be known by experts during the elicitation. Table 10.1 provides general guidance for selecting seed variables. Ideally, the analyst should have access to ongoing studies or domain data which become available shortly after the elicitation. These make great sources for formulating domain-prediction variables. Examples can include data from official reports which will become available shortly after the elicitation takes place. Suppose experts have been asked several questions about the percentage of unvaccinated children in Europe, in the period 2015–2018. The elicitation took place in November 2019, and the WHO official report, which is the only source for these questions, was due to appear in December 2019. Since one of the questions of interest regards the percentage of unvaccinated children in Europe in 2030, we regard the seed questions to be domain questions.

However, this not always possible, and data from recent studies within the subject matter or, less desirable, in adjacent subject matters are often the only option. Typically, data from official, yet not public, reports are used to define calibration questions. For example, existing confidential reports that document outbreaks of Salmonella in different provinces in The Netherlands could be used to define seed questions. If the questions of interest regard the number of cases of infection with Salmonella in the same provinces, then the seed questions are seen as being retrodictions and from the same domain. If, on the other hand, the question of interest regards the number of cases of infection with Salmonella at the national level, or even at the European Union level, the seed questions can be regarded as being from an adjacent

[3] From here on, we will call *questions of interest* the questions related to the target variables.

Table 10.1 Types of seed variables and their desirability. The reasonably desirable options are the ones usually used in practice

	Prediction	Retrodiction
Domain/Subject matter	Most desirable	Reasonably desirable
Adjacent/Contingent subject matter	Reasonably desirable	Last resort

subject matter. Even though the question of interest refers to the same bacteria, it is defined in a different context than the calibration question and can, therefore, be seen as from an adjacent subject matter. Another, more clear, example of using adjacent subject matter calibration questions is the following. Suppose the question of interest refers to the effects of Bonamia ostreae parasite in Ostrea chilensis oysters. Since this parasite–host combination is new, data are lacking and domain calibration questions are not possible. Calibration questions have been chosen to study the effects of different parasite–host combinations, i.e. Bonamia ostreae parasite in Ostrea edulis and Bonamia exitiosa parasite in Ostrea chilensis.

Often, elicitations need to involve two or more sub-disciplines. The set of seed questions should have then a balanced selection of items from each discipline. However, the boundaries between sub-disciplines are sometimes blurry and we are yet to learn how well can experts extrapolate their knowledge to answer questions from adjacent domains. This should be carefully dealt with prior to the elicitation and, if resources allow, consider separate panels of experts to answer different (sub-domain specific) seed questions.

Not only the domain of the seed variables is important, but also their formulation. We argue that the seed questions should be asked in exactly the same format as the questions of interest. There is no reason to believe that good performance on a certain type of task is transferable to different tasks. On the contrary, a couple of studies (Morales-Napoles et al. 2014; Werner et al. 2018) comparing the performance of experts when quantifying one-dimensional distributions using percentiles, with quantifying dependence between these one-dimensional margins, indicated a negative relationship.

Given that the domain and the formulation of the seed questions are appropriate, the next thing to consider is what sort of thinking they trigger from experts. Answering the seed questions should certainly not be a memory test about factual knowledge alone. To be able to differentiate expert performance better, the seed questions should also be as diverse as possible. Experts need to be able to make judgements of appropriate uncertainties, hence the seeds should require experts to think about composite uncertainties, in the same way they would need to do when answering the questions of interest.

The seed questions may be asked before the questions of interest and feedback may be presented to the experts before they start answering the questions of interest. Another format of the questionnaire may have all questions in random order. Some (retrospective) seed questions will be identified as such by the experts, however, the predictive ones may not stand out as seeds. An argument for having a questionnaire where seed questions and questions of interest are randomly intermixed relates to the level of experts' fatigue, as increased fatigue affects the ability of experts to concentrate towards the end of the elicitation exercise.

For continuous quantities, between eight and ten seed questions were recommended (Cooke 1991) independent of the number of questions of interest. We argue that a minimum of 15 should be used when there are not more than 35 questions of interest and at least a one day workshop. These are of course guidelines derived from experience and practice, rather than results of proper studies on experts behaviour and fatigue.

Many of the more recent studies using CM published all questions as supplementary material, but some of the older studies did not necessarily do so. As a future recommendation, aligned with the need for transparency imposed by Cooke's principles of rational consensus, we suggest all questions to be made available. Moreover, identifying and reporting the type of seed variables used, as characterised in Table 10.1, is highly recommended.

10.2.4 Dry-Run

A dry-run of the elicitation is strongly encouraged. Such an exercise is essential in decreasing the linguistic uncertainty (ambiguity), which is almost certainly present in the project description and, most importantly, in the formulation of the questions (of interest and seeds). It is also a good exercise for checking if all relevant information is captured and properly conveyed (in a language that is familiar to the experts). One or two domain experts should be asked to provide comments on all available documents, the questions, the additional information given for each question appreciated, and to estimate a reasonable time required to complete the elicitation.

10.2.5 Elicitation Format

There is no single best way to carry on an expert elicitation using CM. The original setting proposed in Cooke (1991) involves a face-to-face individual interaction between the facilitator and the expert. That is, the facilitator meets separately with each expert, trains them if necessary, discusses practice question(s) and then proceeds to guide the expert through the elicitation questions. Willy Aspinall (personal communication) carried out many of his numerous elicitations in a workshop setting. More recently, a number of elicitations have also been performed remotely, using one-to-one Skype interviews. In such cases, a teleconference with all experts may be held prior to the individual elicitation interviews. During this teleconference, the procedure, scoring and aggregation methods should be explained, and a couple of practice questions should be answered (see Chap. 16, this volume).

If the elicitation is done remotely and the seed questions are retrospective, the calibration exercise needs to be done "face-to-face" and the experts should work with the facilitator (e.g. in individual Skype sessions). The questions of interest can be then finished on a more relaxed time-frame and without the facilitator's virtual presence. However, if all the seeds are predictive, individual (remote) interviews are not a requirement.

Special attention needs to be given to experts' uncertainty training. Reasoning with uncertainty and expressing uncertainty prove to be a challenging endeavour. Practice questions are therefore desirable. Some practitioners choose practice questions from the same domain as the seed variables and questions of interest. Others choose a different subject matter, e.g. questions regarding weather, in order to focus primarily on how experts express their uncertainty.

For more details on the elicitation format, we refer to Sect. 2.4 from Quigley et al. (2018).

10.3 Elicitation with the Classical Model

The many details decided upon in the pre-elicitation stage determine the elicitation itself. These include the number and type of questions, the number and expertise of experts, the type of feedback given to and interaction permitted between experts. Once the required estimates are elicited, they are scored and the scores are used to form weights. Several weighted combinations are calculated; they form several so-called Decision Makers (DM) distributions. It is worth mentioning that a decision maker in this context represents a mathematically calculated distribution which corresponds to a virtual expert. The real decision maker would adopt one of the DMs distribution as their own.

10.3.1 From Assessments to Distributions

It is important to stress again that CM, as largely known from the literature, applies to continuous variables. That is, the elicited seed variables, as well as the variables of interest, are modelled as continuous variables and the questions are formulated in terms of percentiles of continuous distributions. Moreover, all major CM applications made use of continuous variables. As already emphasised in places, this chapter provides an in-depth perspective on the Classical Model when using continuous random variables. Eliciting discrete random variables, in terms of the probabilities of their states, and scoring the experts' performance, even though proposed in Cooke (1991), has scarce applications and has not been implemented in Excalibur.[4] It is

[4] The performance scores are calculated differently for discrete variables. Informativeness is replaced with entropy and the calibration score, even though still based on a similar test statistic, is different

also noteworthy that CM should not be used for mixed types of questions, that are both discrete and continuous. Moreover, the questions (seed and of interest) should be either all continuous or all discrete.

The rest of this chapter refers solely to eliciting continuous random variables.

It is worthwhile discussing first how expert's distribution is actually constructed from the expert's assessed percentiles within the CM. In order to specify expert's distribution, we first need to determine the support of the distribution. Assume N experts provide their assessments. Denote expert's e_i assessments for a given question as q_5^i, q_{50}^i and q_{95}^i for the 5th, 50th and 95th percentiles, respectively, and $i = 1, 2, \ldots, N$. The range $[L, U]$ is given by

$$L = \min_{1 \le i \le N} \{q_5^i, \text{realisation}\},$$
$$U = \max_{1 \le i \le N} \{q_{95}^i, \text{realisation}\},$$

for a given seed variable. Note that L denotes the minimum among all experts' lower bounds and the realisation, whereas U denotes the maximum between all experts' upper bounds and the realisation. For the questions of interest, the lower and upper bounds are determined exclusively by the experts' percentiles, i.e. $L = \min\{q_5^i\}$ and $U = \max\{q_5^i\}$, for $i = 1, \ldots, N$. The support of experts' distributions is then determined by the so-called intrinsic range

$$[L^*, U^*] = [L - k \cdot (U - L), U + k \cdot (U - L)],$$

where k denotes an overshoot and is chosen by the analyst (usually $k = 10\%$, which is also the default value in Excalibur). The intrinsic range, therefore, allows for an extension of the interval determined by the interval $[L, U]$. The extension is symmetrical for simplicity. For some questions, the intrinsic range can be specified a priori by the analyst.[5] For example, when eliciting percentages, a natural intrinsic range is [0, 100].

Each of the expert's distribution is constructed then by interpolating between expert's percentiles such that mass is assigned uniformly within the inter-percentile ranges. Consequently, by assuming a uniform background measure, the distribution of expert e_i is given by

as well, and it requires many more seed variables for reliable estimation. The interest in this topic has been revived recently with a theoretical research on calibration scores (Hanea and Nane 2019).

[5] This is however not possible in Excalibur. Unrealistic ranges obtained in Excalibur need to be truncated externally.

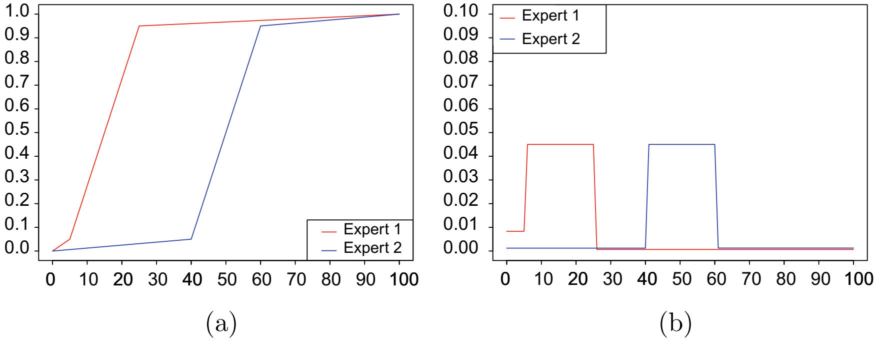


Fig. 10.1 Cumulative distribution functions (**a**) and probability distribution functions and using the intrinsic range [0,100]. (**b**) for two experts whose assessments are (5, 15, 25) (for Expert 1) and (40, 50, 60) (for Expert 2)

$$F_i(x) = \begin{cases} 0, \text{ for } x < L^* \\ \frac{0.05}{q_5^i - L^*} \cdot (x - L^*), \text{ for } L^* \leq x < q_5^i \\ \frac{0.45}{q_{50}^i - q_5^i} \cdot (x - q_5^i) + 0.05, \text{ for } q_5^i \leq x < q_{50}^i \\ \frac{0.45}{q_{95}^i - q_{50}^i} \cdot (x - q_{50}^i) + 0.5, \text{ for } q_{50}^i \leq x < q_{95}^i \\ \frac{0.05}{U^* - q_{95}^i} \cdot (x - q_{95}^i) + 0.95, \text{ for } q_{95}^i \leq x < U^* \\ 1, \text{ for } x \geq U^*. \end{cases}$$

The distribution is piecewise linear on the four intervals determined by the assessed percentiles. Note that the cumulative distribution F_i is continuous. The cumulative distribution and the corresponding density function for two experts with assessments (5, 15, 25) (Expert 1) and (40, 50, 60) (Expert 2) are depicted in Fig. 10.1. The intrinsic range has been assumed [0, 100], which can be considered appropriate as the quantities are percentages.

The above construction of distributions is arguably the most popular method of constructing distributions.

10.3.2 Measures of Performance

CM measures experts' performance as uncertainty assessors. Performance may be regarded as being determined by the properties of experts' assessments that we value positively. Three of these properties are accuracy, calibration and informativeness. Often, in the judgement and decision-making literature, accuracy is understood as the distance from the "best estimate" to the true, realised value (e.g. Einhorn et al. 1977; Larrick and Soll 2006). In the CM, the best estimate is operationalised as the median (the 50th percentile). To avoid difficulties related to estimating average accuracy across multiple seed variables, which will unavoidably be measured on different scales, the CM does *not* score accuracy as defined above. In turn, it scores calibration and informativeness.

Confusingly, from a terminological point of view (in the context outlined above), the CM calibration is also called statistical *accuracy*.[6] We recall the technical definitions of calibration and informativeness and provide an accompanying intuitive explanation.

10.3.2.1 Calibration

Assume there are N experts, $e_1, e_2, \dots, e_N$ and M seed variables/questions SQ_1, $SQ_2, \dots, SQ_M$. Denote expert's e_i assessments on question j as $q_5^{i,j}$, $q_{50}^{i,j}$ and $q_{95}^{i,j}$ for the 5th, 50th and 95th percentiles, respectively; the index j is sometimes omitted for convenience to denote the percentiles assessed for a random question (rather than for a given question j). The notation will then reduce to q_5^i, q_{50}^i and q_{95}^i. For each question and each expert, the probability range is divided into four inter-percentile intervals, corresponding to inter-percentile probability vector $p = (0.05, 0.45, 0.45, 0.05)$. Suppose the realisations of these seed questions are x_1 for $SQ_1, \dots, x_M$ for SQ_M. We may then form the sample distribution of expert e_i 's inter-percentile intervals by simply counting how many of the M realisations fall within each inter-percentile interval. Formally, let

$$s_1(e_i) = \frac{|\{k|x_k \le q_5^i\}|}{M} = \frac{\sum_{k=1}^{M} \mathbf{1}_{\{x_k \le q_5^{i,k}\}}}{M},$$

$$s_2(e_i) = \frac{|\{k|q_5^i < x_k \le q_{50}^i\}|}{M} = \frac{\sum_{k=1}^{M} \mathbf{1}_{\{q_5^i < x_k \le q_{50}^i\}}}{M},$$

$$s_3(e_i) = \frac{|\{k|q_{50}^i < x_k \le q_{95}^i\}|}{M} = \frac{\sum_{k=1}^{M} \mathbf{1}_{\{q_{50}^i < x_k \le q_{95}^i\}}}{M},$$

$$s_4(e_i) = \frac{|\{k|q_{95}^i < x_k\}|}{M} = \frac{\sum_{k=1}^{M} \mathbf{1}_{\{q_{95}^i < x_k\}}}{M}.$$

where

$$\mathbf{1}_{\{x \le a\}} = \begin{cases} 1, & \text{when } x \le a \\ 0, & \text{otherwise} \end{cases}$$

[6] The terminology was changed from calibration to statistical accuracy because of another potential terminological clash with the engineering interpretation of the term calibration.

is the indicator function.

Then $s(e_i) = (s_1(e_i), s_2(e_i), s_3(e_i), s_4(e_i))$, i.e. the empirical distribution for expert i. Note that if the expert assesses the uncertainty effectively, then we expect the distribution of the M counts to be multinomial, with parameters 0.05, 0.45, 0.45 and 0.05. Alternatively, if the realisations are indeed drawn independently from a distribution with percentiles as stated by the expert, then the quantity

$$2MI(s(e_i), p) = 2M \sum_{l=1}^{4} s_l(e_i) \ln \frac{s_l(e_i)}{p_l}, \tag{10.1}$$

is asymptotically distributed as a chi-square random variable with 3 degrees of freedom. Hence, we can score expert e_i as the statistical likelihood of the hypothesis

H_{e_i} : the inter-percentile interval containing the true value for each variable is drawn independently from probability vector p.

In Eq. (10.1), M is the number of seed questions, and $I(s(e_i), p)$ is the Kullback–Leibler divergence (Kullback and Leibler 1951), which Cooke calls the relative information of one distribution with respect to another (e.g. Cooke and Goossens 2008). The relative information score measures how one distribution, s in this case, diverges from another distribution, p here. In other words, if the experts would indeed give values which correspond to the 5th, 50th and 95th percentiles of distributions, on the long run, their sample distribution s should be equal to p. Then $I(s(e_i), p) = 0$ and this should correspond to the highest possible calibration score. As s starts diverging from p, the value of $I(s(e_i), p)$ increases, and the calibration measure should decrease, penalising the fact that the experts are not answering corresponding to the stated percentiles. A simple test for this hypothesis uses the test statistic defined by Eq. 10.1.

The p-value of this hypothesis is defined as the calibration (score or statistical accuracy)

$$Cal(e_i) = Prob\{2MI(s(e_i), p) > r | H_{e_i}\},$$

where r is the value of the expression from equation (10.1) based on the observed values[7] $x_1, \ldots, x_M$. It is the probability, under hypothesis H_{e_i}, that a deviation at least as great as r should be observed on M realisations if H_{e_i} were true.

With a finite, relatively small number of questions, often s cannot equal p. Most of the time, they differ, because of, for example, M being an odd number. An even number of seed questions does not guarantee equality either, for example for the most commonly used number of questions, ten, an expert can achieve a maximum calibration score of 0.83 when $s = (0.1, 0.4, 0.4, 0.1)$.[8] This is important when comparing calibration scores. How different should calibration scores be to conclude that one

[7] If s is equal to p, then $r = 0$ and $Cal = 1$.

[8] The minimum number of questions needed to obtain a calibration score of 1 is 20.

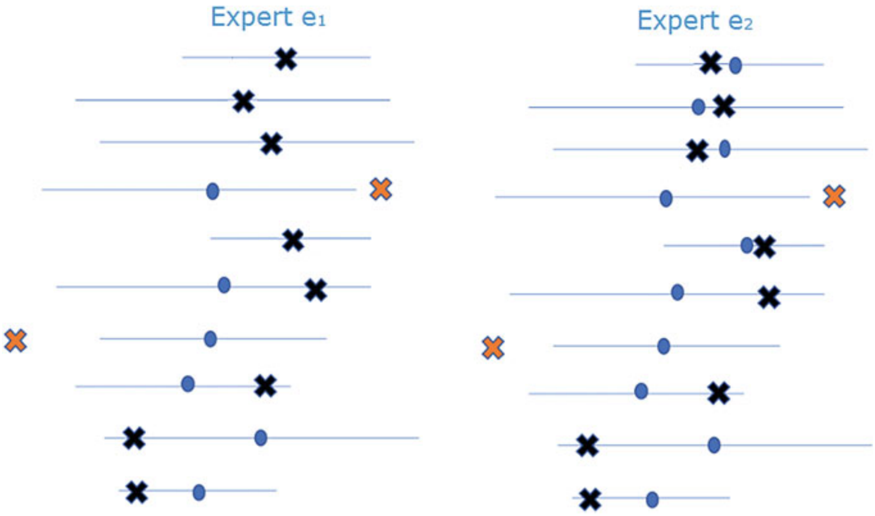
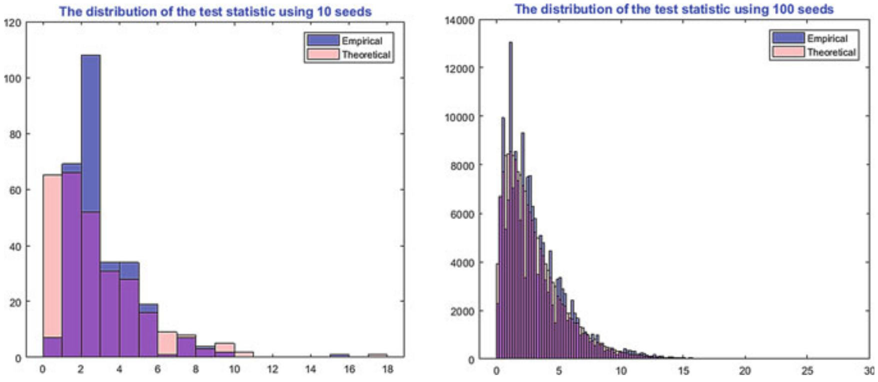


Fig. 10.2 Two experts' assessments on 10 seed questions. The starting and ending points of any line in this graph correspond to the 5th and the 95th percentiles, the blue dot corresponds to the 50th percentile and the cross corresponds to the realisation. The blue dot is not visible when it coincides with the realisation

is much better than another? The answer to this question is not straightforward. The following example illustrates an interesting situation which is slightly unrealistic, but not impossible.

On the right hand side of Fig. 10.2, Expert e_2 gave their percentiles for ten seed questions. The left and right ends of each line correspond to the 5th and the 95th percentiles, respectively. The blue dots correspond to the 50th percentiles, and the crosses correspond to the realisations of the seed variable. The crosses are blue if they are captured within the 90% credible interval, and red is they fall outside this interval. In this example, $s(e_2) = (0.1, 0.4, 0.4, 0.1)$ and expert e_2 achieves the maximum possible calibration score of 0.83. Expert e_1 gave exactly the same estimates for all the questions with the exception of four medians, which happened to coincide with the realisations of those variables (see the left-hand side of the same figure). The empirical distribution of expert e_1 is $s(e_1) = (0.1, 0.6, 0.2, 0.1)$. Expert e_1 is thus penalised as an artefact of the way the empirical distribution is constructed and achieves what seems to be a much lower calibration score of 0.39.

These sort of examples are useful to understand what these differences in calibration scores can mean. In this case, both experts are well calibrated and the 0.44 difference between calibration scores should not be used to say that expert e_2 is much better calibrated than expert e_1. However, when calibration scores are low with one of them below 0.05, the former should be considered as an indication of better performance. For example, if the empirical distribution of an expert is $s(e_3) = (0.3, 0.2, 0.2, 0.3)$, their calibration score is with approximately 0.3 less then expert e_1 calibration, mak-



(a) Ten seed variables

(b) 100 seed variables

Fig. 10.3 Histograms of $2MI(s(e_i), p)$ under null hypothesis that the inter-percentile interval containing the true value for each variable is drawn independently from probability vector p (blue), versus a random sample from a chi-square variable with 3 degrees of freedom (pink)

ing it of order 10^{-2}. Expert e_3 placed most of the mass in the tails of the distribution, which should make one confident in considering them poorly calibrated.

The discussion above about the significance level aims to stress that any calibration above a certain threshold (often chosen to be the familiar 0.05 from classical statistical testing) may be considered a good calibration, and that the calibration score should not be used to differentiate among very fine levels of calibration, but provide rather indicative levels. This is, again, similar to conducting a hypothesis testing, where one does not compare different p-values concluding that a higher p-value produces more evidence to accept the null hypothesis, but one rather compares the p-values with the significance level of, say, 0.05. Consequently, the conclusion is either enough or not enough evidence to reject the null hypothesis H_0.

Another reason for not taking the actual calibration scores and the differences between them too seriously is the asymptotic nature of the test. For ten seed variables, the distribution of the test statistic is quite far from a chi-squared distribution. This is illustrated in Fig. 10.3, where the histogram of the test statistic is determined empirically and compared with the histogram obtained by sampling from a chi-squared distributed variable. The figure on the left-hand side uses ten seed variables and the one on the right-hand side uses 100 which is of course not feasible in practice. The right-hand side histograms in Fig. 10.3 agree not only on a visual level, but also when comparing them using statistical tests. We repeatedly used the two-sample Kolmogorov–Smirnov and the two-sample Cramer–Von Mises tests, and the null hypothesis that the data in the two samples came from the same continuous distribution was not rejected in 98% of the cases.

Calibration scores are absolute scores and can be compared across studies, if these studies use the same number of seed questions. In other words, before comparing calibration scores, it is appropriate to equalise the power of the different hypothesis

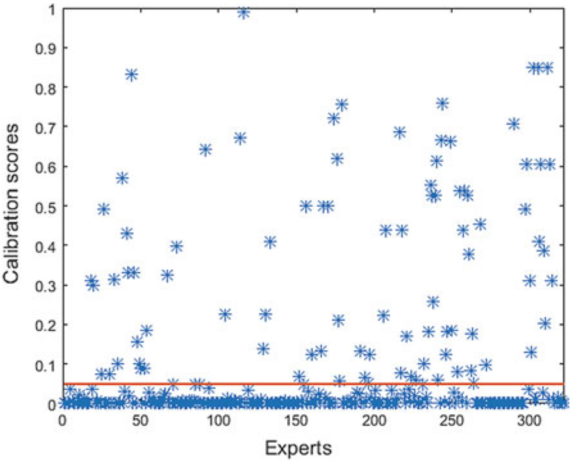


Fig. 10.4 The calibration scores of 322 experts across the pre-2006 studies available in the TU Delft dataset. The red line denotes the 0.05 significance level

tests by equalising the effective number of seed variables. Because the calibration score uses the asymptotic distribution of the $2MI(s(e_i, p)$, we adjust the power by leaving s calculated on M questions but replacing $2M$ by $2M'$, with $M' < M$, M' representing the smallest number of seed variables. In this way, we use all the M seed variables, but *pretend* that the relative information is based on M' rather than M variables. The ratio $\frac{M'}{M}$ is called the power of the calibration test (called *calibration power* in Excalibur). When the number of the seed questions increases, the calibration scores decrease, but are still distinguished if the numerical implementation of the scores is accurate enough. However, Cooke argued (Cooke 1991) that the degree to which calibration scores are distinguished should be a model parameter one can optimise for, and that reducing the power may be important in situations when all experts are very poorly calibrated. When all experts are poorly calibrated (e.g. with calibration scores of the order less than or equal to 10^{-4}, spanning three or more orders of magnitude) with one being better calibrated than the rest, all the weight may go to this one (still very) poorly calibrated expert. By reducing the power, several other combinations may be found optimal and the best of them should be used.[9] However, the accumulation of evidence since 1991 seems to suggest that in such cases an equally weighted combination of experts' distributions will be a much better choice than a combination based on optimising the calibration power.

To close our little parenthesis on the calibration power, we advise reducing the calibration power *only* for comparing calibration scores across studies with different numbers of seed questions.

To give an indication of the range of experts' calibration scores in professional applications, Fig. 10.4 presents just over 300 of experts' calibration scores extracted from the studies collected in the Delft dataset, prior to 2006. The horizontal line

[9] If you do elect to optimise weights using reduced calibration power, you should evaluate performance by introducing these weights as user weights and compare with other combinations *without* power reduction.

corresponds to the calibration score of 0.05, and it is quite clear that the majority (73%) of individual calibration scores are below this level.[10]

A completely different picture will emerge when, in Sect. 10.4.1 of this chapter, we will investigate the magnitude and spread of combinations of experts. Figure 10.9 reveals the improved performance, in terms of the calibration score, of the combination of experts.

10.3.2.2 Informativeness

Along with the calibration score, experts' assessments are evaluated with respect to an information score. The information score is intrinsically connected with determining experts' distribution, given the three percentiles specified by the expert, as constructed in Sect. 10.3.1. The information score reflects how informative expert's distribution is with respect to the background measure used to construct the distribution. If that measure is the uniform distribution, then informativeness is calculated with respect to the uniform. However, when the intrinsic range spans many orders of magnitude, the log-uniform measure is used to construct the distributions. The informativeness of such a constructed distribution is then evaluated with respect to the log-uniform background measure as well.

Both background measures are available in Excalibur and the analyst should choose between the two measures. As a rule of thumb, when the range of experts' assessments for a question spans over four orders of magnitude, then it is advised to use a log-uniform background measure[11].

The background measure is assumed, for now, to be the uniform distribution over the intrinsic range $[L^*, U^*]$

$$U(x) = \frac{x - L^*}{U^* - L^*}, \text{ for } L^* \leq x \leq U^*.$$

One can derive the probability that an uniform random variable with distribution U lies within each of the inter-percentile intervals. Experts assessments with respect to the uniform background measure for each of the four inter-percentile intervals thus yield

$$\begin{aligned}
r_1 =& U(q_5^i) - U(L^*) = \frac{q_5^i - L^*}{U^* - L^*}, \text{ for } x \in [L^*, q_5^i],\\
r_2 =& U(q_{50}^i) - U(q_5^i) = \frac{q_{50}^i - q_5^i}{U^* - L^*}, \text{ for } x \in (q_5^i, q_{50}^i],\\
r_3 =& U(q_{95}^i) - U(q_{50}^i) = \frac{q_{95}^i - q_{50}^i}{U^* - L^*}, \text{ for } x \in (q_{50}^i, q_{95}^i],
\end{aligned}$$

[10] Similar pictures presented in a slightly different format are shown in Colson and Cooke (2017).

[11] There is no theory behind the choice of the background measure. It is chosen on the basis of experiences and can later be subjected to sensitivity analysis.

$$r_4 = U(U^*) - U(q_{95}^i) = \frac{U^* - q_{95}^i}{U^* - L^*}, \text{ for } x \in (q_{95}^i, U^*].$$

With respect to expert's distribution $F(\cdot)$, let

$$\begin{aligned} f_1 &= F(q_5^i) - F(L^*) = 0.05, \\ f_2 &= F(q_{50}^i) - F(q_5^i) = 0.45, \\ f_3 &= F(q_{95}^i) - F(q_{50}^i) = 0.45, \\ f_4 &= F(U^*) - F(q_{95}^i) = 0.05, \end{aligned}$$

The information score of expert e_i for question j is then determined by

$$I_j(e_i) = \sum_{k=1}^{4} f_k \ln \frac{f_k}{r_k}.$$

Writing the information score in terms of expert's assessments and the intrinsic range gives

$$I_j(e_i) = 0.05 \ln \frac{0.05(U^* - L^*)}{q_5^i - L^*} + 0.45 \ln \frac{0.45(U^* - L^*)}{q_{50}^i - q_5^i} + 0.45 \ln \frac{0.45(U^* - L^*)}{q_{95}^i - q_{50}^i} + 0.05 \ln \frac{0.05(U^* - L^*)}{U^* - q_{95}^i},$$

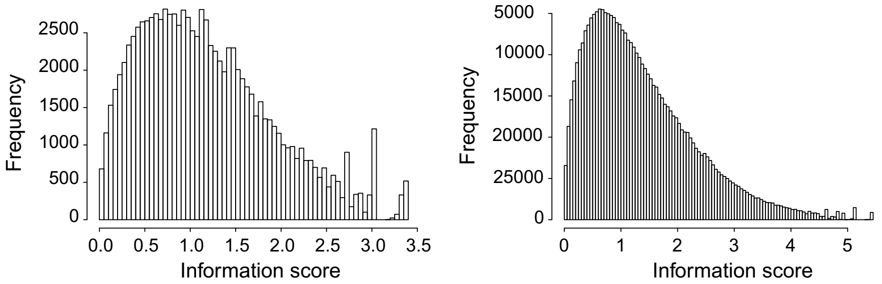
which can be re-written somewhat more compactly

$$I_j(e_i) = 0.05 \ln \frac{0.05}{q_5^i - L^*} + 0.45 \ln \frac{0.45}{q_{50}^i - q_5^i} + 0.45 \ln \frac{0.45}{q_{95}^i - q_{50}^i} + 0.05 \ln \frac{0.05}{U^* - q_{95}^i} + \ln(U^* - L^*), \tag{10.2}$$

as in Cooke (1991). The information score is a strictly positive function, which can take, in principle, arbitrarily large values. It can be observed in (10.2) that the closer expert's assessments are, the larger $I_j(e_i)$ will be. One would wonder, however, how large can the information score be, in practice, and how does the distribution of information scores looks like. We have investigated the behaviour of information scores from simulated data, as well as from expert elicitations data from previous studies.

Firstly, the simulations have been performed assuming an intrinsic range of [0, 100], as for the elicitation of percentages, and are depicted in Fig. 10.5a. Only integer values have been assumed for the experts' assessments, in order to simplify calculations. Furthermore, simulations of information scores over an intrinsic range of [0, 1000] and the histograms can be found in Fig. 10.5b.

While for an intrinsic range of [0, 100], information scores obtained are not larger than 3.5, when the intrinsic range extends to [0, 1000], the maximum observed information score is around 5.8. Repeated simulations have produced similar results for the information scores. As mentioned beforehand, the intrinsic range of [0, 100] corresponds to integer percentage assessments, whereas the intrinsic range of [0, 1000] corresponds, for example to eliciting percentages up to the first decimal.



(a) Intrinsic range of [0,100].

(b) Intrinsic range of [0,1000].

Fig. 10.5 Histograms of information scores over an intrinsic range of [0, 100] (**a**) and [0, 1000] (**b**)

The information score of an expert over all seed questions is defined as the average of information scores

$$I(e_i) = \frac{1}{M}\sum_{j=1}^{M} I_j(e_i).$$

Notice that the information score can be computed for the seed questions as well as for the questions of interest, whereas the calibration score can only be computed for the seed questions. Moreover, note that, while the calibration score of each expert is computed independently of other experts' assessments, the distribution of experts, and hence the information score depends on all experts' assessments, which makes informativeness a group dependent measure.

Finally, it should be once more emphasised that the information score reflects how informative is the expert's distribution is with respect to the background measure, which is usually assumed to be the uniform distribution. While the information score could be thought of as a measure of spread in the expert's assessments, that is, in fact, not quite the case. Consider the following examples of experts' assessments, as depicted in the Table 10.2 below.

Even though Expert 3 assessments are quite spread, the percentiles result in a skewed distribution, which is quite informative with respect to the background measure. The information score is almost the same as for Expert 2, where the probability mass function is concentrated between 40 and 60. There is a significant difference in the information score between Experts 1, 2, 3 and Expert 4. Whereas the highest information score is attained by Expert 1, the difference with Expert 2 and 3 is not that large. The cumulative distribution function and the probability density function of the 4 experts are depicted in Fig. 10.6.

The information score can now be heuristically tied with expert's distribution, namely with how discrepant expert's distribution is from the uniform distribution. For example, it is quite obvious that Expert 4 (brown) is the least discrepant from the uniform distribution (black). Similarly, Expert 1 (red) is the most discrepant and has therefore the highest information score among the 4 experts. Additionally, it is quite

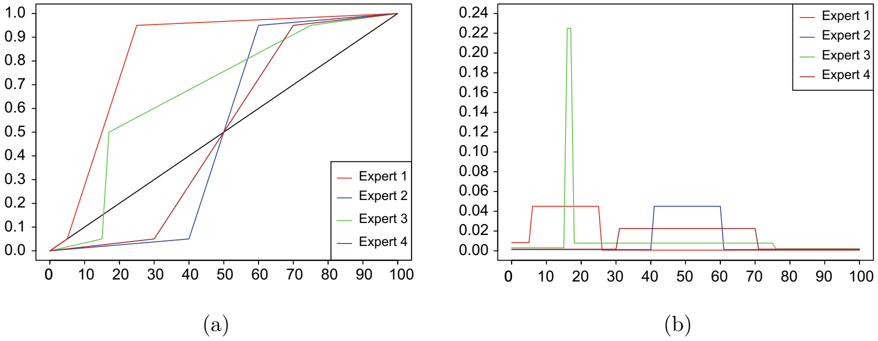


Fig. 10.6 Cumulative distribution functions (**a**) and probability density functions (**b**) for four experts whose assessments are included in Table 10.2

hard to evaluate and compare the information scores of Experts 2 (blue) and 3 (green). Their cumulative distribution functions are quite distinct, whereas the information scores are almost the same.

Obviously, the higher the information score, the more informative the expert is and an expert with high information score is preferred over an expert with a low information score, assuming they have the same calibration. One can however wonder when is an information score low, that is, when is an expert considered uninformative. Of course, an expert whose assessments coincide with the percentiles of the uniform distribution will have an information score of zero. When the assessments differ from the uniform percentiles, one could think that a test can determine whether the differences are statistically significant or not. A number of tests can quantify the difference between two distributions. Cramér-von Mises test, for example, evaluates the integrated quadratic difference between two distributions. The distributions of all four experts whose assessments are included in Table 10.2 are statistically significantly different from the uniform distribution, according to the Cramér-von Mises test, when using 100 or 1000 observations. An inspection of several examples leads to the conclusion that information scores as low as 0.15 lead to the rejection of the null hypothesis that expert's assessments come from a uniform distribution. Furthermore, an assessment of 10, 35 and 90 for the three percentiles leads to an information score of 0.1, and the p-value of the Cramér-von Mises test is 0.21. However, it should be born in mind that these results are dependent on the intrinsic range, which has been chosen [0, 100] for our example.

Another question that might arise is whether information scores are significantly different from a statistical point of view. This is nicely exemplified with the four experts' assessments above, that is, whether an information score of 1.15 is significantly higher than an information score of 0.55. Cramér-von Mises test between Expert 2 distribution (blue) and Expert 4 (brown) distribution as depicted in Fig. 10.6 leads to a p-value of 0.25, whereas the p-value for the test between Experts 3 and 4 is less than 2.2×10^{-16}. This shows that determining statistically significant differ-

Table 10.2 Example of four experts' percentage assessments

	5%	50%	95%	Information score
Expert 1	5	15	25	1.21
Expert 2	40	50	60	1.14
Expert 3	15	17	75	1.15
Expert 4	30	50	70	0.55

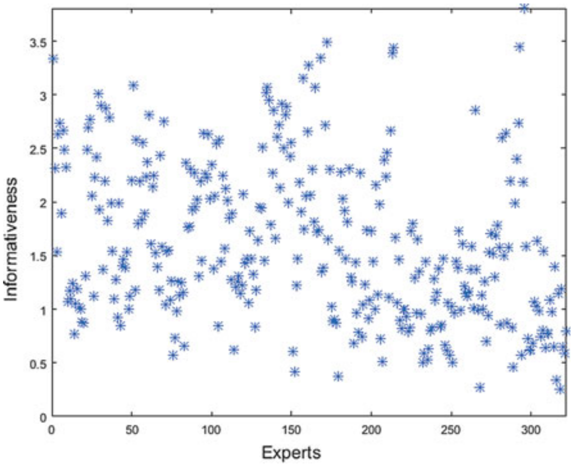


Fig. 10.7 The information scores of 322 experts across the pre-2006 studies available in the TU Delft dataset

ences between information scores is arguably an important question that cannot be answered without using more refined metrics.

To get an idea about the possible values and spread of information scores from expert elicitation data, we plotted information scores obtained by the experts taking part in the studies collected in the Delft dataset, prior to 2006. All scores are between 0.25 and 3.81 and half of these scores are larger than 1.47.

10.3.3 Combined Scores to Form Global and Item Weights

Measuring performance serves multiple purposes. Apart from differentiating between experts' performance, scores can be used to form weights which will then be used to construct a differentially weighted linear combination of distributions over the target variables. These mathematically aggregated distributions are considered to be the rational consensus distributions. They can be thought of as virtual experts whose "opinions" incorporate all experts' opinions, weighted according to their validity. An equally weighted linear combination is another virtual expert. These virtual experts can be treated as any other expert and their constructed opinions can be scored in the same way as experts' opinions. The final aim of this exercise is to find the virtual expert who performs the best. Before discussing the different virtual experts, let us return to how the scores presented in the previous sections can be combined and used as weights.

CM accounts for both calibration score and informativeness and proposes a combined score, which is the product of the calibration and the information score and it uses a cutoff level α, below which calibration scores are undesirable. The calibration score is often described as being a *fast* function, which means that its value changes quickly with the addition of every seed question and its associated response. Informativeness, on the other hand is said to be a *slow* function, which means that it is less sensitive to a small change in the number of questions. When multiplied, the calibration will dominate the value of the combined score, therefore CM values the calibration score more in comparing experts. This is also intuitively desired, as one would not prefer an informative over a poorly calibrated expert, which reflects only overconfidence. The combined score for expert i is given by

$$CS(e_i) = Cal(e_i) \cdot I(e_i) \cdot \mathbf{1}_\alpha \left(Cal(e_i)\right),$$

for $i = 1, \ldots, N$ and $\alpha \geq 0$; the weight of expert i will be proportional to their score

$$w_i = \frac{CS(e_i)}{\sum\limits_{k=1}^{N} CS(e_k)}, \tag{10.3}$$

for $i = 1, \ldots, N$. Experts with calibration scores below α will receive weight zero and their judgements will not be directly used in the final linear combination of opinions. However, all experts' assessments determine the support of all variables, therefore, all experts contribute to the virtual expert's distribution. A value α larger than zero ensures that the weights are asymptotically strictly proper. For detailed information on scoring rules, see Cooke (1991).

Note that the information score is actually calculated per question (item), and then averaged across all questions. This suggests that a combined score can be computed for each expert and seed variable

$$CS_j(e_i) = Cal(e_i) \cdot I_j(e_i) \cdot \mathbf{1}_\alpha \left(Cal(e_i)\right),$$

for $j = 1, \ldots, M$ and $i = 1, \ldots, N$. The information score $I_j(e_i)$ denotes how informative expert i is on question j. This combined score leads to the weights

$$w_i^j = \frac{CS_j(e_i)}{\sum\limits_{k=1}^{N} CS_j(e_k)},$$

for expert i and question j, where $i = 1, \dots, N$ and $j = 1, \dots, M$. The weights are called "item weights", and they are calculated per item, per expert. Thus, an expert can receive different weights for each seed variable. It should be born in mind, however, that the calibration score remains the same for each seed variable, therefore, dramatic changes in the item weights should not be expected, especially for experts with very low calibration scores. Furthermore, these weights are potentially more attractive, as they allow an expert's weight to be higher or lower for individual items/questions/variables, according to their knowledge about each question. Knowing less is usually translated into choosing percentiles further apart, and by doing that, lowering the information score for that item. The combined score for expert i is then different for each question j.

In contrast, the weights in (10.3) are referred to as global weights. For both global and item weights, calibration dominates over informativeness; the information score serves to modulate between more or less equally calibrated experts, with one exception, which will be discussed in the next section.

10.4 Post Elicitation

As mentioned in the previous section, the performance-based weights are used in CM to combine experts' judgements using a linear pool. The aggregation of expert distributions is usually referred to as a Decision Maker (DM). We reiterate that a DM in this context is a mathematically calculated distribution which corresponds to a virtual expert. The real decision maker would adopt this distribution as their own, representing rational consensus.

The performance-based weights distinguish between global and item weights, which lead to two DMs, the Global Weight Decision Maker (GWDM) and the Item Weight Decision Maker (IWDM). Moreover, different GWDM and IWDM combinations can be obtained by choosing different values for the cutoff α parameter. The α values which lead to distinct GWDM and IWDM are, in fact, the calibration scores of the experts. Using α equal to the smallest calibration score results in the combination of all experts' assessments into the DMs. Choosing the next larger calibration value translates into forming DMs using all but one expert. Choosing the largest calibration as a cutoff level translates into DMs which are the same as the best-calibrated expert. We distinguish between GWDM and optimised GWDM; GWDM uses $\alpha = 0$ (but it is essentially the same as using α equal to the smallest calibration which is usually larger than zero), therefore accounts for all experts' assessments, whereas optimised GWDM uses α such that the combined score of GWDM is maximum. Similarly, we have IWDM and optimised IWDM.

For the IWDM, the weights are different for each question, hence IWDM uses a set of weights. If GWDM uses a vector of weights, IWDM uses a matrix of weights, where each row represents the vector of weights corresponding to each question, of interest or calibration. Concluding, for GWDM, experts' weights are constructed exclusively based on the calibrations questions. IWDM uses, alternatively, weights that are constructed both on calibration questions, as well as on questions of interest. More specifically, the weights for each question of interest are computed using the calibration score and experts' information score of the question of interest.

The aggregation of expert distributions can also be done by using equal weights, which gives the equal-weight decision maker, denoted by EWDM.

Finally, it is worth mentioning that even though CM aggregates experts' distributions, other approaches are possible, such as aggregating experts' percentiles. A discussion between emerging differences in DM's distributions as well as DM's performance when aggregating distributions versus percentiles has been addressed in Colson and Cooke (2017).

10.4.1 DMs and Their Scores

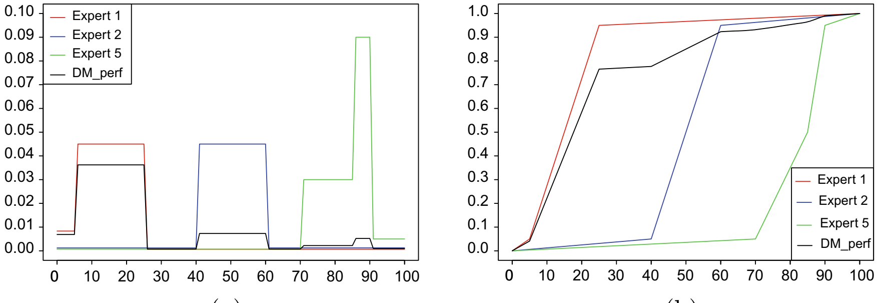
The final, and perhaps most important use of the performance-based scores is to evaluate the performance of the many DMs and be able to choose the best one, as measured by performance, which is expressed in terms of the combined score defined in Section 10.3.3. This is arguably the only valid way of motivating one choice of aggregation over others available.

DM distributions for the questions of interest are used as a final output of the elicitation study. DM can however be regarded as an expert itself, albeit virtual, and therefore one can derive its assessments also for the seed questions. These assessments can be evaluated with respect to the calibration and information score, just as for any other expert. The calibration score and informativeness of DM can be compared to single experts' performance. Moreover, both GWDM and IWDM can be optimised by choosing the value of α which maximises the combined score of the resulting DM. The combined scores of GWDM, IWDM and EWDM can be compared; the combined scores are available in Excalibur and they are a standard output of CM studies.

Excalibur also allows the users to export the DMs percentiles, which can then be used to derive the DMs distribution and plot it along with the other experts' distributions. Figure 10.8 presents the cumulative distribution functions and the density functions of three experts along with the GWDM. Expert 1 and 2's assessments can be found in Table 10.2, whereas Expert 5's assessments are 70, 85 and 90. The normalised weights are 0.8, 0.15 and 0.05, for Expert 1, Expert 2 and Expert 5, respectively.

DMs distributions can be evaluated in terms of the performance scores. The range of DMs' calibration scores in professional applications can be seen in Fig. 10.9, where the scores for EWDMs, the optimised GWDM and optimised IWDMs of 74 studies from the Delft dataset are shown[12]. The horizontal line corresponds to a

[12] There are 79 professional studies for which the DMs' scores were reported in Colson and Cooke (2017), Cooke and Goossens (2008). We were able to identify, re-run and reproduce scores for 74 of them.



(a) (b)

Fig. 10.8 Probability distribution functions (**a**) and cumulative distribution functions (**b**) of three experts along with DM

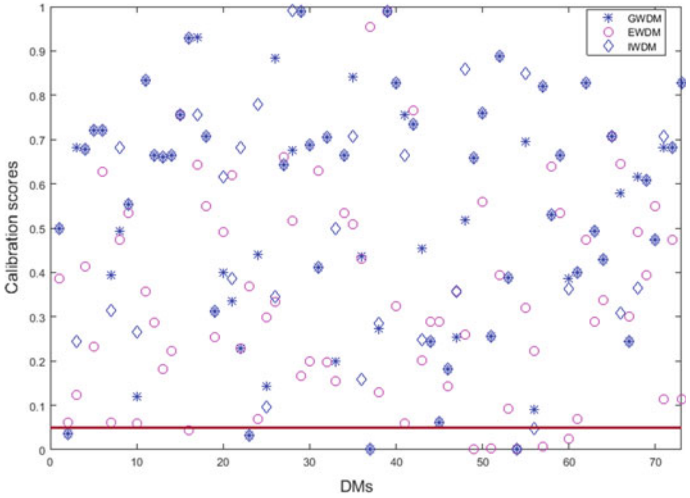


Fig. 10.9 The calibration scores of 222 DMs (74 EWDM, 74 optimised GWDM and 74 optimised IWDM) across studies available in the Delft dataset. The red line denotes the 0.05 significance level

calibration score of 0.05 and, contrary to the individual scores (see Fig. 10.4), the minority (6.7%) of DMs' calibration scores is below this level.

We consider separately the EWDMs and the GWDMs and analyse their performance. This evaluation of the performance is usually referred to as an in-sample validation. That is, the performance of DMs is evaluated on the questions that were used to determine the DMs. Figure 10.10 shows the GWDM scores on the x-axis and the EWDM scores on the y-axis. The horizontal and vertical lines indicate the 0.05 significance level, which can be regarded as a threshold for the calibration score. Very rarely one combination is below this threshold while the other is above. The main diagonal represents equal performance from the calibration view point, and again the two DMs are equally calibrated in very few cases. Given the discussion in

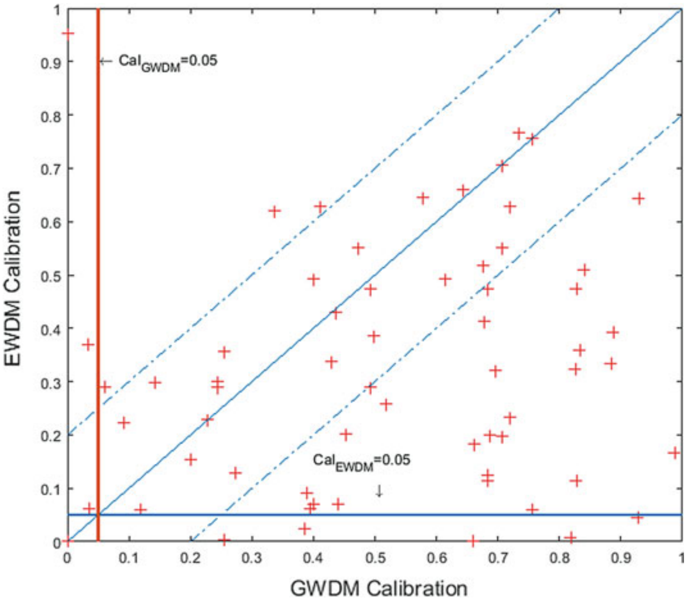


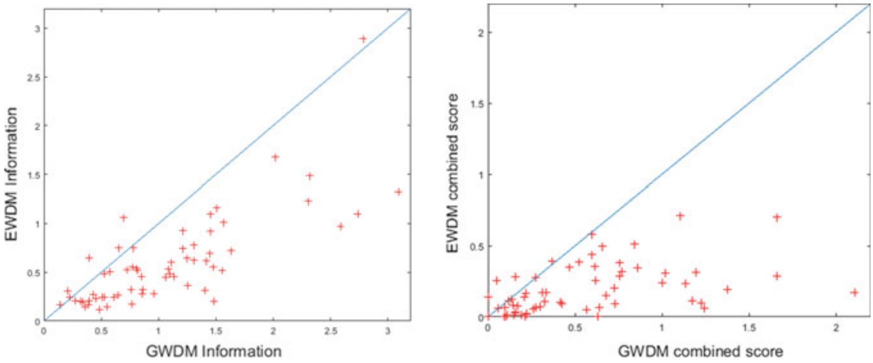
Fig. 10.10 Pairs of 63 calibration scores for optimised GWDMs versus EWDMs across the studies from the Delft dataset using at least 10 seed questions

Sect. 10.3.2.1 about small differences in the calibration scores, we may consider a region around the main diagonal, where we cannot distinguish between calibration scores (see the area bounded by dashed lines in Fig. 10.10). We consider only the studies which used at least ten seed variables (63 out of the 74 used above). It results that 41.27% of the scores fall within that region, and in 50.79% of the cases, the GWDM calibration score is clearly better than the EWDM's calibration score. In only 7.94% of the studies was the EWDM's calibration better than the GWDM's. Some would consider this as irrefutable evidence that the optimised GWDM combination is either as good or better than the EWDM.

The picture changes dramatically when we consider the information scores. These are shown in Fig. 10.11a. The vast majority of the scores are higher for the GWDM, pattern which is repeated when looking at the combined score (see Fig. 10.11b).

Item weights sometimes improve over global weights. In the same dataset of 74 professional studies (that is all studies we initially considered and not just those with more than ten seed questions), the informativeness of the IWDM is larger than the informativeness of GWDM in 57.1% of the studies, IWDMs' calibrations are only 20.6% of the times larger than that of the GWDMs. IWDMs' combined scores are larger than the PWDMs score for 41.3% of the studies.

Of course, the above analysis only serves as an in-sample validation of our intuition that performance-based combinations are at least as, or more calibrated than, and certainly more informative than the equally weighted combinations. Out of



(a) Pairs of 63 information scores for optimised GWDMs versus EWDMs across the studies from the Delft dataset using at least 10 seed questions.

(b) Pairs of 63 combined scores for optimised GWDMs versus EWDMs across the studies from the Delft dataset using at least 10 seed questions.

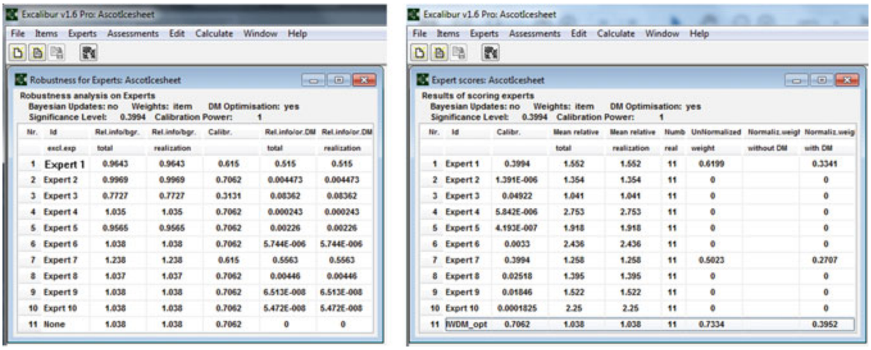
Fig. 10.11 Optimised GWDMs versus EWDMs information scores (**a**) and combined scores (**b**) across the studies from the Delft dataset using at least 10 seed questions

sample validation studies confirming the same results have been published in Colson and Cooke (2017), and the random expert hypothesis has been investigated in Chap. 3, this volume. An ultimate proof that the observed differences in scores are indeed important would be the possibility to use the different combinations in their respective decision problems and confirm that such differences in performance result in differences in decisions. Unfortunately, this does not seem to be possible. Maybe future SEJ studies should follow up with such an analysis.

10.4.2 Optimised DMs

Optimised performance-based DM's have been considered in the analysis of the professional studies in the previous subsection. Even though clarified and discussed with every opportunity, the optimisation procedure (which ensures that we are using a proper scoring rule, at least asymptotically) seems to still make analysts and young facilitators nervous, because this procedure is perceived as excluding experts (by assigning them zero weight) from the final combination of judgements.

Weight zero does ***not*** mean value zero. Most of the time, this means that those experts' knowledge was already contributed by other experts. The value of unweighted experts is seen in the robustness of the answers against the loss of experts. Excalibur has the option to perform such a robustness analysis and to recalculate the scores that would have been obtained if experts were completely excluded (rather than weighted zero) from the analysis. One of the very important contributions experts make is in determining the support of the variables. All experts contribute to these ranges and, when one expert's assessments are not taken into account, both the cali-


Excalibur v1.6 Pro: AscotIcesheet
File Items Experts Assessments Edit Calculate Window Help
Robustness for Experts: AscotIcesheet
Excalibur v1.6 Pro: AscotIcesheet
File Items Experts Assessments Edit Calculate Window Help
Expert scores: AscotIcesheet

Robustness analysis on Experts
Bayesian Updates: no Weights: item DM Optimisation: yes
Significance Level: 0.3994 Calibration Power: 1

Nr.	Id	Rel.info/bgr.	Rel.info/bgr.	Calibr.	Rel.info/or.DM	Rel.info/or.DM
	excl.exp	total	realization		total	realization
1	Expert 1	0.9643	0.9643	0.615	0.515	0.515
2	Expert 2	0.9969	0.9969	0.7062	0.004473	0.004473
3	Expert 3	0.7727	0.7727	0.3131	0.08362	0.08362
4	Expert 4	1.035	1.035	0.7062	0.000243	0.000243
5	Expert 5	0.9565	0.9565	0.7062	0.00226	0.00226
6	Expert 6	1.038	1.038	0.7062	5.744E-006	5.744E-006
7	Expert 7	1.238	1.238	0.615	0.5563	0.5563
8	Expert 8	1.037	1.037	0.7062	0.00446	0.00446
9	Expert 9	1.038	1.038	0.7062	6.513E-008	6.513E-008
10	Exprt 10	1.038	1.038	0.7062	5.472E-008	5.472E-008
11	None	1.038	1.038	0.7062	0	0

(a) Robustness analysis for experts in the ice sheet application detailed in [3].

Results of scoring experts
Bayesian Updates: no Weights: item DM Optimisation: yes
Significance Level: 0.3994 Calibration Power: 1

Nr.	Id	Calibr.	Mean relative	Mean relative	Numb	UnNormalized	Normaliz.weigt	Normaliz.weig
			total	realization	real	weight	without DM	with DM
1	Expert 1	0.3994	1.552	1.552	11	0.6199		0.3341
2	Expert 2	1.391E-006	1.354	1.354	11	0		0
3	Expert 3	0.04922	1.041	1.041	11	0		0
4	Expert 4	5.842E-006	2.753	2.753	11	0		0
5	Expert 5	4.193E-007	1.918	1.918	11	0		0
6	Expert 6	0.0033	2.436	2.436	11	0		0
7	Expert 7	0.3994	1.258	1.258	11	0.5023		0.2707
8	Expert 8	0.02518	1.395	1.395	11	0		0
9	Expert 9	0.01846	1.522	1.522	11	0		0
10	Exprt 10	0.0001825	2.25	2.25	11	0		0
11	IWDM_opt	0.7062	1.038	1.038	11	0.7334		0.3952

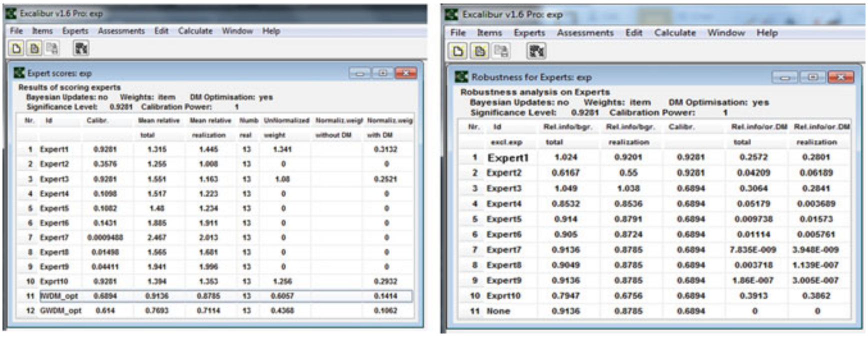
(b) Item Weights Decision Maker Optimised combination for the ice sheet application detailed in [3].

Fig. 10.12 Weight zero does ***not*** mean value zero

bration scores and the information scores of the remaining experts may change. This sometimes results in a worse calibrated DM. Below is one such example from the ice sheet application published in Nature Climate Change.

Figure 10.12a shows a snapshot from Excalibur obtained when clicking on the Robustness (experts) button. Row i corresponds to the scores that would have been obtained if Expert i was not part of the expert panel. The last row shows the scores obtained when all experts are involved. Figure 10.12b shows the optimal combination of experts when item weights are assigned. Only experts 1 and 7 are weighted in the optimal combination. However, the robustness analysis shows that if one of them is removed from the analysis, there is only a slight, irrelevant (given the number of seeds) decrease in calibration. However, if expert 3, whose weight is zero in the combination, is completely removed from the panel, the calibration drops from 0.7 to 0.3.

In the example above, the optimised IWDM (and GWDM) uses a combination of the two best-calibrated experts from the panel. In this case, as in many other cases, the optimised combination affords a higher calibration score than the two experts individually. Even though this seems intuitive, it is not always the case. Hence, there are cases when the optimised DM performs worse than the best expert. The reason behind this is the following: when the optimised DM is used, the optimisation is based on the calibrations scores alone. When there are two (or more) experts with the same best calibration, the optimised DM includes them all in the final combination, independent of the differences between their information scores. Their respective weights will be differentiated using their information scores, but this may still result in "optimal" DM whose calibration (or even combined score) is worse than the best experts' calibration. An explanation for this behaviour may be what Cooke calls a "peculiar" sort of correlation, which "has never been observed in practice" (page 197 from Cooke 1991). However, since the book was written, this phenomenon was observed in practice, even though in a different context than the one explored in


Excalibur v1.6 Pro: exp
File Items Experts Assessments Edit Calculate Window Help
Expert scores: exp
Excalibur v1.6 Pro: exp
File Items Experts Assessments Edit Calculate Window Help
Robustness for Experts: exp

Results of scoring experts
Bayesian Updates: no Weights: item DM Optimisation: yes
Significance Level: 0.9281 Calibration Power: 1

Nr.	Id	Calibr.	Mean relative	Mean relative	Numb	UnNormalized	Normaliz.weigt	Normaliz.weig
			total	realization	real	weight	without DM	with DM
1	Expert1	0.9281	1.315	1.445	13	1.341		0.3132
2	Expert2	0.3576	1.256	1.008	13	0		0
3	Expert3	0.9281	1.551	1.163	13	1.08		0.2521
4	Expert4	0.1098	1.517	1.223	13	0		0
5	Expert5	0.1082	1.48	1.234	13	0		0
6	Expert6	0.1431	1.885	1.911	13	0		0
7	Expert7	0.0009488	2.467	2.013	13	0		0
8	Expert8	0.01498	1.565	1.681	13	0		0
9	Expert9	0.04411	1.941	1.996	13	0		0
10	Exprt10	0.9281	1.394	1.363	13	1.256		0.2932
11	IWDM_opt	0.6894	0.9136	0.8785	13	0.6057		0.1414
12	GWDM_opt	0.614	0.7693	0.7114	13	0.4368		0.1062

Robustness analysis on Experts
Bayesian Updates: no Weights: item DM Optimisation: yes
Significance Level: 0.9281 Calibration Power: 1

Nr.	Id	Rel.info/bgr.	Rel.info/bgr.	Calibr.	Rel.info/or.DM	Rel.info/or.DM
	excl.exp	total	realization		total	realization
1	Expert1	1.024	0.9201	0.9281	0.2572	0.2801
2	Expert2	0.6167	0.55	0.9281	0.04209	0.06189
3	Expert3	1.049	1.038	0.6894	0.3064	0.2841
4	Expert4	0.8532	0.8536	0.6894	0.05179	0.003689
5	Expert5	0.914	0.8791	0.6894	0.009738	0.01573
6	Expert6	0.905	0.8724	0.6894	0.01114	0.005761
7	Expert7	0.9136	0.8785	0.6894	7.835E-009	3.948E-009
8	Expert8	0.9049	0.8785	0.6894	0.003718	1.139E-007
9	Expert9	0.9136	0.8785	0.6894	1.86E-007	3.005E-007
10	Exprt10	0.7947	0.6756	0.6894	0.3913	0.3862
11	None	0.9136	0.8785	0.6894	0	0

(a) Optimised Decision Makers for a recent defence application detailed in [16].

(b) Robustness analysis for experts in a recent defence application detailed in [16].

Fig. 10.13 The optimised DM is not alway optimum

Table 10.3 Correlation matrix of the three best-calibrated experts

	Expert 1	Expert 3	Expert 10
Expert 1	1	−0.07	0.55
Expert 3	−0.07	1	−0.24
Expert 10	0.55	−0.24	1

Cooke (1991). We conjecture that these situations occur when the experts' answers are correlated in a certain way; however, it is not clear yet what this "certain way" may mean. In a recent application detailed in Hemming et al. (2019), there were three experts who received the best possible calibration (0.928) score to be obtained on 13 seeds (which is the number of seeds used for this elicitation). Even though it is common for two (or even three) experts to have the same calibration score, it is rather unusual for three of the experts to have the same best calibration score Fig. fig:defencespsrobDM.

Returning to the ice sheet example, we note that the combination of the three best experts (experts 1, 3, and 10) leads to poorer performance for both the GWDM and the IWDM. However, taking one of the best-calibrated experts out of the combination restores the score of the DMs to equal that of the best-calibrated experts. This is true only when we take expert 1 out of the analysis, as shown in Fig. 10.13b. The dependence structure between these three experts in Table 10.3 is depicted.

Expert 1's assessments seem to be positively correlated with those of expert 10 and uncorrelated with those of expert 3. The two experts whose combination would be better calibrated seem to be slightly negatively correlated (even though on 13 samples this correlation is not significantly different than zero). The correlation values were calculated based on the medians of the experts rather than all three quantiles, in a similar way to the calculations performed in other studies that investigated dependence between experts' assessments (see Kallen and Cooke 2002; Wilson and Farrow 2018).

There is an unequivocal need for more research into these issues and more awareness of the possibilities.

10.5 Closing Remarks

This chapter draws attention to some (maybe less discussed) aspects of the theoretical background of CM. One of these aspects is the misinterpretation of the differences between calibration scores. Another one regards the intuitive relation between the wideness of the uncertainty bounds and the information score. The chapter also aims to provide a thorough overview of practical aspects and choices that practitioners face before and during the elicitation process.

"The qualifier *structured* means that expert judgement is treated as scientific data, albeit scientific data of a new type" (Cooke 1991). The name of the method itself the "Classical Model" emphasises the close connections with classical statistics. Furthermore, the method has auspiciously laid grounds for further statistical endeavours, such as goodness of fit and validation. If one regards the DM's performance as a goodness of fit measure, then the optimised DM's distributions are constructed such that they best fit the expert data. The evaluation of the performance-based DM has also been referred to as an in-sample validation. Furthermore, notable effort has been undertaken (Colson and Cooke 2017) to validate CM out-of-sample. The scores of performance-based DM's are hence evaluated on questions that have not been used to construct DM's distributions.

Despite the demanding nature of CM, the results from the studies show that the effort of forming and using performance-based combination of experts' distributions is definitely worthwhile.

References

Aspinall, W. (2008). Expert judgment elicitation using the classical model and excalibur. *Briefing Notes*.

Aspinall, W. (2010). A route to more tractable expert advice. *Nature*, *463*, 294–295.

Aspinall, W., & Bamber, J. L. (2013). An expert judgment assessment of future sea level rise from the ice sheets. *Nature Climate Change*, *3*, 424.

EFSA. (2014). Guidance on expert knowledge elicitation in food and feed safety risk assessment. *EFSA Journal*, *12*(6). (Parma, Italy).

Bolger, F., & Rowe, G. (2015a). The aggregation of expert judgment: Do good things come to those who weight? *Risk Analysis*, *35*, 5–11.

Bolger, F., & Rowe, G. (2015b). There is data, and then there is data: Only experimental evidence will determine the utility of differential weighting of expert judgment. *Risk Analysis*, *35*, 21–26.

Clemen, R., & Winkler, R. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis*, *19*, 187–203.

Cooke, R. M. (1991). *Experts in uncertainty: Opinion and subjective probability in science*. Environmental ethics and science policy series: Oxford University Press.

Colson, A. R., & Cooke, R. M. (2017). Cross validation for the classical model of structured expert judgment. *Reliability Engineering and System Safety*, *163*, 109–120.

Cooke, R. M., & Goossens, L. H. J. (2000). Procedures guide for structural expert judgment in accident consequence modelling. *Radiation Protection Dosimetry*, *90*(3), 303–309.

Cooke, R. M., & Goossens, L. H. J. (2008). TU Delft expert judgment data base. *Reliability Engineering and System Safety*, *93*(5), 657–674.

Cooke, R., Kraan, B., & Goossens, L. (1999). Rational consensus under uncertainty: Expert judgment in the EC-USNRC uncertainty study. In K. Andersson (Ed.), *NEI-SE-308*. Sweden.

Einhorn, H. J., Hogarth, R. M., & Klempner, E. (1977). Quality of group judgment. *Psychological Bulletin*, *84*(1), 158.

French, S. (1985). Group consensus probability distributions: A critical survey. In J. M. Bernardo, M. H. De Groot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics* (pp 182–201). Elsevier North Hollan.

Hanea, Anca M., McBride, Marissa, Burgman, Mark, & Wintle, Bonnie. (2018). The value of performance weights and discussion in aggregated expert judgments. *38*, 03.

Hanea, A. M., & Nane, G. F. (2019). Calibrating experts' probabilistic assessments for improved probabilistic predictions. *Safety Science*, *118*, 763–771.

Hemming, V., Walshe, T. V., Hanea, A. M., Fidler, F., & Burgman, M. A. (2018). Eliciting improved quantitative judgments using the idea protocol: A case study in natural resource management. *PLOS ONE*, *13*(6), 1–34.

Hemming, V., Hanea, A. M., Armstrong. N., & Burgman, M. A. (2019). Improving expert forecasts in reliability. Application and evidence for structured elicitation protocols. *Quality and Reliability Engineering International*. (Accepted in September, 2019).

Hoffmann, S., Devleesschauwer, B., Aspinall, W., Cooke, R., Corrigan, T., et al. (2017). Attribution of global foodborne disease to specific foods: Findings from a world health organization structured expert elicitation. *PLOS ONE*, *12*(9), 1–26.

Kallen, M. J., & Cooke, R. M. (2002). Expert aggregation with dependence. In *Proceedings of the 6th International Conference on Probability Safety and Management* (pp. 1287–1294).

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*(1), 79–86.

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, *52*(1), 11–27.

Morales-Napoles, O., Hanea, A. M., & Worm, D. T. H. (2014). Experimental results about the assessments of conditional rank correlations by experts: Example with air pollution estimates. In R. D. J. M. Steenbergen, P. H. A. J. M. van Gelder, S. Miraglia, & A. C. W. M. Vrouwenvelde (Eds.), *Safety, Reliability and Risk Analysis: Beyond the Horizon* (pp. 1359–1366). CRC Press.

Van Elst, N.P. (1997). Betrouwbaarheid beweegbare waterkeringen [reliability of movable water barriers]. In WBBM report Series 35. Delft University Press.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., et al. (2006). *Uncertain judgments: Eliciting experts' probabilities*. London: Wiley.

Quigley, J., Colson, A., Aspinall, W., & Cooke, R. M. (2018). Elicitation in the classical model. In L. C. Dias, A. Morton, & J. Quigley (Eds.), *Elicitation: The science and art of structuring judgment* (pp. 15–36). International series in operations research & management science Cham: Springer.

Werner, C., Hanea, A. M., & Morales-Napoles, O. (2018). Eliciting multivariate uncertainty from experts: Considerations and approaches along the expert judgement process. In L. C. Dias, A. Morton, & J. Quigley (Eds.), *Elicitation: The science and art of structuring judgment* (pp. 171–210). International series in operations research & management science Cham: Springer.

Wilson, K. J., & Farrow, M. (2018). Combining judgments from correlated experts. In L. Dias, A. Morton, & J. Quigley (Eds.), *Towards a general theory of expertise: Prospects and limits* (Vol. 261). Cham: Springer.

Chapter 11
Building on Foundations: An Interview with Roger Cooke

Gabriela F. Nane and Anca M. Hanea

1. **To start off, which applications of Structured Expert Judgment (SEJ) were most important for disseminating the Classical Model (CM), in your opinion?**

The applications of which I am aware are summarized on my webpage http://rogermcooke.net/, so I won't give separate references. Christian Preyssl got us started with applications at ESTEC. Certainly, the nuclear work in the 1990s was most influential in working out dependence elicitation and probabilistic inversion. The EU Procedures Guide (Cooke and Goossens 2000) emerged from that work and standardized the methods. There was a complex mating ritual between the European and American teams, but it all worked out in the end. I was not involved in Willy Aspinall's volcano work, but that certainly was very fecund. The fine particulate study led by John Evans was a very important foray into public health applications. Willy led several subsequent applications in this area. The recent ice sheet applications are very important in bringing SEJ uncertainty quantification to the climate discussion.

2. **You originally studied Philosophy of Science at Yale—can you tell us a little bit about that, and about how that study influenced the later development of your thinking? For a Philosopher, you have done a lot of work on real-life problems.**

I started in Philosophy of Physics. Not only was that program weak at Yale, but I struggled with the basics. Put a glass of water on a table. There are two invisible

G. F. Nane (✉)
Delft University of Technology, Delft, Netherlands
e-mail: g.f.nane@tudelf.nl

A. M. Hanea
University of Melbourne, Melbourne, Australia

A. M. Hanea et al. (eds.), *Expert Judgement in Risk and Decision Analysis*,
International Series in Operations Research & Management Science 293,
https://doi.org/10.1007/978-3-030-46474-5_11

forces acting; gravity is pulling the glass down and the table is pushing back up with, miraculously, the exact same magnitude in the opposite direction so that nothing happens. Take away the table and the glass falls, take away gravity and nothing happens. To do the simple physics exercises, you have to know the code, and know which questions not to ask. I eventually learned the code, but retained the sense that there were flaws in the story. The real puzzle for me was why "force" worked and "phlogiston" didn't. Why "space" and "time" worked but "ether" didn't.

I switched to Philosophy and spent a lot of time on the Greeks, the Scholastics, and Enlightenment philosophers, especially Kant and Hegel. The great philosophers construct a coherent system for understanding everything. In so doing they start with the natural language and progressively re-wire it so that concepts successively acquire new meanings, defined in evolving contexts. You can't understand it piecemeal; you just have to keep going until it all starts making sense. Once you "get it" you can see everything in a new way, like a conversion experience. Most people have at most one conversion experience which they then regard as apodictic. If you study philosophy, you go through several....it helps. In retrospect, that's one of the great things I learned in philosophy, that and how to read seemingly unintelligible texts.

Here's an anecdote: a math colleague and I were trying to learn atmospheric dispersion modeling. Atmospheric chemists have their own code, which mathematicians find inscrutable. We started with an elementary text. The colleague would come to something he didn't understand, stop, and look for another text to explain. That sequence doesn't converge. I would just keep reading until their code starts to become intelligible.

The great Systematic Philosophies have at their core theory of knowledge. What knowledge is determines what we can know; 'what we know' and 'how we know' are very tightly coupled. For Plato, knowledge was acquired by the direct intuition of a soul sufficiently purged of false beliefs. For the Scholastics, knowledge is Reason applied to Divine Revelation. For Kant, Newton's mechanics and Euclidean geometry enjoyed a level of certainty not attainable by induction from observations: they are necessarily imposed on our perceptions of nature by our knowledge apparatus—or so he thought. He was wrong about that, but he was right that, to turn a philosophical phrase, knowledge and the knowledge of knowledge are inseparable. For Hegel, knowledge is the self-consciousness of trans-personal spirit. Another story altogether.

3. **Foundations of probability played a big role in your thinking, could you elaborate?**

The flaws in classical mechanics began to extrude themselves in the later nineteenth century and people felt that language was a big part of the problem. In an effort to separate pure definitions and mathematics from deliverances from experience Heinrich Hertz gave the first axiomatization of classical mechanics in a proto-formal language. He found the notions of force and absolute space–time superfluous and unhelpful. Such formal approaches were cross-fertilized with activity at the foundations of mathematics—another story. Ernst Mach invented "semantic analysis" whereby notions must obtain a semantic pedigree tracing them to elementary sensations before they are serviceable to science. It emerged that concepts like phlogiston, force, absolute space–time lacked semantic pedigrees. Propositions assigning them properties are not unknowable, they are meaningless. The power of that insight emerges when you contemplate all the unknowable things people believe. The revolutions of relativity and quantum mechanics drew heavily on semantic de-constructions. Mach himself believed that atoms also lacked a semantic pedigree and that propositions about atoms were therefore meaningless. Atoms, however, proved very useful. Indeed where would modern physics be if confined to Mach's semantic strictures?

Philosophy of science emerged as an effort to articulate the scientific method and thereby determine what science is and is not. Is risk analysis science, psychoanalysis, creationism, economics? Terms like leptons and quarks do not have operational meaning in the narrow sense as they are not directly linked to measurements, yet they seem to be ok. What about Freud's id, creationism's intelligent designer, economists' representative consumer? What about randomness, fuzzy membership, degrees of possibility? You see where this is going.

The demarcation of science and non-science is closely bound up with the problem of "theoretical terms": articulate a semantics in which terms without direct operational meaning, nonetheless acquire meaning in a given theory. There is a load of active literature on this, which I have tried to boil down to a simple formula (see Cooke 2004)."

> The operational meaning of "*degree of possibility*" in the proposition: "*The degree of possibility that the Loch Monster exists is 0.0031416*" is the set of non-tautological propositions not containing "*degree of possibility*" which that proposition implies.

What about "uncertainty," what does it mean? In the natural language, it means different things in different contexts, including ambiguity, ambivalence, confusion, distrust, unpredictability, and indecisiveness. Anyone wishing to "represent uncertainty" in a scientific context must do some serious re-wiring. As often happens, a scientific reconstruction of a term in the natural language captures only part of its native meaning. Compare "force" in physics and in the natural language.

L.J. Savage's foundation of subjective probability is a superlative example of rational reconstruction in science. He provides axioms describing rational preference with clear operational meaning for the primitive terms. Strong arguments support his axioms—maybe not as strong as arguments for the axioms of Zermelo Frankel set theory but very strong nonetheless. He then proves that the preferences of a rational individual can be represented as expected utility, where a personal probability (aka subjective degree of belief) is uniquely determined and the utility function is unique up to a positive affine transformation. All my students had to learn these proofs, not only to understand uncertainty but also to understand how to extend the purview of science. The germ of Savage's rational reconstruction comes from Ramsey (1931).

Others may protest that uncertainty means much more than subjective probability. Duh. However, if you want to quantify, say ambiguity, you must provide operational meaning telling us whether it is, e.g., positive, invariant under monotone, affine or ratio transformations, etc. Those invariances must be derived from the operational meaning of the primitive terms. At a conference in Paris, a leading light presented his new definition of uncertainty which unbeknownst to him, allowed uncertainty to take negative values. The theologians would love that. There have been many variations on Savage's axioms, just as there have been many variations on Zermelo–Frankel set theory, but they all remain variations around a core theory that is suitable for applications. There are also countless "alternative representations" of uncertainty that lack any foundation whatsoever.

4. **How did the idea of a rational consensus emerge—can you describe what it is and why you think it is useful to policy and decision-makers?**

We come to the theme of extending the purview of science. Traditional philosophy of science pretends that, within the context of justification, science deals only with certainties and reasons deterministically. It isn't so. Society is increasingly confronting decisions with large uncertainties with consequences impacting our survival. We all know the myriad ways in which private interests can and will exploit uncertainty to further their own aims. We must bring 'decision making under uncertainty' within the purview of science. Savage provides necessary but not sufficient conditions for rational decision making under uncertainty. Indeed, ANY subjective probability combined with ANY utility is rational in the sense of Savage. Rationality in science, whatever that means, is much more restrictive. The challenge is to bring science-based restrictions into Savage's model, at least with respect to probability, such that all subjective probabilities are not equal. Utility is another problem. Validation is not hopeless but much less active than the probability component of rational decisions (see Neslo and Cooke 2011).

I first encountered the term rational consensus in a book by Keith Lehrer and Carl Wagner (Rational Consensus in Science and Society 1981). It is similar to that of De Groot (1974), discussed in Experts in Uncertainty. Participants assign probabilities to events and weights to each other's probabilities, leading to an equilibrium distribution. There's nothing scientific about it IMO, and it is not remotely practical. Experts are overworked and underpaid. They're not going to travel long distances to sit together and reach 'dialectical equilibrium' as a prerequisite for weighing each other.

However, the term rational consensus stuck in my mind and I sought a more science-based meaning. The idea is that experts construct their rational consensus. They quantify their degrees of belief as subjective probabilities for both the variables of interest and for calibration variables taken from their field. They are scored as statistical hypotheses with respect to statistical accuracy and informativeness. The theory of strictly proper scoring rules, appropriately generalized, converts their scores into weights. The combination scheme satisfies necessary (not sufficient) conditions for the scientific method. Rational consensus means that experts pre-commit to the results of the combination. They needn't adopt the result as their personal probability. However, withdrawing from the rational consensus imposes a proof burden of showing how the necessary conditions were violated or should be improved. The necessary conditions are traceability, neutrality, fairness, and empirical control. The last is of course the most consequential, it implements Popper's idea of falsifiability.

Fairness excludes pre-judging experts, neutrality corresponds to proper scoring rules, traceability means that all steps in the calculation must be open and reproducible.

Tony O'Hagan's question 'is rational consensus a subjective probability, if so whose?' gets the simple answer: it is the personal probability of any rational agent whose preference representation as expected utility has a personal probability agreeing with the rational consensus.

5. **Can you tell us something about the types of risk problems that you were thinking about when you started developing your ideas about expert judgment?**

The topology of the problems was defined in the Rassmussen Report (USNRC 1975) and evolved through three generations as described in (Cooke 2013). We have panels of order 10 experts assessing up to 100 uncertain quantities. Discrete events are sometimes assessed, but most variables are effectively continuous. The Rasmussen report did a good job on traceability. Publishing all the expert raw data made visible the very large differences between experts, thereby raising the issues of combination and validation. The Rasmussen report selected the distributions used in the report in a rather inscrutable fashion. In the second generation studies, experts' rationales were cataloged and their distributions were combined with equal weighting. The third generation in which I participated added performance measurement, empirical validation, dependence modeling, and probabilistic inversion.

6. **Do you think over the years research on SEJ methods has focused on the right areas of EJ? How important do you think the social sciences side of EJ is?**

Classical Model (CM) drew heavily on decision science research from 1950 to 1990. Publication of the Delft SEJ database (Cooke and Goossens 2008) spawned good research, starting with the special of RESS (Cooke 2008). Wisse et al. (2008) looked at moment based elicitations, rather than quantile elicitation. Lin and Bier (2008) regressed expert calibration on study parameters and found an 'expert effect', suggesting that differences in expert statistical accuracy are not explained by random fluctuations. Perhaps the most productive was Clemen's critique. In addition to raising all the familiar questions regarding calibration variables, he introduced the issue of cross-validation. His method is Remove One At a Time (ROAT): calibration variables are removed one at a time and the recomputed Decision-Maker (DM) assesses the excluded calibration variable. Thus, with ten calibration variables, each is assessed by a different DM using weights based on the non-excluded items and scored for performance on the excluded items. Clemen (2008) analyzed 14 cases in this way and found only 9 (62%) in which CM outperformed equal weighting (EW), which was not statistically significant. Clemen's numbers checked out and I spent quite a bit of time analyzing this. On typical data sets, removing one calibration variable can change an expert's calibration score by a factor 2 or 3, hence ROAT can upweight experts who assessed the excluded item badly. Doing this for all variables introduces a significant bias against performance weighting. I finally found a simple example that made this very clear Cooke (2012) and Colson and Cooke (2017). Colson and Cooke (2017) give a complete discussion of the ROAT cross-validation exercises with CM. These exercises used own code which was not benchmarked against our publically available code EXCALIBUR, (http://www.lighttwist.net/wp/excalibur) and contained wildly divergent scores. We spent a lot of time trying to figure out what they were doing, even going so far as to obtain and analyze their codes where possible. Those studies can be bracketed (e.g., Lin and Cheng 2008, 2012; Flandoli et al. 2011).

Eggstaff et al. (2014) performed a very serious cross-validation on the 62 studies available at the time. They took every non-trivial subset of calibration variables as a training set to initialize the CM and scored performance on the complementary set. With 15 calibration variables, there are 32,766 splits of training/test sets. Abby Colson and I worked with Lt. Col. Eggstaff for some time until we got an exact agreement with EXCALIBUR. This is the only cross-validation code for which this has been done, to my knowledge. We used this code for the cross-validation of the post-2006 studies, and still, use it. There are many subtle issues involved in such studies, but the upshot is that performance weighting (PW) outperforms equal weighting (EW) out-of-sample on 72% of studies, similar to Colson and Cooke (2017). Using the recommend training set size of 80% excludes training sets with very low power and pushes the fraction to 78%. Including the most recent studies brings the number to 84%. The bias in ROAT is very roughly the difference between 62% and 84%. The hypothesis that PW and EW are not statistically distinct is rejected at the 1.6E–6 level. All this activity emerged from Clemen's critique.

Researchers at George Washington University are exploring a new idea. EW is based on the idea that one expert is as good as another. If that were true, then a randomized panel should do just as well as the original panel. In other words, we could construct new experts by randomly scrambling the original expert assessments and it would perform just as well statistically. Initial results roundly reject the random expert hypothesis. This approach is potentially more powerful than cross-validation because it doesn't require splitting the calibration variables. We're now comparing median predictions. It turns out that averaging medians (equally or performance weighted) yield markedly higher prediction errors than using medians of equally or performance weighted combinations as shown in the Chap. 3, this volume. Moreover, performance weighting outperforms equal weighting in point predictions. Hence, even if one is only interested in point predictions, it is better to quantify uncertainty, measure performance, and performance weight the experts' distributions (Cooke et al. 2020).

These are active mathematical research themes. Other active themes include dependence modeling (Werner et al. 2017), dependence elicitation (Morales et al. 2008), stakeholder preference, and probabilistic inversion (Neslo and Cooke 2011). The social sciences have also made enormous contributions to this field, for example, the many publications of the Eugene Oregon school, of Kahnemann and Tversky, and of Fischoff. I got updated on the social science themes as lead author for the chapter on Risk and Uncertainty for IPCC AR (5).

Once we know how to measure expert performance, research into the best training methods would be very helpful. This would probably link with risk communication research. It's a topic I hope the social scientists will pick up.

I would also like to see a good psychometric experiment that tests the Elsberg paradox where there is no information asymmetry between the experimental subject and the experimenter. The Elsberg paradox shows that people prefer a lottery with objective probability (½, ½), to a lottery with probability unknown to the subject (but known to the experimenter). I suspect that a large part of the "ambiguity aversion" effect is due to "manipulation aversion" when the subject knows that the experimenter knows more than (s)he does. For example, let the subject choose an odds ratio (1 − r)/r, and let a fair coin determine which side of the lottery the subject will play. Now, the probability of winning is equally uncertain to the subject and experimenter alike. Do subjects still prefer a fair coin toss? By how much? and if you decrease the win on the fair coin from $10 to $9.90?

7. **Do you think that the definition of SEJ from your book would need a revision? And are there any methods except the Classical Model (CM) that you would think are part of the SEJ group of methods/protocols?**

CM has stayed pretty much the same, the only change from the book is that information is measured as relative information with respect to a background measure instead of inverse entropy—this was just for cosmetics to make the role of the background measure more visible. Relative information is a familiar concept, inverse relative entropy less so. Keeping the model unchanged helped build up a large database of SEJ applications.

Re other methods: The IDEA protocol for discrete events is a very promising initiative (Hanea et al. 2016). It combines the CM with Delphi-like feedback rounds. Philip Tetlock's good judgment project (Unger et al. 2012) has had success forecasting current events measuring performance with the Brier score and successively eliminating experts until a small subset of "super forecasters" is found. The time and resources (in a number of experts) required to preclude application to science and engineering problems. Eliminating experts is a form of performance weighting. There have been very many proposals that do not attempt to validate their performance. To all these, I say *Why Not?* It's getting harder to pretend that validation is impossible.

8. **Your book and other references contain numerous practical suggestions about performing an elicitation. Have any of these advices (technology, remote, etc.) changed with time?**

The book says that elicitations should not exceed on hour. I would now say one-on-one elicitations must not exceed four hours. Four hours are grueling. Other formats are now employed, including remote elicitation with e-tools.

9. **Your Classical Model has been criticized by some because of the way it combines different paradigms to uncertainty—for example in using classical statistical tests as a tool to construct a judgmental probability distribution. Since you have impeccable credentials in the philosophy of science you will be entirely aware of this, and of the other "rough and ready" choices you**

made in designing this approach. Is the classical model grounded in science or is it a part of what some may call "decision engineering"?

Familiarity with foundations teaches that the combination of experts is not a mathematics problem—the axioms of probability will never tell us how to combine experts. Its also not a problem of personally expected utility maximization and Bayesian approaches never achieved lift-off (see Cooke 1991). The expert problem is akin to an engineering problem. We define the objectives and look for a design that optimizes performance. In our case, the objective is to promote rational consensus through science-based uncertainty quantification. We use first principles, the axioms of probability, and second principles, Savage's axioms, but they obviously won't give us a working design. Tertiary principles like the marginalization property leading to the linear pool, and quaternary principles like scoring rules, P-values, and Shannon relative information are also needed. Finally, we need to apply common sense. Any arbitrary choice of the analyst should be manipulable in the code and available for robustness analysis. Examples are the choice of background measure, the choice of calibration power, and choice of P-value cut-off.

Some mathematicians don't appreciate the difference between mathematics and engineering: a bicycle obeys Newton's laws but doesn't follow from them. The design of a bicycle involves many decisions motivated by different considerations; the wheels could be a millimeter larger, the saddle a millimeter smaller, etc. A design always mixes physics, psychology, economics, etc. Complaints from academics about ad hoc-ness and methods mixing are like someone refusing to ride a bicycle because the optimal bicycle cannot be derived from Newton's laws. Such righteousness is most laudable.

10. **Many decision-makers and social scientists are familiar with the measures 'accuracy' and 'precision'. How do those relate to the CM?**

CM's performance measures of statistical accuracy and information do not map neatly onto the terms "accuracy" and "precision," which are familiar to social scientists. Accuracy denotes the distance between a true value and a mean or median estimate, and precision denotes a standard deviation. While appropriate for repeated measurements of similar variables, these notions are scale-dependent and therefore not useful in aggregating performance across variables on vastly different physical scales. For example, how should one add an error of 10^9 colony-forming units of campylobacter infection to an error of 25 micrograms per liter of nitrogen concentration in the water? Expert judgments frequently involve different scales, both within one study and between studies. For this reason, the performance measures in the Classical Model are scale-invariant. That said, the exhaustive out-of-sample analysis of Eggstaff et al. (2014) found that the realizations were closer to the PW combination's median than the EW combination's median in 74% of the 75 million out-of-sample predictions based on the TU Delft data. Such non-parametric ordinal proximity measures, proposed by Clemen (2008) are not used to score expert performance, as the scores strongly depend on the size of the expert panels.

11. **Maybe the biggest criticism of the CM is the lack of representativeness of the seed questions for the questions of interest and the way performance is measured on those seeds.**

The claim that performance on calibration variables cannot represent performance on variables of interest is just a bald assumption. Scientists don't traffic in bald assumptions; they look at the evidence for or against the statement that performance on the calibration variables predicts performance on the variables of interest. Clemen (2008) is the only critic who used valid code, to my knowledge. Since the representativeness question gets asked on virtually every application, I have a standard answer. Suppose you have two experts, one is very accurate statistically and very informative on the calibration variables, the other is massively overconfident with abysmal statistical accuracy. Would you give them equal weight on the variables of interest? If your answer is "yes," then calibration variables have failed in their function.

The CM is subjected to empirical control in-sample on every application, including leave-one-out robustness on experts and calibration variables. It is validated out-of-sample with cross-validation. In some studies, the actual variables of interest have been observed post hoc (Goossens and Cooke 2008). There are new ideas in the pipeline. My hope is that other approaches will also be motivated to address validation. For example, why shouldn't the Delphi method validate itself? The early studies did compare predictions with reality with very uneven results (see Cooke 1991). Has the record improved? Why wouldn't practitioners of the Sheffield method attempt to validate their results against observations? Why shouldn't proponents of imprecise probabilities say what a good imprecise probability assessment is, and measure how well their methods perform? If one degree of possibility is as good as another, one imprecise probability interval as good as another, one fuzzy membership as good as another, then why go to all the trouble? Just use Happy Numbers, i.e., numbers that make you happy.

In view of all the research into validation, the claim that validation is impossible becomes a bit fatuous. Expert judgment is a raucous field with practitioners from very diverse backgrounds. Applying the CM requires a level of numeracy that many analysts may find challenging. Indeed, the analyst has to understand what a likelihood ratio is, what Shannon information is, what a proper scoring rule is. (S)he has to explain the CM to the experts and write it up intelligibly. Perhaps most importantly, (s)he must be able to explain why (s)he is NOT following any of the other approaches in circulation…Bayesian averaging, quantile averaging, consensual probabilities, imprecision, fuzziness, degrees of possibility, Delphi, etc. Of course, mathematicians know that the CM is not a heavy lift and many scientists and engineers have become adepts. However, to a non-numerate person, it may look like a very heavy lift.

People often ask if experts can game the system. It is theoretically possible and would be easy to spot. In all our applications there was one expert in one panel for whom we suspected that his business interests were informing his assessments.

12. **What do you see as the greatest challenges facing EJ practitioners in the coming years?**

Finding enough qualified analysts.

13. **One of the comments to the "Structured expert judgment" post from 2015 by Judith Curry said: "Uncertainty, like love, cannot be quantified. There is nothing to measure."**

Ha Ha, sounds funny but isn't. Such cavalier attitudes towards uncertainty unwittingly license all the defective modes of dealing with uncertainty with which we now struggle (Cooke 2015). I'll mention three books to underscore this. The first is Oreskes and Conway (2010) "Merchants of Doubt." They detail the massive resources spent by private interests to create doubt and derail government regulation. Does smoking cause cancer? NOT PROVEN!! look at this research sponsored by Tobacco Lobbies. Such tactics work as long as the public is unwilling and unable to reason under uncertainty. The question isn't whether we KNOW that smoking causes cancer; of course, we don't. We don't KNOW that $F = ma$. The question is how much smoking raises the risk of cancer. We see the same small set of "experts" pandering to private interests, whether it is supersonic transport, smoking, air pollution, or climate change.

The second book is Malcom Nance (2018) "The Plot to Destroy Democracy." Nance is a veteran cybersecurity specialist in the US Government and gives a detailed account of Russia's propaganda machine. The story goes back to Lenin but has taken a new form in the age of the internet. Hundreds and hundreds of Russians work 24/7 concocting lies targeting specific groups and pushing them onto the internet. A plausible lie gets a preamble in known facts embellished with things the target wants to believe, then topped off with a complete fabrication. This formula was applied to Pizza-gate (Hilary Clinton ran a pedophile trafficking ring from the basement of a Washington Pizza parlor) of which Michael Flynn Jr wrote: "until #PizzaGate is proven false it'll remain a story." Less well known was, e.g., the fake news headline in St. Mary Parish, Louisiana "Toxic fume hazard warning in this area until 1:30PM." Despite being a proven Russian hoax, a Wikipedia page and YouTube video showed ISIS claiming responsibility with Burqa's waiving guns. And then D.J. Trump's "The Art of the Deal" written by Tony Schwartz. Trump reveals his tactic of "truthful hyperbole" (https://www.fastcompany.com/3068552/i-call-it-truthful-hyperbole-the-most-popular-quotes-from-trumps-the-art-of-the-deal). I interpret it thus: any proposition not provably false which suits your interests should be repeated as often as possible, challenging the adversaries (there are always adversaries) to disprove it. Since they can't, your story will win. I'm not saying that Savage can deliver us from all this. I am saying that peoples' unwillingness to reason probabilistically makes it possible to influence their behavior by pushing the proof burden to the side you want to lose. 'You haven't PROVED that smoking causes cancer','You haven't PROVED that climate change is real', 'You haven't PROVED that Russia hacked the US election', etc.

An interesting aside on the Russian story: it was the Dutch AIVD that pinpointed the source of Russian troll farm COZY BEAR to a building in Moscow. They even counter-hacked the security cameras on one particular floor of the building and observed the Russian spies using the system.

14. **Finally, if you could organize a dinner party with 3 or 4 'great thinkers' who influenced your development of the classical model, who would you invite and why?**

Learning to reason probabilistically will be an event in the cognitive history of Man comparable to the formulation of deterministic reasoning in Aristotle's Logic. The great hero here is Frank P. Ramsey. His "Truth and Probability" (written in 1926, published in 1931) is a bolt of sheer genius. Let's also include John von Neuman (Theory of Games and Economic Behavior 1944) and Lenard Jimmy Ogashevitz (aka Savage)(The Foundations of Statistics 1954). But not for dinner—nobody could get along with von Neumann.

The most important people at the inception of CM were Louis Goossens, Max Mendel, and Simon French. Early adapters from the first hour were Willy Aspinall, Tim Bedford, Jan van Noortwijk, Matthijs Kok, Dmitri Solomatin, Gordon Woo, Tom Mazzuchi, and Christian Preyssl. Follow on forces include Dorota Kurowicka, Anca Hanea, Tina Nane, Oswaldo Morales, Jim Hammitt, John Evans, Abby Colson, John Quigley, Justin Eggstaff, Rene van Dorp, Arie Havelaar, and Ben Ale. These would also need to be invited; we will need a Banquet Hall. Then we can also invite all the colleagues who performed the applications, Kim Thompson, Radboud Duintjer Tebbens, Juoni Tuomisto, Nicole van Elst, Daniel Puig, Frank van Overbeek, Xi Quing, Maurits Bakker, Rabin Neslo, Daniel Lewandowski, Sandy Hoffmann, Matt Gerstenberger, Maart Janssen, Augusto Neri, Eric Jager, Ben Goodheart, Juliana Lopez de la Cruz, Julie Ryan, Maartin Nauta, Marion Whitmann, David Lodge, John Rothlisberger, Arno Willems, Jim Smith, Fred Harper, Steve Hora, Mark Burgman, Elizabeth Beshearse, Raveem Ismail, Vicki Bier, Bernd Kraan, Ben Koch, Daniela Hanea, Christoph Werner, Bis Bhola, Michael Oppenheimer, Jonathan Bamber, Bob Kok, Monika Forys, Michael Tyshenko, Maartin Nauta, Karin Slijkhuis … with apologies to everyone I forgot.

Recalling all these people and their contributions is quite humbling. BTW, didn't we have just such a banquet in July 2017?

References

Clemen, R. T. (2008). Comment on Cooke's classical method. *Reliability Engineering & System Safety, Expert Judgement, 93*(5), 760–765. https://doi.org/10.1016/j.ress.2008.02.003.

Colson, A. R., & Cooke, R. M. (2017). Cross validation for the classical model of structured expert judgment. *Reliability Engineering and System Safety, 163,* 109–120.

Cooke, R. M. (1991). *Experts in uncertainty: opinion and subjective probability in science*. New York: Oxford University Press.

Cooke, R. M. (2004). The anatomy of the Squizzle—the role of operational definitions in science. *Reliability Engineering and System Safety, 85*(2004), 313–319.

Cooke, R. M. (2008) Special issue on expert judgment, Editor's Introduction Reliability Engineering & System Safety, 93(5), Available online 12 March 2007.

Lin, S. W., & Cheng, C. H. (2009). The reliability of aggregated probability judgments obtained through Cooke's Classical Model. *Journal of Modelling in Management, 4*(2), 149–161.

Cooke, R. M. (2012). Pitfalls of ROAT cross-validation: comment on effects of overconfidence and dependence on aggregated probability judgments. *Journal of Modelling in Management, 7*(1), 20–22.

Cooke, R. M. (2013). Uncertainty analysis comes to integrated assessment models for climate change… and conversely. *Climatic Change, 117*(3), 467–479.

Cooke. R. M. (2015). Messaging climate change uncertainty. *Nature Climate Change*, 5(1).

Cooke, R. M., & Goossens, L. J. H. (2000). Procedures guide for structured expert judgment Project report EUR 18820EN, Nuclear science and technology, specific programme Nuclear fission safety 1994–98, Report to: European Commission. Luxembourg, Euratom. Also in Radiation Protection Dosimetry Vol. 90 No. 3.2000, 64 7, pp. 303–311.

Cooke, R. M., & Goossens, L. H. J. (2008). TU delft expert judgment data base. *Reliability Engineering & System Safety, Expert Judgement, 93*(5), 657–674.

Cooke, R. M., Marti, H.D. & Mazzuchi, T. A., (2020) Expert Forecasting with and without Uncertainty Quantification and Weighting: What Do the Data Say? *To appear in International Journal of Forecasting.*

DeGroot, M. (1974). Reaching consensus. *Journal of the American Statistical Association*, *69*, 118–121.

Eggstaff, J. W., Mazzuchi, T. A., & Sarkani, S. (2014). The effect of the number of seed variables on the performance of cooke's classical model. *Reliability Engineering & System Safety, 121*(January), 72–82.

Flandoli, F., Giorgi, E., Aspinall, W. P., & Neri, A. (2011). Comparison of a new expert elicitation model with the classical model, equal weights and single experts, using a cross-validation technique. *Reliability Engineering & System Safety, 96*(10), 1292–1310.

Hanea, A. M., McBride, M. F., Burgman, M. A., Wintle, M. A. (2016). Classical meets modern in the IDEA protocol for structured expert judgement. *Journal of Risk Research*, *21*(4), Published online: 09 Aug 2016.

Lehrer, K., & Wagner, C. (1981). *Rational consensus in science and society*. Dordrecht: D. Reidel.

Lin, S. W., & Cheng, C. H. (2008) Can cooke's model sift out better experts and produce well-calibrated aggregated probabilities? In *IEEE International Conference on Industrial Engineering and Engineering Management*, IEEM 2008, pp. 425–29.

Lin, S. W., & Cheng, C. H. (2012). Effects of overconfidence and dependence on aggregated probability judgments. *Journal of Modelling in Management*, *7*(1), 6–22. Lin, S. W., & Bier, V. M. (2008). A study of expert overconfidence. *Reliability Engineering & System Safety, Expert Judgement*, *93*(5), 711–21.

Nance, M. (2018). *The plot to destroy democracy*. New York: Hachette Book Group.

Neslo, R. E. J., Cooke, R. M. (2011). Modeling and validating stakeholder preferences with probabilistic inversion, Appl. Stochastic Models Bus. Ind. Games and Decisions in Risk and Reliability Analysis, *27*(2), 71–171, First published: 04 April 2011.

Morales, O., Kurowicka, D.,& Roelen, A. (2008). Eliciting conditional and unconditional rank correlations from conditional probabilities. *Reliability Engineering & System Safety*, *93*, 600–710. Available online 12 March 2007, Volume 93, Issue 5, May 2008.

Oreskes, N., & Conway, E. M. (2010). Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming (1st U.S. ed.). New York: Bloomsbury Press.

Ramsey, F. (1931). Truth and Probability" in Foundations of Mathematics and Other Logical Essays, London, originally written in 1926.

Savage, L. J. (1954). *The foundations of statistics.* Wiley.

U.S. Nuclear Regulatory Commission. (1975). “Reactor Safety Study.” WASH-1400, NUREG-75/014. Washington, D.C.

Ungar, L., Mellors, B., Satopää, V., Baron, J., Tetlock, P., Ramos, J., & Swift, S. (2012) The good judgment project: A large scale test of different methods of combining expert predictions. *2012 AAAI Fall Symposium Series.*

von Neumann, J., Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton University Press.

Werner, C., Bedford, T., Cooke, R. M., Hanea, A. M., & Morales-Nápoles, O. (2017). Expert judgement for dependence in probabilistic modelling: A systematic literature review and future research directions. *European Journal of Operational Research, 258*(3), 801–819.

Wisse, B., Bedford, T., & Quigley, J. (2008). Expert judgement combination using moment methods. *Reliability Engineering & System Safety, 93*(5), 675–686.

Part III
Process, Procedures and Education

Structured expert judgement is effective not just because of the theory on which it is built but also because of the processes and procedures that surround its implementation. The chapters in this part discuss and illustrate those processes and procedures, and ways of training analysts to use them.

Chapter 12
Scientific Advice: A Personal Perspective in Dealing with Uncertainty. An Interview with Prof Dame Anne Glover, in Conversation with Tim Bedford

Anne Glover and Tim Bedford

INT: **Anne, the purpose of this discussion is to understand more about the role of expertise in the scientific advisory mechanisms for government and more specifically about the role of a CSA as an intermediary between the policy world and the science world. Obviously, your CSA roles have given you unique experience of interacting at the highest levels with public authorities. I'm sure there are lots of different aspects that we could cover, but for the purposes of this conversation I am particularly interested in your general experience of the use of scientific expertise, the expectations that government—whether civil servants or ministers themselves—have of scientific advice mechanisms and how those match expectations of scientists, and the ways in which both groups do or do not articulate uncertainty.**

RES: I think it's important to say that in the posts I've held, where my responsibility has been about providing scientific advice to governments, it was the first time government had such a role. Although I had a broad network to rely on in terms of procuring advice and helping to synthesise that advice to provide expert judgement, what was not expert was the receiving end of that advice. The value of expert judgement is two-fold. One part of it is how you procure that expert judgement including what methodology you use to approach that, and the second is the capacity to absorb and use that advice. For example, the recipient of the advice may already have a bias regarding what they want to hear, often referred to as policy-led evidence, which is a difficult thing to overcome when you are trying to provide evidence in order to underpin

A. Glover · T. Bedford (✉)
University of Strathclyde, Glasgow, UK
e-mail: tim.bedford@strath.ac.uk

A. M. Hanea et al. (eds.), *Expert Judgement in Risk and Decision Analysis*,
International Series in Operations Research & Management Science 293,
https://doi.org/10.1007/978-3-030-46474-5_12

an evidence-based policy. In my experience, not in every case, but there is a very strong, probably philosophically led, approach to how you should go about procuring evidence. It starts with the question you ask. A fictitious example I might give to illustrate the point is that a European Commissioner might be thinking over the weekend: "I'm really concerned about financial instability within society, how people manage money and get themselves into debt and how this impacts more largely on banks wishing to lend in an inappropriate way, to very high risk to people who are prepared to take on debt". The Commissioner might be thinking about that whole issue and then decide "OK, I think probably one way of dealing with that would be to restrict the use of credit cards in the European Union and that would be a good policy announcement". Making such announcements could be regarded as one measure of success for a Commissioner. So they might come in on Monday morning and say to their officials, the civil servants, "Find me the evidence that the use of credit cards enhances the likelihood of debt amongst families that cannot hope to repay that debt". If there's expertise within the area then the officials will try and provide that, but more likely than not they won't have the expertise so they will go to a consultant to say "We would like to look at evidence implicating credit card use as a cause of financial instability" and what you get is a report telling you just that. This is an example of asking the wrong question which pre-supposes the answer you want to get.

INT: **So there are a couple of things bound up in that, one is the framing of the initial question, how narrow or how broadly you frame that question, and the other aspect of it is the fact that the politicians have got the right and duty to make their value judgements that they've been chosen to make surely?**

RES: Yes, but at the end of the day it would not be expert judgement if they just had a hunch so it's not based on anything. It would just be trying to camouflage a narrative of "I don't like credit card companies and I think they encourage unsustainable personal debt" by the procurement of biased evidence to confirm the hunch. [As a CSA] if you're not framing your request for evidence in a proper way then the person who is asking you for the advice can easily be undermined by being challenged on the relevance and diversity of the advice sought.

INT: **Is one of the wider roles of a CSA then, to try and help them frame questions in a better way?**

RES: Ideally yes. You want to understand the issues that are concerning politicians so that you might offer them as much relevant and useful evidence or advice as possible. An example might be a concern a politician had about the increase in number of people in hospitals with respiratory disease and their wish to understand why that was. The understanding might then offer some policy interventions to improve the situation

A good example would be around passive smoking where you might identify a linkage between the environment patients were exposed to and they are being admitted to hospital with respiratory disease. The evidence would be built up to implicate exposure to second-hand smoke as being the primary cause of their illness. That would be a straight forward provision of evidence which would allow a politician to consider policy interventions to reduce the exposure of citizens to secondary smoke. A more challenging example might be where politicians don't want to hear the evidence because they've already decided what they want to do about an issue. A good example of that in Europe is that the vast majority of European citizens across all 28 Member States do not wish to have genetically modified food on their shelves and so the politicians would quite like to regulate against selling GM food in Europe or importing GM food for human consumption. But, if we presume that policy should be based on evidence, it is not possible to find credible evidence to support the claim that GM food is harmful. Politicians then might change the nature of the question to ”can you be 100% certain that GM food is safe?” Now if you just ask that question my answer as an advisor would be “no” but you could ask a supplementary question which is “can you be 100% sure that non GM food is safe?” My answer would be the same “no” because I can't be 100% certain for either. But they want the uncertainty in the first case and they don't want to hear it in the second case, and so the political solution is to say that we don't have enough evidence to judge the safety of GM food. In fact, by any objective measure, we've got more evidence about the safety of GM food than we do for any type of food. This is a good example of trying to legitimise a policy on the basis of saying there is insufficient evidence, whereas, in other areas, policy might be made on the basis of scant evidence if there was a political imperative.

INT: **So do scientific advice mechanisms fear to tread in areas where there's a very strong cultural preference?**

RES: No, I think if you fear to tread and you alter your advice you're dead as a scientific advisor.

INT: **I wasn't saying you would alter your advice, but is it less welcome in those areas.**

RES: I think it is probably not welcome but the value in speaking to somebody who's seeking advice is to say that sometimes you'll like the evidence that is presented to you and sometimes you won't, but the evidence will be independent of the level of political bias or philosophical bias that they use.

INT: **But biases can sometimes arise in the interpretation of that evidence, or in the way evidence is extrapolated to new situations?**

It is also worth mentioning that science is not value-free either. Scientists have values, and we shouldn't pretend that we can be absolutely objective when it comes to the advice or the evidence that we put forward because we as scientists can also be selective in how we address evidence. I challenged my own approach to evidence after a politician gave me a book on unconscious bias saying “you might find this interesting”. Of course, I knew about unconscious bias but I like reading so I read the book and it made me think. If I continue with the example of GM, then I realised that I've looked extensively at the consensus of evidence around the safety of GM technology in order to produce new food varieties and I am convinced that it's as safe, if not safer, than any other technology we might use to produce a new variety of food, using conventional plant breeding technologies. What was interesting to me is that I recognised that if a paper was published that concluded that GM food is dangerous then I went through the methodology and every part of that paper with a fine toothcomb because it did not agree with my previous thinking and my judgement on the evidence that was available. But if another paper came out that said GM food is safe, I didn't go through it with the same degree of scrutiny. So that's why I'm saying my own values and confidence in my own judgement affected my approach. You have to remind yourself that both papers require the same degree of scrutiny or you're not doing your job as an advisor or expert.

INT: **In public debate, there is sometimes a resistance to accepting evidence, you might call a conscious bias, and I wonder what the underlying reasons are for this**

RES: Often people have very strong views that they haven't dissected in their own minds e.g. on GM. Someone against would probably agree there is no evidence against the technology but they might agree that it hasn't been used in a beneficial way. Many organisations against the use of GM raise issues like the impact on bio-diversity, increased use of herbicides, impact on small farmers. But these things didn't arise from the technology, but from the way, it was used by Monsanto. The technology is conflated with the business practice. The call to ban GM is a simple message that seems to be driven by this.

INT: **From a theoretical decision-making point of view when you're making decisions under uncertainty there is the separation on the one hand between the uncertainties which are supposed to be measured by probabilities and could be assessed by expertise or by experimentation perhaps and on the other hand the value judgements—trade-offs and so on—which are being made by the politicians. You are observing that there are biases on both sides which could affect the outcome.**

RES: Yes. You can get well-respected scientists who will take opposite views on topics where there is demonstrable uncertainty such as the impact of low levels of endocrine-disrupting chemicals (EDCs) in the environment. One scientist might say that we should ban all use of EDCs because the potential impact to humans could be very harmful and another scientist who would vehemently disagree because there is no evidence to demonstrate such harm at low-level exposure and that they are really valuable chemicals for society. They try to undermine each other's arguments and, of course, both groups are good scientists, but they try to discredit each other. None of this is helpful to the policymaker as they can choose what advice they wish to use.

INT: **So this is actually the scientific method in all its glory at work as run by humans rather than purist exponents of the scientific method.**

RES: Yes, it is run by humans. The evidence has a degree of purity around it but where it is synthesised into advice or judgement, other factors come into play.

INT: **So you just described very nicely there the ways in which scientists argue especially when they are within their scientific area. If they are pulled into a policy advice area where they know the purpose is to try and understand the level of uncertainty and to follow some kind transparent process, do they change behaviour?**

RES: They can select evidence or highlight uncertainty in order to pursue the hypothesis they favour. For example, think of a hydrologist who is expert in how water flows through rocks and they are asked about whether it is safe to allow fracking. If that scientist really doesn't want to see fracking perhaps because of a rationale such as "you shouldn't be using shale gas or any fossil fuels because it's polluting and I'm worried about climate change". Even although a scientist considers that in a particular environment fracking might be quite safe they may pull in evidence from other examples to allow them to say "ah yes it was demonstrated here that there was pollution of the water table by a fracking process". We should be very conscious of the selective use of evidence. So scientists, because they are human, can easily conflate things in their own mind which affects at the end of the day what they decide to say.

INT: **So how does the person that requires the policy get access to more unbiased scientific judgement?**

RES: Let me give you an example of one approach. There are two groups of reputable scientists who have diametrically opposed views about what European Union policy should be on endocrine-disrupting chemicals and I asked them to come and have a meeting with me on the topic, in the absence of policy officials (because I didn't want them trying to influence the policy officials by what they said). I set the agenda around defining what the issues were and examining the evidence to identify where there was consensus and where the disagreement lay. At the conclusion of the meeting, both sides felt able to sign up to a statement of what they agreed and where the uncertainties lay. However, two issues became clear. The first is that each group of scientists had a constituency and they were nervous that they would be seen as relaxing their views if they did not maintain their fixed views. The second issue was that both groups felt they had a role in suggesting what the policy should be based solely on the evidence they provided. At this point, I think there is value in a third party (this could be a CSA or an advisory panel with no fixed views or constituency to serve) assessing the evidence to draw out where the uncertainty lies and what the impact of that uncertainty may be. Based on the outcome of this, the policymaker can develop different policy options (possibly bringing in other non-science-based evidence such as economic impact, public opinion, etc.) each with an impact assessment. Ultimately, the politician will choose which direction policy will move and evidence will not be the only factor being considered.

Politicians find it very difficult to deal with uncertainty (and risk) in these situations. There's a famous quote from Winston Churchill saying that what he most wanted in life was a one-handed Chief Scientific Adviser because the problem with Chief Scientific Advisers is they always say "well on the one hand …." and that's not what he wanted to hear. He wanted to hear that the evidence is clear that he should do one thing or another. But most scientists are reluctant to talk in terms of certainty. Also, I think you should start getting worried if scientists or scientific advisors are setting the policy as, unlike politicians, they do not have a democratic mandate. However, society does pay for their expertise by training them so we should value the evidence and analysis they can provide.

INT: **So it's interesting that you said politicians don't like to talk about risk and they don't like to talk about uncertainty but when you look at things like the national risk register which we have in the UK that uses exactly that language of risk and uncertainty, so it does play quite a big role somewhere in government?**

RES: I think this is a slightly different issue than providing scientific advice. It is very important to have a risk register and some idea of how those risks might be mitigated. That could require direct action by the Government or could require a policy intervention, which should then be evidence-based. Quantifying risk in terms of likely impact is also crucial to allow the targeting of resources when the amount of resource is always limited.

INT: **So do things like the national risk register help start the conversation about risk and how you manage those risks?**

RES: Yes, but I can't ever remember a time where there was a public discussion around the national risk register although more recently, citizens have engaged and demanded more discussion around issues such as climate change. Climate change is an example of an issue that is very difficult for politicians because of the timescale surrounding the issue. It is a significant challenge to make difficult policy decisions that may result in substantial change for citizens in the timeframe of the electoral cycle in order to safeguard the environment for future generations. They face not getting the credit for their actions (e.g. banning the use of private cars) where the benefit will be felt 20 years down the line.

INT: **One sometimes feels that uncertainty is used to avoid a discussion around controversial issues, even when the risks are low. To support good public policy should we avoid framing issues purely in terms of risk?**

My biggest frustration in trying to drive evidence-based policymaking through the provision of scientific advice is the lack of transparency in the process. It is right that other types of evidence than scientific evidence should be used. These might include economic evidence, ethical views, political consideration, public opinion, etc. The biggest hindrance to improving the quality of evidence in policymaking is the lack of transparency about what evidence is used and what's not used.

So on the GM issue, how good it would be if in the EU and the UK ministers said "all the evidence tells us that this technology is safe and can be used and has been used safely, however, we also know that our citizens have demonstrated no demand for this technology. So given there is no demand, we will not approve cultivation". What that does is allow citizens to say "Oh, I've never wanted it but you're saying it is safe—is there any advantage to having it for me?". Then there might be conversations around reducing the climate impact of agriculture by not using any pesticides or herbicides which come from petrochemical manufacture but we could genetically modify the crops so that they were resistant to the pests in the first place so you didn't have to use the chemicals, you might call that something like GM Organic. You don't have to add chemicals but it's not a conventionally bred crop, it's gene-edited or genetically modified, that is an option, are you interested in considering that? This is the start of a dialogue and of course what you rely on is the expert judgement from the scientist to say how effective GM organic crops would be in reducing the contribution of agriculture to greenhouse gas emission.

INT: **And there must be a strong role for social science as well?**

RES: Social scientists are crucial here but not as an add on, as a fundamental part of the process.

INT: **It may be going off-topic slightly but things like the citizens' assemblies have been mooted as ways as dealing with kind of politically controversial issues, I wonder if they would have a place here?**

RES: Yes, actually it is interesting because Scotland has just introduced citizens assemblies and I am very much in favour of this. It has worked well in Ireland and in Denmark. The role of scientists would be to provide evidence and explanation to the assembly. I also think there is value in these assemblies being run at arm's length from Government as this will help the population have confidence in the debate being held. Trust in the process is very important. Evidence from other assemblies is that the wider population appreciate the time their fellow citizens are investing in the assembly and to get to the bottom of the evidence presented and they feel more open-minded to the recommendations coming from the assembly at the end of the process. I do like the idea of that we hear the citizen's voice because they bring a different view of the evidence and other values to the table.

INT: **You were discussing the fact that experts might have very different kinds of opinions about things. There is a school of thought in the expert judgement literature that says that if experts come together and they evaluate the evidence for long enough they'll eventually come round to sharing the same perspectives and even agree to the same probability assessment. Is that something that's reasonable or is the only way to get them to agree to lock the door and not give them lunch until they agree?**

RES: I've probably got more examples where people have very entrenched views about the evidence and the validity of particular types of evidence and so on. If they have an entrenched viewpoint, they are usually reluctant to be open to discussing somebody else's view of the evidence. It's probably because by the time it gets to those discussions, views have become quite polarised. If you were generating the evidence in a laboratory and there was discussion about the interpretation of the evidence, that would generally be used in a positive way and constructive tension between the two might unpick the differences and improve the analysis. But we must understand that scientists or experts are also human and have human frailties, they don't want to be humiliated, proven wrong. They may feel they have a vested interest in their evidence being used and relied upon and their expert judgement being valued by a group of citizens. You must be wary about that.

INT: **So that's very interesting because I would have expected you to say, "as a professional scientist that scientists will do their best to be as objective as they can be" but what you're saying is that in many cases actually they're swept up, they're also political animals as well, so does that…**

RES: Of course, in an ideal world you would want scientists to be utterly objective but that's hard to achieve. They've got values and sometimes those values come to the fore and might result in the selective use of evidence.

INT: **As we've discussed there are various approaches to expert judgement, some of which broadly speaking rely on creating consensus, other's which don't bother with that but which try to test how effective the judgements are of experts on similar sorts of questions. Does what you've just been saying have a bearing on the kind of approaches that you could try to use in certain types of problems, maybe the more public and maybe the more controversial ones?**

RES: It's always an interesting process trying to procure evidence about a topic. First, you might want to know if there is an established viewpoint on the issue. Let's take climate change as an example. You could ask "is there a consensus around the evidence supporting manmade climate change" and the answer would be yes. It is a useful consensus as 98-99% of scientists agree with this statement. There might be disagreements around the role of the ocean as a carbon sink or the role of cloud cover in affecting warming. So, there might be disagreement on mechanisms, etc., but the consensus is useful because it can highlight to those asking the question that we have the evidence we need to act. I would always try to identify if there is a consensus but not to force one.

INT: **If you were to get a group of different experts who have been selected may be from different scientific schools of thought do you see, as a synthesiser of expert views, that some of those groups are giving you either more useful or more reliable evidence than other ones, or is that something that's a little bit too far away from a CSA's expertise to be able to make judgement on?**

RES: Well in the EU they constitute things called expert panels, for example, in the European Food Safety Authority (EFSA) and they will look at topics such as the safety of food additives and the expert panels peer review the evidence available. The expert panels will also be asked to declare any potential conflicts of interest in order to highlight any potential for biased views. The management of these expert panels is important as I think it is useful to hear from those who may have biased views (e.g. a scientist working for a company that makes food additives in this case) but you might excuse them from any decision-making on the recommendations made. You could argue that we might all have some degree of bias but scrupulous transparency is helpful here.

INT: **Some people also use concepts of calibration and information of experts to assess the effectiveness of experts in making useful judgements of uncertainty. Calibration measures their ability to assess uncertainties, for example with 95% of outcomes contained within their 95% uncertainty bands, and information measures the relative narrowness of uncertainty bands. Do those concepts make sense or is there something that they're missing?**

RES: This approach could be useful in expert panels—where there is often a broad range of experts. You could do some analysis of how individual experts have performed with respect to how they approach uncertainty. I know this approach has been tried—although I do wonder how acceptable it is to panels of experts!

I am not sure this approach would have helped me with some of the divergent views I had to deal with. For example, in examining the impact of endocrine-disrupting chemicals we had large differences of views: some experts thought there could be catastrophic outcomes, while others thought the uncertainty wasn't significant. As CSA I needed to get them to agree on what the starting point was. This is the value of a CSA and helped to explain why their views diverge. It then enabled us to talk about the likelihood of the scenarios and how we mitigate the consequences.

However, I recognise that there can be different "personalities" of an expert. I've been in situations where there has been a vocal and sometimes antagonistic person that doesn't listen to others. I don't find that useful as part of an evidential framework. You need experts who are challenging and are willing to be challenged.

INT: **In Cooke's method these performance measures are used essentially to downweight experts who are less good (according to those measures) at assessing uncertainty. The resulting weighted average is called the rational consensus uncertainty estimate. Is that a useful summary of the expert evidence?**

RES: It could be. I always thought it was valuable to find where the weight of evidence lay, but it was also necessary to advise the government minister of the full spectrum of opinions, so they also appreciated if there was a range. If I could explain why a particular expert was underperforming then we could point this out and say that we might want to take this into account when placing weight on that view.

But in practice, other factors might contribute to what evidence a minister might use. They might give weight to an opinion which they liked to hear, and they often prefer to use uncertainty around the science to avoid the decision they don't like. It's very difficult to contradict this argument (the one referring to uncertainty), although it can be an abuse of the science and lead to questionable outcomes.

Experts help us by giving us a degree of confidence in a policy that delivers the desired outcome—this is where we need to get to. Policies aren't perfect as the evidence can change over time. For example, with EU biofuel policy—the incentive to grow crops for biofuels had an unintended consequence of land traditionally used for edible crops being taken out of that use. The real trick is to develop policies that are resilient and open to additional evidence that might become available with time.

INT: **The Precautionary Principle also influences the way that policy is developed from the scientific assessments. How has this affected the process?**

RES: When it was first introduced in the European Commission it was intended to allow us to take advantage of the developments of new technologies while minimising the risk from them. But it got reinterpreted over time as a brake on the use of new technologies. The challenge is to develop a policy that allows you to take advantage of new technology while taking account of the risks and to evolve that policy as time progresses and more evidence becomes available.

INT: **You've talked around a wide range of the kind of things that come up on the plate of a chief scientist. I don't know if this is possible but is there a percentage estimate you could give of the proportion of questions which come to the chief scientist which are potentially amenable to the use of these expert judgement methods, is that large or small or does the question not… is it too difficult to answer?**

RES: I think it is probably too hard to say because I think in some instances where you might seek consensus because you can see that there is perhaps very little spread of opinions and you are trying to make it simpler for the recipient of the advice. Then in other areas, you need to be very clear about the level of uncertainty so not seek consensus as that would be doing ministers a disservice if you tried to synthesise a consensus view for them because that could easily be undermined at a later stage and would undermine the minister's confidence in the evidence. You need to be able, to be honest about where there is significant uncertainty and then help a minister to both understand the extent, and the impact, of the uncertainty. In general, I feel that the use of expert judgement methods might have the greatest value in the step before a CSA gets involved and it is for the CSA to translate the output of the expert panels.

INT: **…and so there is also the issue of how grave the impact is?**

RES: I've mentioned the importance of the impact of uncertainty a lot and that is because sometimes the evidence has a high degree of uncertainty but it really doesn't matter because the impact, if you get it wrong, is quite low. So, for example, with endocrine-disrupting chemicals, you might have a consensus suggesting that particular uses represent a low risk but to some receptors (e.g. a developing foetus) the impact would be enormously high. Although the risk overall may be judged as low the potential harm could be very high. Ministers need to know this as it will likely impact the decisions they may make.

A Chief Scientific Adviser or a scientific advisor in many ways is a translator of evidence. That is a key part of the role because you should be able to understand the language of your peers, but you need to translate that into something meaningful for the person that is asking for your advice. There is always the danger that you lose some nuance or sophistication on the way in order to make it accessible.

I'm sure you get good Chief Scientific Advisers and not so good ones. Like everything in life, there will be people who will be particularly good at talking things through, but most of them particularly if they're independent, e.g. not part of the civil service, will be as truthful as they can be and will not fear to tell a minister something they really don't want to hear. It's much easier for a Chief Scientific Adviser to provide unwelcome evidence to a minister than it is for an official, because an official works for that person, they have a career, whereas for a Chief Scientific Adviser their main focus should be on working to be true to the evidence and not considering how it's going to be received.

I used to meet with all the UK Chief Scientific Advisers on a regular basis when I was CSA for Scotland. We did not discuss much expert judgement; we talked about what advice was being asked for and where evidence was likely to be sought. We also discussed how evidence could be brought to the fore when the policy was being developed. We didn't test ourselves in the same way or ask each other the questions that you're asking me now.

INT: **Yes, and there are two possible explanations, one is that these kind of structured expert judgement methods are just simply not known and the other explanation is that they are known a bit but they are either too difficult or too expensive to apply.**

RES: I think some of us did know about them but there is a broad spectrum of approaches to being a Chief Scientific Adviser. Some are incredibly collegiate and draw from a very wide pool of expertise. Others feel they can make expert judgements themselves without substantial input. In my view, the latter approach is a weaker one.

INT: **As a CSA there are probably times when you have had responded in the heat of a crisis—how do manage to respond at speed while maintaining a scientific perspective?**

RES: An example is when the Icelandic volcano Eyjafjallajökull erupted, I got a very anxious phone call from Alex Salmond who was First Minister of Scotland at the time. He said "a volcano's erupted in Iceland, volcanoes erupt all the time but there's a complete closure of air space in Scotland and that has big impacts for citizens trying to travel and for business. Why is this happening?" My reaction was to say I don't know but I'll find out because that is a perfectly reasonable question. I spoke to the NERC British Geological Survey who have expert volcanologists who understand about volcanic eruptions and they explained why the closure of airspace was necessary for this instance. I asked if theirs was the consensus view or would anyone disagree. Once I understood the information and that there was a high degree of confidence in it, I was able (within a couple of hours) to get back to the First Minister to explain what was happening and why. I didn't provide him with the detailed technical information but rather "modern jet engines are designed to burn fuel efficiently above their melting point. The reason they don't melt is that they are engineered with fine capillaries through the engine so when it's moving forward at speed, the air cools the engine. If you put glassy dust which is being expelled from Eyjafjallajökull into those capillaries it melts and blocks the capillaries after a period of time. The engine will melt and fall off the aeroplane and that's the reason that we've stopped aeroplanes flying in Scottish airspace" and he said, "OK that's fine"

My approach there was not to get a consensus view. It was to look at the rationale that people were using in order to ban flights in Scottish airspace. The plumes of dust from the volcano had been mapped. We knew that there would be a residence period for any plane within that dust and then work out what impact that would have on a modern jet engine. An initial precautionary approach was being adopted because potential consequences were severe. It was also accepted that restrictions might be relaxed as more evidence became available.

INT: **Yes, but of course there are other issues which are highly contested and more controversial and where there are ranges of opinion.**

RES: And even in the case of the volcanic eruption, propeller aircraft took air samples and this allowed experts to refine what the average concentration of dust particles was, the movement of the plume of dust was modelled and flight paths could be proposed that would avoid 95% of the glassy ash and calculations were that that might reduce air cooling if you did go through by a certain percentage but that it wouldn't melt. It would damage the engine.... and of course, the damage could be cumulative if you kept on flying the same engines through that over a period of time. What is important is to be able to make evidence-based decisions quickly as ministers need to act and to communicate about the issue.

INT: **Its clearly a challenging role, so is there a course of training for a chief scientist?**

RES: No, I suspect the training is being a scientist because all of us do it, you do it whether you're in big collaborative projects or you're in peer review panels where you're having to discuss evidence and you're having to weigh up the different scenarios… your life as a research scientist probably is the training to allow you to facilitate the prioritisation or use of evidence in a case like that particularly quickly. You've got time to be much more cerebral as a CSA and it's a very small proportion of a scientific advisor's role to respond to an emergency like that. I think it only happened to me a couple of times in the Scottish Government and only once at the European Commission.

Chapter 13
Characteristics of a Process for Subjective Probability Elicitation

John Quigley and Lesley Walls

Abstract The elicitation of subjective probabilities from experts can be critical in determining a course of action when making decisions under uncertainty. A sound process to elicit probabilistic judgement is necessary to ensure that good quality data are used to inform the decision-making, as well as to provide protection to those accountable for the consequences of the determined actions. We synthesise the characteristics of a good elicitation process by critically reviewing those advocated and applied. We compare the processes inherent in the guidance produced by two professional bodies to exemplify the manner in which the characteristics manifest themselves in practice. We examine whether standardisation is meaningful given the maturity of processes for the elicitation of subjective probability.

13.1 Introduction

Big data and the digital age have not removed the need for nor diminished the importance of expert judgement; observed data are history and expert judgement is the future. We still require expert judgement to support decisions where observed data are few or non-existent. Also we can require expert judgement in situations where observed data are abundant since the relevance of the past to the future can be assessed only with expertise (Hora 2007; Quigley and Walls 2018). This is not likely to change as more observed data are collected. Further, for situations with little or no observed data, we believe that concepts like *black swans* (Taleb 2007), *perfect storms* (Junger 1997) or *deep uncertainty* (Cox 2012) should not be an excuse for superficial thinking about possible future events (Dias et al. 2018). We encounter many problems where there exists relevant expertise and for which the problem characteristics are

J. Quigley (✉) · L. Walls
University of Strathclyde, 130 Rottenrow, Glasgow G4 0GE, Scotland
e-mail: j.quigley@strath.ac.uk

L. Walls
e-mail: lesley.walls@strath.ac.uk

A. M. Hanea et al. (eds.), *Expert Judgement in Risk and Decision Analysis*,
International Series in Operations Research & Management Science 293,
https://doi.org/10.1007/978-3-030-46474-5_13

measurable in theory but not in practice; these conditions are ideally suited for expert judgement (Cooke and Goossens 2008).

We are concerned with the elicitation of quantitative subjective judgement, specifically the expression by experts of their beliefs in the form of subjective probability distributions. Such measures do not come naturally to people and so we require a process to facilitate the expression. Research has indicated that there is a need for formal elicitation to extract and quantify judgements since people, even experts, are unable to provide accurate data simply on request; see, for example, Cooke (1991), Meyer and Booker (2001).

Since the work of Tversky and Kahneman (1974), there has been awareness of the biases and heuristics people apply in decision-making under uncertainty that can result in poor probability assessments. Examples include contextual biases and heuristics such as anchoring, availability and representativeness. Other challenges associated with assessments made by people include issues such as groupthink (Janis 1971), group polarisation (Myers and Lamm 1976), overconfidence (Soll and Klayman 2004) and difficulties associated with communicating knowledge in numbers and probabilities (Gigerenzer and Edwards 2003). Inappropriate and ill-informed elicitation can amplify biases by relying on subjective and unreliable methods for selecting experts (Shanteau et al. 2002), asking poorly specified questions (Wallsten et al. 1986), ignoring protocols to counteract negative group interactions (Janis 1971) or applying subjective or biasing aggregation methods (Aspinall and Cooke 2013; Lorenz et al. 2011).

An elicitation process design should address these known issues. However, according to Burgman (2004), Krueger et al. (2012), Kuhnert et al. (2010) and Regan et al. (2005), amongst others, informal methods for elicitation persist. French (2012), Choy et al. (2009) and Krueger et al. (2012) report that few elicitations provide sufficient detail to enable review and critical appraisal. The consequences of poor judgement are misinformed decision-makers as illustrated by Wilson (2017) who reported a 52% hit rate in 95% intervals from his investigation of selected expert judgement studies.

Reported expert probabilistic assessments have been conducted for almost 50 years. Early studies include WASH-1400 concerning nuclear reactor safety (United States Nuclear Regulatory Commission 1975) that applied methods further developed into NUREG-1150 (United States Nuclear Regulatory Commission and others 1990). Approaches to elicitation continue to be developed and expert probability assessments remain a key input to policy and decision-making today. Examples include the following: Determination of volcanic eruption-related fieldwork risks (Christophersen et al 2018); pollination uncertainty to inform policymaking for ecosystems (Barons et al. 2018); the combined effect of the meteorology and oceanography (also known as metocean) in offshore engineering (Astfalck et al. 2018); assessment of technology uncertainty during aerospace product design (Hodge et al. 2001); role of technical expert panels in probabilistic seismic risk analysis (Budnitz et al. 1998); expert judgement underpinning influential global environmental policies (Hemming et al. 2018) such as International Union of Conservation

Nature (IUCN) Red List (IUCN, 2012) and Inter-governmental Panel on Climate Change (IPCC) Assessments (Mastrandrea et al. 2010).

A sound process to elicit judgement for such problems is necessary to inform the decision-making. In addition, a sound process can also provide protection to those accountable for the consequences of the determined actions. Consider the 2009 L'Aquila earthquake tragedy in Italy where 309 lives were lost (for details, see Nature 2011 and Science 2012, 2014). This case highlights the need for transparent, rigorous and widely accepted processes for assessing uncertain events. On appeal in November 2014, only the government official continued his prison sentence as the responsible person for the risk communication, while the six scientists who provided expert advice were acquitted. Nevertheless, in the original trial in October 2012, the six scientists as well as the government official, who participated in Italy's National Commission for the Forecast and Prevention of Major Risks six days prior to the earthquake, were sentenced to six years in prison for manslaughter. The prosecution argued that the expert advice from the Commission resulted in 30 people deciding to stay indoors contributing to their death; the scientists were brought to trial originally because of poor practice and the presiding judge ruled their analysis superficial. Additional criticism of this risk assessment has been made by the President and General Secretary of the International Seismic Safety Organisation (ISSO) concerning the lack of independence amongst expert judgements (Martelli and Mualchin 2012).

Our contribution is to characterise more generally what makes a good elicitation process by critically reviewing relevant literature and reported applications. Our intent is to inform others responsible for developing future elicitation processes for specific purposes and contexts of the characteristics of a good elicitation process. Section 13.2 describes the seminal work of the Stanford Research Institute (SRI) in constructing an interview process to address a variety of known biases commonly encountered with the elicitation of subjective probabilities from experts. Section 13.3 extends our discussion of the issues surfaced in the previous section through a wider review of subsequent work and organises these issues into the emergent characteristics of a good elicitation process. Section 13.4 compares the elicitation guidance documents from two professional societies to illustrate and assess how the general characteristics manifest themselves for different purposes. Section 13.5 explores whether standardisation of an elicitation process for subjective probability is useful given the maturity of practice. Section 13.6 presents our concluding discussion.

13.2 Stanford Research Institute Elicitation Process—The Genesis

Although the RAND Corporation developed formal approaches for the elicitation of expert judgement in the 1960s, these were not for probabilistic judgements (Dalkey and Helmer 1963; Dalkey 1967, 1969). Spetzler and Stael von Holstein (1975) were the first to report an elicitation *process* for subjective probabilities grounded in practice by the Decision Analysis Group at the Stanford Research Institute (SRI). Previously, research had focused upon methods to encode probability assessments that concentrated more narrowly on the quantification of expert uncertainty rather than wider processual considerations; see, for example, Hampton et al. (1973) and Winkler (1967). By broadening the scope to position encoding methods within an elicitation process, the so-called SRI five-stage approach seeks to identify potential biases and minimise their impact on the quantitative assessment. Since the SRI process is concerned with structuring an interview with the expert, some important aspects, such as expert selection, are not considered. We describe the five stages—namely, *motivating, structuring, conditioning, encoding* and *verifying*—since these provide a common basis that has informed many subsequent elicitation processes.

13.2.1 Motivating Stage

The SRI advocates that the process design should address motivational biases, such as management and expert bias. Management bias occurs when an expert provides goals rather than judgement. For example, an expert states the aspiration that there will be no weaknesses in a system by time of manufacture, rather than providing an assessment of their beliefs about the likely occurrence of weaknesses. Expert bias occurs when a person becomes overconfident merely because they are called an *expert*. During this stage, the intent is to determine if there are motivations for the expert to, consciously or unconsciously, adjust probability assignments based on perceived rewards.

Motivational biases can be identified through discussion, where the interviewer develops a rapport with the expert and discusses openly any payoffs that might be associated with the probability assignment as well as possible misuses of the information; for example, single-number predictions are often interpreted as commitments. The interviewer should make clear to the expert that no commitment is inherent in a probability distribution, and that complete judgement from the expert is sought. Additionally, the interviewer introduces the encoding task to the expert by explaining both the importance and purpose of probability encoding in relation to the decision, and clarifying the difference between probabilistic and deterministic predictions.

13.2.2 Structuring Stage

Structuring involves defining the event under consideration to minimise ambiguity in the questions and to explore how an expert thinks about the quantity for which probability judgement is to be elicited. The aim is to manage possible cognitive bias by simplifying the complex task of assigning probabilities by disaggregating the quantity of interest into more elementary variables (Armstrong et al. 1975). However, the unpacking principle (O'Hagan et al. 2006), also known as subadditivity (Tversky and

Koehler 1994), may be the consequence. This refers to the situation where the more detailed the description of the event, the greater the likelihood assigned to it. For example, an expert may provide an assessment for the probability of a component failing and subsequently during the elicitation process provide probabilities associated with causes of that failure that may result in a component probability exceeding the initial assessment.

The quantity of interest needs to be specified so that a measurement scale can be determined. There is a need for precise thinking about how the quantity of interest will be realised. For example, if the exchange rate between two currencies next year was to be assessed, then it would be necessary to specify the exact time next year for the measurement as well as where the currencies would be exchanged, as banks, stock exchanges and tourist agencies all buy and sell currencies and offer different rates.

It is important to choose a scale that is meaningful to the expert. One important consideration when selecting the quantity of interest is feedback to the expert. This is considered crucial for calibrating the expert and should be event-specific (Fischhoff 1989; Bolger and Wright 1992; Ferrell 1994). In other words, the feedback must be with respect to assigning probabilities to particular classes of relevant events and not only feedback on the ability of the expert to assign probabilities to any situation. To increase the effectiveness of feedback in terms of learning, conditions that influence the event should re-occur as often as possible (Fischhoff 1989; Kadane and Wolfson 1998). Therefore, the factors on which the measure is conditioned should be as few and general as possible. The structure of the quantity of interest may need to be expanded so that the expert does not have to model the problem further before making each judgement.

Structuring should encourage the expert to think carefully about the event before the actual encoding session begins by probing and clarifying issues concerning relevant and irrelevant background information.

13.2.3 Conditioning Stage

Information relevant to assessing the probabilities is discussed to address issues such as availability bias and anchoring and adjustment. Availability bias refers to the influence that easily recalling examples can have on the assessment of probability, such as overestimating the likelihood of a disaster because it has devastating consequences unrelated to its frequency. Anchoring and adjustment is a heuristic, where people base their judgement on a piece of information (i.e. the anchor) and adjust for the assessment. For example, an expert making a series of assessments will provide an initial assessment for the first quantity of interest, and all subsequent assessment will be adjustments; anchoring and adjustment can lead to overconfidence and other judgement errors (Kahneman et al. 1982). Such discussions can also form part of the structuring stage for these and other possible biases such as the conjunction fallacy, whereby people guess that the odds of two or more events co-occurring are greater than the odds of any one of the events occurring alone because the co-occurrence appears more representative (Tversky and Kahneman 1983).

The conditioning stage aims to encourage the expert to think fundamentally about their judgement, understand how they make probability judgements and through revealing the information that seems most available, what (if any) anchors are used and what unstated assumptions are being made. Experts can be asked to specify the most important bases for their judgement to identify anchors as well as exploring more extreme situations.

13.2.4 Encoding Stage

This stage refers to the actual method for elicitation of the probability distribution for the quantity of interest. A popular encoding procedure for distributions is the fractile method where the expert assesses the median value of their subjective probability distribution along with, say, the (25th,75th) and the (5th, 95th) percentiles. The order in which these quantities are elicited should start with the extreme values first and progress towards the central values to avoid a central bias (Seaver et al. 1978).

After percentiles of the distribution have been assessed, graphical techniques can be applied to enhance the quality of the distribution (Chaloner et al. 1993). Once these probability values have been elicited, then a parametric distribution might be sought to maximise fit.

Spetzler and Stael von Holstein (1975) provide the following suggested steps in encoding.

(a) Ask for extreme values—deliberate use of availability to counteract central bias.
(b) Ask for scenarios that lead to realisations beyond extreme values—makes outcomes more available to an expert so more likely to assign higher values to extreme outcomes to address central bias.
(c) Assign probabilities to scenarios—increases variability in overall distribution.
(d) Choose values and assign probabilities—do not choose a significant value for assessment as this will lead to anchoring, but choose values experts will be comfortable with assessing.
(e) Construct Cumulative Distribution Function (CDF).
(f) Fit curve.

If the expert is to assess multiple quantities of interest, then in step (d) it is recommended the probabilities (i.e. percentile) are fixed and the values elicited so that numbers are not provided upon which an expert might anchor.

13.2.5 Verifying Stage

Since a subjective probability distribution has been elicited, the interviewer now guides the expert though a review of the distribution to ensure it reflects his/her expressed belief. If it does not, then additional elicitation is required. Verification is accomplished by showing the expert the implications of the interviewer's interpretation of their response.

Two common activities performed at this stage include visualising the probability density function and comparing equi-probable intervals from the CDF. Asking which interval the expert would bet on supports verification because the expert should be indifferent to betting between intervals if the CDF reflects their belief. It is suggested this activity should be repeated three to five times.

Verification is required to ensure that an expert has provided a reflection of his/her true beliefs. If problems are encountered, then the previous stages are to be repeated.

13.2.6 Extensions of the SRI Process: Aggregation and Discretisation

Miley Merkhofer, manager of the Decision Analysis Research Program at the SRI between 1975 and 1983, reported an extended SRI process that included a sixth and seventh stage, namely, Aggregation and Discretisation, respectively (Merkhofer 1987).

For situations where multiple experts are assessing the same quantity of interest, then individual probability distributions may need to be aggregated; evidence suggests that combined judgement can improve assessment quality (Ashton and Ashton 1985). There are two approaches to aggregation—mathematical and behavioural. The former implies the experts should not influence each other's decisions (Ferrell 1985). The latter requires experts to share their judgement and re-assess their distributions and includes techniques such as Delphi (Ferrell 1985) and Nominal Group Technique (Moore 1987), see Gosling (2018). There are several mathematical approaches to aggregation most of which aim to evaluate a weighted average across the experts. See Cooke (1991) for a fuller discussion as well as more recently developed methods such as Wisse et al. (2008).

Rather than encoding using, say, the fractile method, it can be necessary to treat continuous random variables as discrete. Discretising refers to techniques for fitting continuous distributions to the elicited data while preserving important moments. This is accomplished by dividing the range of all possible values for the uncertain variable into intervals, selecting a representative point from each interval and assigning that point the probability that the actual value will fall within the corresponding interval. The moments can be preserved through, for example, Gaussian quadrature techniques (Miller III and Rice 1983).

13.2.7 Managing Bias

Throughout the SRI process, the interviewer explores the potential for bias with the expert and takes steps to manage bias by careful consideration of issues as discussed above.

At the time of development of the SRI in the 1970s and 1980s, such consideration of expert bias was the only approach. During the 1990s, Roger Cooke introduced a more formal means of measuring bias where seed variables are used, and experts assess quantities that are unknown to them but known to the analyst; see, for example, Cooke (1991), Quigley et al. (2018).

13.3 Characteristics of an Elicitation Process

We now synthesise the characteristics of good elicitation processes based on a review of two classes of publications. First, proposals from the scientific literature, which describe largely positive attributes we seek an elicitation process to possess. Secondly, insights gained from published criticisms of practical applications, which are indicative of pitfalls to avoid when designing and implementing an elicitation process.

A variety of literature sources have been drawn upon including books, studies, critical reviews as well as journal articles. Books focusing on expert judgement include Cooke (1991), Meyer and Booker (2001), Ayyub (2001) and O'Hagan et al. (2006). The U.S. Nuclear Regulatory Commission has published several relevant documents. These include NUREG-1150 (United States Nuclear Regulatory Commission and others 1990) which reports how to estimate the uncertainties and consequences of severe core damage accidents in selected nuclear power plants for which Keeney and Von Winterfeldt (1991) provide a critical appraisal. NUREG/CR-6372 (US Nuclear Regulatory Commission 1997) provides guidance on the use of expert judgement for seismic hazard analysis and is accompanied with practical guidance from NUREG-2117 (Kammerer and Ake 2012); lessons from more recent application of these guidelines are discussed in Siu et al. (2015). Cooke and Goossens (2008) review various elicitation applications, while Shephard and Kirkwood (1994) provide an in-depth description of an elicitation case study. Walls and Quigley (2001) describe how the SRI model informed an elicitation process for assessing uncertainty in product development, for which Hodge et al. (2001) reflect upon the lessons learnt from the perspective of multiple participant roles. Additional references are given to the literature in relation to specific issues discussed below.

We acknowledge that a specific situation will require a particular elicitation process. Figure 13.1 summarises some questions likely to be asked by any analyst approaching elicitation of subjective probabilities and indeed is based on the questions the authors themselves posed in such a situation. Within the diagram, the specific questions are grouped around the who, what and how of approaching an elicitation process. Here, we seek only to discuss general characteristics of good practice that will be transferable across a variety of problem domains where subjective probability assessment of some quantity of interest is required.

Figure 13.2 illustrates our structured collation of the issues emerging from the literature. We show that the *process* should be grounded in scientific *principles*, while taking account of the *purpose* of elicitation that determines the quantities of

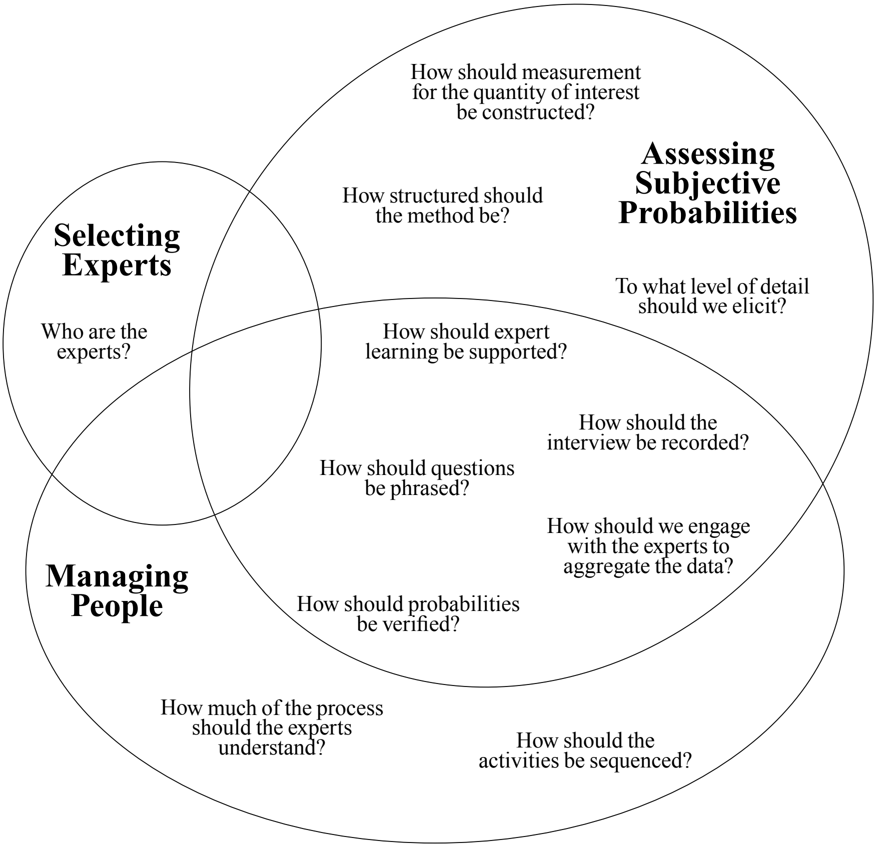

How should measurement for the quantity of interest be constructed?
Assessing Subjective Probabilities
How structured should the method be?
Selecting Experts
To what level of detail should we elicit?
Who are the experts?
How should expert learning be supported?
How should the interview be recorded?
How should questions be phrased?
How should we engage with the experts to aggregate the data?
Managing People
How should probabilities be verified?
How much of the process should the experts understand?
How should the activities be sequenced?

Fig. 13.1 Example questions posed by an analyst when designing an elicitation process

interest for which *probabilities* are to be elicited. The *people* whose assessments will be elicited will be sourced from the *purpose* context, and their expertise will depend on the quantities for which *probabilities* are required. The inner design, plan and implement loop reflects that an elicitation is not necessarily a linear process. Our following discussion reflects the inter-dependency between issues to be considered when developing a good elicitation process.

13.3.1 Principles

Cooke (1991) proposes that expert judgement processes should be subject to the following principles.

Scrutability/accountability: All data, including experts' names and assessments, and all processing tools should be open to peer review and the results must be reproducible

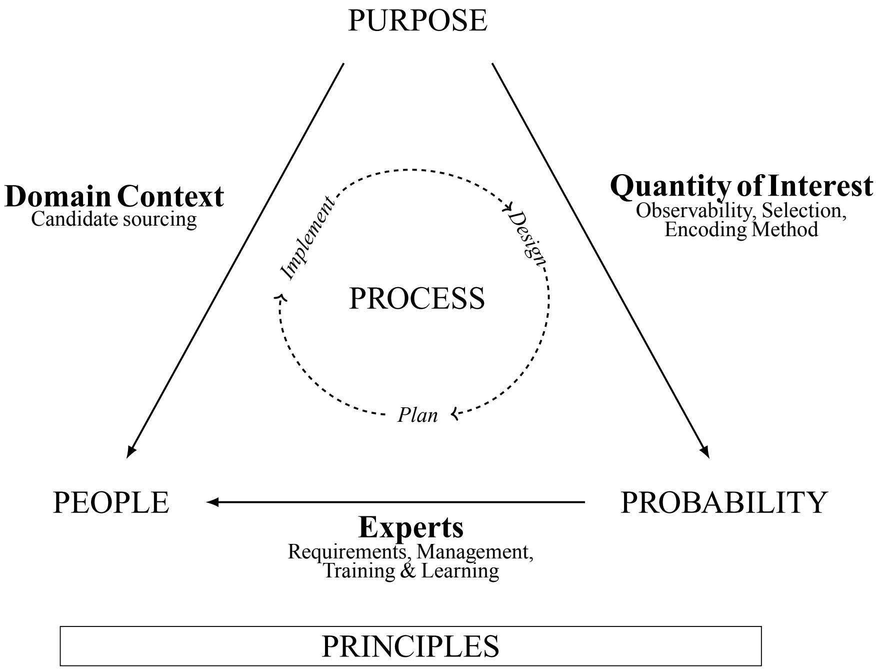


Fig. 13.2 Key characteristics of a good elicitation process emergent from the literature

by competent reviewers. It is not sufficient to present synthesised summary measures of expert assessments only and all subjective probabilities should be traceable back to the individual expert. Cooke (1991) argues strongly for publishing the names of experts for public decision-making but acknowledges the potential disadvantage in relation to conflict of interest for private firms.

Empirical control: Expert assessments should be susceptible in *principle* to empirical control. Scientific statements should be falsifiable in principle and, while such a test may not be feasible, it should be possible. Essentially this principle is guarding against an expert being free to say anything, and inferring one subjective probability is as good as another.

Neutrality: The method of elicitation should encourage experts to state their true opinions. Cooke (1991) suggests the Delphi technique, as reported by Sackman (1975), is an example where experts are encouraged not to deviate too far from the median of the group, as well as a providing a process where experts are required to self-assess their judgement implying there is no incentive for honesty (see Brockhoff 2002).

Fairness: All experts should be treated equally a priori. Note that this does not prevent unequal treatment of experts a posteriori. In contrast to the fairness principle, some

Bayesian methods for combining expert judgement require the decision-maker to assess the reliability of an expert prior to the elicitation. Further, there is a lack of guidance upon what basis such an assessment can be made. Even if such guidance did exist, then the fairness principle would be violated.

13.3.2 Purpose

According to Kammerer and Ake (2012) and the US Nuclear Regulatory Commission (1997), the purpose of an elicitation study is to provide a representation of the centre, body and range of views of the informed expert community regarding the quantity of interest. The output of an elicitation process involving multiple experts is not consensus but integration, since there is no one correct answer. The elicitation leads to the construction of what has been termed by Siu et al. (2015), as a community probability distribution.

The purpose of a process is more than just facilitating the elicitation of structured expert judgement while minimising biases in the resulting assessments. Importantly and additionally, the process should enable judgements to be subject to review and critical appraisal (French 2012). This need for a process to be transparent and repeatable means that documentation is a key enabling activity. It is the recording of the goal of the exercise, the design of the process and the judgements obtained to an appropriate level of detail and clarity that enables the process to be repeated (Cooke 1991).

Documentation has a variety of purposes (Bonano et al. 1990), including improving decision-making, enhancing communication, facilitating peer review, avoiding biases in judgement, unambiguous identification of the current state of knowledge and providing a basis for updating. A key strength of formal expert elicitation is in the documentation of the complete process as well as of the elicitation results and reasoning (Keeney and Von Winterfeldt 1991).

13.3.3 Probability Assessment of the Quantity of Interest

We consider a set of issues relating to the definition of meaningful quantities of interest for which probability assessments are required and the nature of modelling choices to be made in relation to how such assessments are obtained.

13.3.3.1 Observable Quantity of Interest

Many models contain parameters that are both unobservable and uncertain. The variable to be elicited should be related to an observable quantity, at least in principle. Cooke and Jager (1998) and Frijters et al. (1999) examine how to accomplish this.

For a relative frequency, for example, a large virtual population could be imagined and appropriate random selections considered. This should assist in ensuring use of a consistent definition for the quantities of interest (Keeney and Von Winterfeldt 1991).

13.3.3.2 Selection of Quantity of Interest

While an audit of available data may preclude the necessity of quantifying some variables, care is needed when assessing the relevance of the data (Siu et al. 2015). It can be useful to remind the expert in situations where empirical data are available, say test data, that judgement is needed to assess the relevance of that data to the practical field situations of interest. Mechanical processing of empirical data without consideration of its relevance towards the specific conditions under consideration may not result in appropriate representations of uncertainties.

Quantities of interest can be organised into similar groups to help reduce the cognitive burden by minimising the number of mental models required in assessment (Quigley and Walls 2010). However, this could introduce the bias of anchoring, so care should be taken.

Resources will be constrained and elicitation can be time-consuming; hence, better planning can result in better data. Careful expression of judgement can be a fatiguing process for an expert, especially when relevant data are sparse (Siu et al. 2015). The number of quantities of interest that an expert can meaningfully quantify in a session is limited (Keeney and Von Winterfeldt 1991); hence, it might not be possible to quantify each, and so screening the variables may be necessary (Bonano et al. 1990). Consideration should be given to the number of parameters that can be realistically assessed so that they can be prioritised for judgement. Alternative strategies will need to be considered for the remaining variables, informed by their importance towards the decision-making and the associated range of uncertainty. Bonano et al. (1990) suggest three experts being involved in parameter selection: specialists with subject matter knowledge, generalists with expertise in modelling and experts in sensitivity analysis.

13.3.3.3 Method for Encoding Probability

Requesting experts to provide their estimates in the form of a set of pre-designated quantiles can protect the expert against anchoring to the provided values as well as creating some consistency across questions that may lead to efficiencies in assessments. However, providing such judgements requires a degree on introspection and many experts do not think naturally in terms of the quantiles (Siu et al. 2015). Processes should remain flexible in accepting expert input in the form the expert feels most representative of their beliefs. When there are multiple experts, then compromise might be required.

There are important advantages in using parametric probability distributions to represent expert judgement. Three advantages, in particular, are intrinsic smoothing of the expert assessments, interpolation between assessments and extrapolation beyond the assessments. However, care is needed to avoid force-fitting a parametric model to expert assessments (Siu et al. 2015). A parametric probability distribution, such as the Normal distribution, provides an infinite range of probabilistic assessments and, typically, an expert only provides a few assessments. Given that the implications of a parametric model choice may not be apparent to an expert, then it is important that the elicitation process should include activities to check model adequacy for the expert assessment. It is particularly important that the analyst checks that the probability distribution elicited is a good fit in the part of the function (e.g. the tails) that will drive the decisions.

Considering only a limited number of parametric models to represent an expert's belief should be avoided. For example, situations where multi-modal distributions (representing the possibility of distinct, competing "models of the world") accurately represent the state of knowledge can be missed if the elicitation process focuses only upon a limited number of parametric models, especially the common uni-modal distributions. Therefore, the elicitation process should ensure that any mathematical representations of probability assessments do not unduly distort an expert's beliefs for the sake of convenience.

13.3.4 Managing the People Participating in the Elicitation

We now characterise the different roles of participants in the elicitation process and, in particular, discuss issues relating to the management and training of experts.

13.3.4.1 Classes of Participants

Each participant involved in the elicitation process should be clear about his/her role, the aims of the exercise and should be made aware of how his/her judgements will be used, i.e. how their data will inform particular decisions (Siu et al. 2015).

Who is an expert? Multiple definitions exist, with many focusing upon expertise, typically gained through experience in a particular field. For example, Ferrell (1994) defines an expert to be

> a person with substantive knowledge about the events whose uncertainty is to be assessed

while Meyer and Booker (2001) define an expert as

> a person who has a background in the subject matter at the desired level of detail and who is recognised by his/her peers or those conducting the study as being qualified to solve the questions.

These definitions implicitly assume experts can be accessed if required. When gathering experts from a constrained pool, then the definition by O'Hagan et al. (2006) might suffice since

> an expert may, in principle, just mean the person whose judgements are to be elicited.

In addition to having expertise in the domain problem, we require an expert who can express their uncertainty accurately as a subjective probability. Being able to accurately assess uncertainty is not the same as being a subject matter expert. One may know less, but be more capable of expressing this degree of uncertainty quantitatively. Hora and Von Winterfeldt (1997) suggest six criteria for identifying experts: tangible evidence of expertise; reputation; availability and willingness to participate; understanding of the general problem area; impartiality; and lack of an economic or personal stake in the potential findings.

In order to elicit a wide spectrum of judgement, we may use a group of experts with diverse knowledge that encompasses all facets of scientific thought on a particular problem. This should help to identify areas of interest that may be missed with a small group of experts or with a group of experts from a specific school of thought. See, for example, Hogarth (1978), Clemen and Winkler (1986), Broomell and Budescu (2009), Larrick et al. (2011).

For further details, see Bolger (2018) who provides a detailed consideration of experts and their selection.

Beyond the expert, there are other participants in the elicitation process. Bedford et al. (2006) identify two additional roles of decision-maker and analyst. The decision-maker is the problem owner and the one who is ultimately responsible for any decision and wishes to be informed of the uncertainties that exist. The analyst is the person responsible for identifying the necessary experts, the events of interest and developing the elicitation protocol. Others describe similar additional roles. O'Hagan et al. (2006) make a distinction within the analyst role between that of a facilitator and a statistician. The facilitator manages the interaction with the experts and should be an expert in facilitation, while the statistician is an expert in probability and gives training to the experts, validates the results and provides feedback. However, O'Hagan et al. (2006) state that these roles can be merged. Booker and McNamara (2002) also identify the role of advisor-expert. An advisor-expert is someone who helps to support experts by offering technical support. This support may be in the area of identifying the appropriate experts or areas of interest that we wish to elicit judgement about. Within professional guidance, NUREG/CR-6372 identifies three different expert roles. A resource expert to present data, models and methods in an impartial manner. A proponent expert as an advocate for a specific model, method or parameter. An evaluator expert who will objectively examine available data and models, challenge technical bases, underlying assumptions and, where possible, test the models against observations.

13.3.4.2 Managed Experts

Assessments based on the aggregation of multiple experts' judgements are reported to be more accurate than predictions based on an individual's judgement (e.g. Page 2007; Soll and Larrick 2009). Therefore, we need to consider how experts should be managed, particularly their interaction. That is, should experts communicate and, if so, how?

The experts (as representatives of the informed technical community) will evaluate the available evidence (e.g. numeric data, models, theories and scientifically accountable positions) to inform their judgements. The selection of experts is to ensure a breadth of the collective state-of-knowledge. The extent to which experts should discuss the assessment of a quantity of interest varies by approach. Behavioural approaches such as Gosling (2018) advocate a facilitated discussion to arrive at the community probability distribution, while performance-based approaches as described in Quigley et al. (2018) propose experts form their assessments independently. Additionally, others propose hybrid approaches (Hanea et al. 2018; Hemming et al. 2018).

There is evidence in some contexts (Siu et al. 2015) that experts can be reluctant to quantify their beliefs as well as to share with fellow experts. Therefore, a process needs to consider the socio-technical nature of elicitation to help put experts at ease, thereby encouraging them to openly share their point of view, even if it is not shared by others. Challenges to proponent positions are important to enhance a group's understanding but need to be managed carefully.

When managing experts, there are three "i's" to consider when designing the process activities (Siu et al. 2015).

(a) *Independence*—judgement should be based upon an individual's expertise; judgement should not be influenced by the organisation that the expert represents.
(b) *Interaction*—if a behavioural aggregation approach is undertaken, then the process of evaluation, elicitation and integration is achieved through interaction amongst experts.
(c) *Integration*—the process should emphasise integration (rather than consensus) of individuals' interpretations or judgement.

The advantage of a performance-based approach is the ability to discriminate between the quality of the experts' quantitative judgement through testing during the interview; this is not possible in a behavioural elicitation workshop.

13.3.4.3 Training and Learning

It is acknowledged by, amongst others, Bonano et al. (1990),Keeney and Von Winterfeldt (1991), US Nuclear Regulatory Commission (1997) that both an expert's willingness to provide numerical estimates and the quality of their assessments can be improved through training. Such training should explain the meaning of subjective probability, raise awareness of well-known sources of bias and provide *meaningful* exercises on which to practice. To be meaningful to engage the experts, such exercises should align with the problem under investigation.

Bonano et al. (1990) suggest three key tasks be conducted during the training session: First to familiarise the expert with the process and motivate them to provide formal judgements. Second to provide experts with practice at expressing their judgement. Third to educate the experts on potentials for bias.

The quality of subjective probabilities from experts is dependent on both the expert's experience and the method of elicitation. If the expert lacks experience, the prior distributions will be uninformative or misleading, regardless of the elicitation approach employed. Poorly designed elicitation techniques may degrade the quality of information provided by experts. Fischhoff (1989) proposes the following four necessary conditions to support improving judgement skills.

(a) *Abundant practice with a set of reasonably homogeneous tasks*—to assist the expert in developing their judgemental skills on the relevant task.
(b) *Clear-cut criterion events for outcome feedback*—learning requires feedback to the expert, but this can be challenging to evaluate if the judgements are components of complex systems (natural, social or biological).
(c) *Task-specific reinforcement*—performance should be based on the wisdom of their judgement; be aware if there are implicit rewards for the experts. e.g. did they bring good news? Did they disrupt plans?
(d) *Explicit admission of the need for learning*—using titles such as expert can result inhibit learning.

Fischhoff (1989) also points out that often judgements concern events that are not realised for years, which provide little opportunity to learn about the quality of such judgements.

13.3.5 Process Considerations

The elicitation process is more than the means by which the method to obtain the probability assessments is implemented with the selected experts, say by an interview or some other means. We examine issues that are important in creating a coherent process that allows design choices about the probability and people aspects, as discussed in the previous sections, to be meaningfully planned and implemented.

13.3.5.1 Core Activities

The process should account for key activities that add value to the quality of the data collected. Such activities include the recruitment of experts, the framing of questions, the elicitation and aggregation of their judgements, using procedures that have been tested and clearly demonstrated to improve judgements (e.g. Cooke 1991; Mellers et al. 2014).

In particular, the following activities are core to the process.

(a) *Preparation*—this will entail the development of the following: problem statement, project plan, expert panel, reading material, package of data available and elicitation procedures.
(b) *Pilot study/Expert training*—it is essential that all experts share a common understanding of the problem and the specific quantities to be estimated, as well as being trained in using probability. Moreover, the intended use of the outcomes, the elicitation process and the participants' roles need to be explained.
(c) *Expert elicitation*—depending on the approach undertaken, this could be in the form of a group workshop or individual interviews.
(d) *Combine judgements*—depending on the approach undertaken, this could be during the group workshop through interaction or by the analyst following all interviews.
(e) *Feedback*—to all experts.
(f) *Documentation*—participation needs to be appropriately documented, specifically which experts were involved in assessing which quantities as it would be misleading to identify a panel of experts and the resulting assessment only.

13.3.5.2 Tactics for Sound Process Management

Providing guidance on the underpinning reasons for each activity in the process allows the analyst to make better informed design decisions. Hence, explication of the process logic and the role of each activity is important because, if not, then users might approach the process rather superficially through lack of detailed understanding and so inadvertently introduce substantial variations in the elicitation outcomes.

Elicitation processes are lengthy and require the expert to concentrate for a considerable length of time, which can result in compromising the level of accuracy in the elicited probability (Shephard and Kirkwood 1994). The process design should manage experts so that they spend a greater fraction of time on issues of greatest uncertainty. This will avoid a common tendency of spending time on aspects of the problem where data exist and the problem is well understood. Having experts document and bring their written rationales to the elicitation will facilitate the clarification of substantive issues and reduce time (Cooke and Goossens 1999).

13.3.5.3 Checking

The analyst should perform credibility checks to ensure that the probability assessments provided are consistent with an expert's beliefs. This means not only the elicited values, but also the implications of how the analyst is interpreting the judgements by having the expert reflect both the underlying quantity of interest and also the data that will be realised (Keeney and Von Winterfeldt 1991).

When assessing multiple quantities of interest, it is important to check for trends across the values for each to determine if there are any indicators of anchoring and adjustment bias (Siu et al. 2015).

It is also possible to include checks within and beyond the time frame of the elicitation to estimate the predictive accuracy of judgemental probability assessments of uncertainties. For example, "test" quantities of interest for which realisations will be obtained within the time frame of the elicitation provide a means to understand the degree to which an expert is calibrated (Anderson et al. 2015), while having a forward-looking activity to monitor and record any realisations of the quantities of interest enables empirical control, even if only in principle.

13.4 Comparison of Two Elicitation Processes

There are several guidance documents for elicitation processes from a variety of professional or academic sources available in the public domain. We consider the guidance on elicitation processes produced by the European Food Safety Authority (EFSA) in 2014 and the Institute and Faculty of Actuaries (IFoA) in 2015. We select these because they are examples of practice that allow us to illustrate the diversity in application domain as well as the variation in the scope, repetitiveness and level of process prescription. After summarising the salient elements of the guidance for these two processes, to the level expressed by the respective documents, we compare their characteristics in relation to those discussed in Sect. 13.3.

13.4.1 European Food Safety Authority (EFSA) Guidance

The European Food Safety Authority (2014) has developed a detailed process that also includes procedures for expert judgement elicitation within a project, as shown in Fig. 13.3. EFSA is responsible for food safety risk assessment in Europe and operates independently of European legislative, executive institutions and EU Member States. Hence, it is separate from risk management or policymaking. EFSA is a regulator and so deals with expert problems or, occasionally, textbook problems (Hartley and French 2018).

The EFSA process comprises three main phases—initiation, pre-elicitation and elicitation—which are each managed by a different group—working, steering and elicitation group, respectively.

The Working Group defines the problem and justifies the need for an elicitation. This first step requires consideration of all of the relevant model parameters, and to determine which require expert elicitation and which do not. Thus the Working Group prepares a document of the background information.

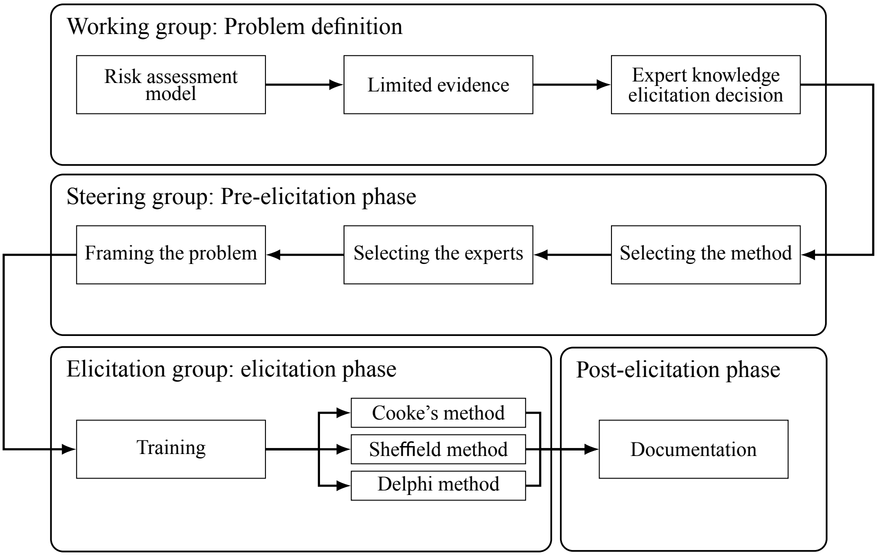


Fig. 13.3 Key phases of the EFSA elicitation process. (Adapted from European Food Safety Authority 2014)

The Steering Group can be a subset of the Working Group and will comprise scientists, experts on elicitation and administrative staff. Their remit is to plan the elicitation process by designing the elicitation protocol. This group specifies the questions suitable for expert elicitation, defines expert profiles and selects the experts and elicitation method as well as the Elicitation Group. Procedures are given for three elicitation methods: the Classical Model, which EFSA calls Cooke's method, Sheffield and Delphi methods which we outline below.

The Elicitation Group typically comprises one or two elicitors with additional administrative support who are familiar and experienced with the selected elicitation protocol. All direct contacts with the experts are made by the Elicitation Group, so members should have a neutral position on the elicitation question. To enhance trust and guarantee confidentiality in ambiguous or conflictive situations, the Elicitation Group should be independent of all parties involved. This group is responsible for executing the elicitation method as well as providing training for the experts.

The evidence dossier is a key part of the guidance to capture the evidence regarding each quantity of interest to be elicited. Expert judgement should not differ because experts have access to different data; difference in opinion should be due to different expertise and interpretation of data. Therefore, data to which the experts have access should be documented and shared. Such documentation should not be too large since it challenges the experts in assimilating all the evidence as well as pointing out the weakness (e.g. small sample sizes), and it can also make the expert anchor on the provided evidence and fail to consider counter facts. The documentation should include any new evidence submitted by experts prior to the elicitation.

Documentation is made public since EFSA upholds the three principles of repeatability, transparency and confidentiality. Three types of report are produced. The result report to summarise the findings; a technical support document to detail a full description of the process and its execution; and expert feedback which is a confidential summarising the input from each expert.

Disclosing personal data that might identify individual experts with their judgements is neither an objective of the process nor necessary to fulfil transparency requirements and may discourage experts from taking part in the process or influence their responses. Participating experts are assured on the confidential treatment of their individual answers, where reports will include who took part, what was said but not who said it.

The Sheffield method is a behavioural aggregation method, where experts participate in a facilitated workshop to create a subjective probability distribution for each quantity of interest. Once the training session has been conducted, the workshops progress through four stages for each quantity of interest. An initial review of evidence is followed by each expert individually assessing their judgement on the quantity. These individual judgements are shared and discussed amongst the group. Aspects of individuals' distributions which are different are discussed within the group and rationales elicited. Then the group judgement is formed as one distribution to represent the view of the rational observer. See Gosling (2018) for a detailed description.

The version of the Delphi method included in the EFSA guidance uses pools of experts but, to minimise adverse group effects, it restricts interpersonal interaction by controlling the flow of information. Experts do not meet, instead they exchange their beliefs and assessments through the facilitator. The facilitator summarises the group's views to the experts and invites each to revise their judgements. See European Food Safety Authority (2014) for a detailed description.

The Classical Model is a performance-based method, where experts work with the analyst independently and without interaction with other experts to assess the uncertainty in the unknown quantity of interest as well as for other variables for which the answer is known to the analyst but not the expert, known as seed questions. Seed questions provide an opportunity of assessing the quality of the responses provided by the experts. See Cooke (1991), Quigley et al. (2018) for more details.

13.4.2 Institute and Faculty of Actuaries (IFoA) Working Paper

The IFoA is the only UK chartered professional body dedicated to educating, developing and regulating actuaries based both in the UK and internationally. Actuaries serve the public interest by conducting analysis where there is uncertainty of future financial outcomes. Solvency II is an EU Directive that came into effect on 1 January 2016 and primarily concerns the risk of insolvency of EU insurance companies. The associated judgement by actuaries in applying the principles of Solvency II prompted a working party for the IFoA to present a paper providing a practical framework regarding expert judgement processes, including their validation, for repeated assessment of risk (Ashcroft et al. 2016). The views expressed in the publication are not necessarily those of the IFoA.

A key motivation for the paper was a lack of transparency on the use of expert judgement within the profession, which is one where judgements have significant impacts on risk assessments and subsequent decisions taken. The authors consider knowledge to be socially constructed so that common judgement can be created through mediation of experiences and ideas. As such, their process is designed to facilitate the pooling of experience and ideas, and not necessarily in consensus.

What we shall label as the IFoA process has five key stages, as shown in Fig. 13.4, and which we discuss below.

First, there is a preliminary assessment to determine whether a formal expert judgement process is relevant. This involves considering whether the nature of the judgement is within the scope of an expert judgement process.

Second, the problem is defined. The problem is articulated and its scope is defined. The current level of understanding of the problem is determined to develop an expert brief. Terminology should be made clear to ensure a consistent interpretation of the problem, which is especially important if using external experts. Potential experts

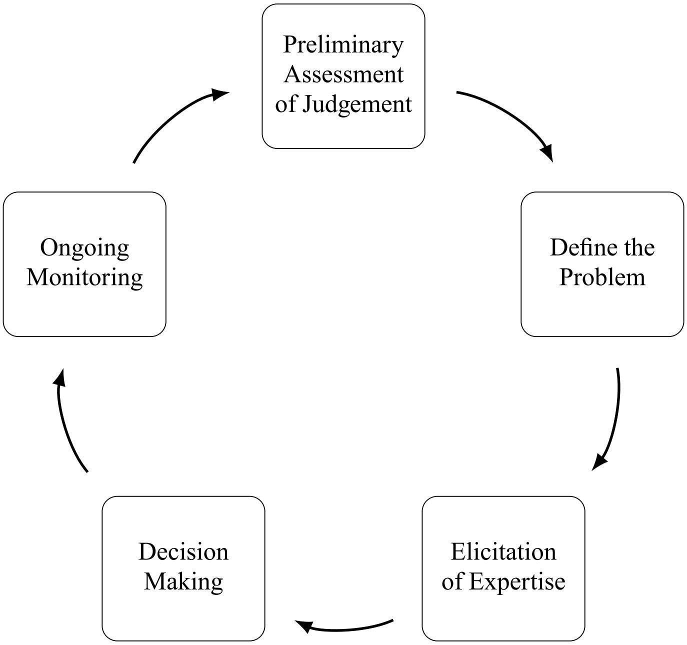


Fig. 13.4 Key stages of IFoA elicitation process. (Adapted from Ashcroft et al. 2016)

are identified, and an initial plausible range of the values of the quantity of interest are assessed.

Third, elicitation of expertise is designed and conducted. The method for elicitation is chosen and will depend upon the nature and importance of the problem; this will include the methods for both encoding the quantities of interest and for combining the views across experts. Documentation is required to describe the available data, assumptions, principles, methodologies and models applied in arriving at a recommendation and on any potential limitations.

Fourth is the decision-making. The decision-makers should review all information (which might include confidential data not known to the experts) and expert judgements to ensure consistency. Decision-makers should set out their thought process clearly, explicating how they are making use of the expert judgement, making clear how they weight the relative importance of information and identify triggers for non-scheduled reviews. This practice is intended to help facilitate a multi-layer governance structure through transparency. A final decision on the judgemental assessment of the quantity of interest is recommended; an overall plausible range, and a summary of the rationale for this, should be communicated back to the experts. This provides an opportunity to flag any serious concerns they may have and which can then be fed back to the decision-makers.

Fifth, there is on-going monitoring. A robust system should be created to monitor the validity of the probability assessment, reflecting on scope of its application, appropriateness of assumptions and triggers for review.

13.4.3 Comparison

The guidance provided by EFSA and the paper from the IFoA naturally differs due to the distinct nature of the problems addressed by the two organisations. EFSA has developed more detailed guidelines aligned with their own organisational need, while IFoA provides higher level guidance to be used by various insurance companies. Conceptually, the processes advocated by both organisations have elements in common, such as problem structuring, an initial evaluation and probability assessment. Since the nature of the problem addressed by the IFoA is on-going, it explicitly continues monitoring after the initial probability assessment, unlike EFSA which assumes a one-off project.

We now compare the two documents in relation to the characteristics of an elicitation process identified in Sect. 13.3.

Principles: Neither process explicitly contradicts any of the principles, but each document supports the principles in varying degrees. Both processes include detailed discussion on documentation and governance. Both allow for processes where expert assessment could be falsifiable with further data; this is either explicitly stated as a goal or implicit in the construction of the elicitation question. The EFSA guidance explicitly states neutrality as a required feature of an elicitation process but how this is ensured is not stated, while the IFoA only mentions the need to manage bias. Fairness is implicit in both processes.

Purpose: Both documents provide a clear statement of purpose. The IFoA has developed guidance for a specific purpose, while EFSA has developed guidance for use in a variety of projects within their remit. EFSA has more clearly identified groups with associated responsibilities of the process. The IFoA acknowledges that multilayer governance structure may vary by institution. Both emphasise the needs for documentation of the elicitation process.

Observable quantity of interest: The EFSA guidance explicitly requires quantities of interest to be observable in principle, whereas the IFoA does not mention this.

Selection of quantity of interest: Both processes advise on the use of data as well as consider an initial assessment of the uncertainty associated with the quantities of interest. These can then be used in a sensitivity analysis prior to conducting the subjective probability assessment by informing prioritisation of the variables to quantify.

Method of encoding: Only the EFSA process explicitly provides guidance on probability distribution fitting with appropriate checks in place and that non-parametric approaches are also available.

Classes of participants: While the EFSA elicitation of the expert's judgement is led by what they term an elicitor, the IFoA considers various formal roles that need to be fulfilled such as the decision-maker, coordinator and validator in addition to the expert. They differentiate between internal and external experts depending upon whether the expert works for the organisation making the risk assessment.

Managed experts: The IFoA process does not provide detailed guidance on the approach to managing groups of experts to the same extent as the EFSA process. The IFoA suggests one may use Delphi or Nominal Group Technique but provides no guidance on the management of interaction. In contrast, the EFSA guidance provides choice depending on whether or not interaction is desired.

Expert training and learning: Both processes acknowledge that some experts will require training in expressing subjective probabilities as well as explicitly requiring that expert feedback is given. However, the nature of the assessments being made implies that meaningful feedback on the predicted assessments is not considered.

Core process activities: Only the EFSA process provides guidance at the level of detail described in Sect. 13.3.5.1.

Tactics for process management: EFSA provides guidance on identifying and managing elicitation fatigue by experts, while the IFoA advises on efficient structuring the elicitation questions to address this issue.

Checking: Both processes provide guidance on feedback to experts as well as validation of their probability assessments. The IFoA process only requires that this be documented but it does not provide guidance on how to validate expert judgement, whereas the EFSA process states that “validation requires eliciting uncertainty on variables whose true values will be known within the time frame of the study” (p. 159 European Food Safety Authority 2014).

13.5 The Value of an Elicitation Process Standard?

Following our comparison of practical guidance, and in light of our abstraction of issues emergent from the literature, we now explore whether it is meaningful to characterise a standard process for elicitation.

Standards represent a voluntary acceptance of the rules. Interestingly, the creation of international standardisation bodies, such as ISO, is grounded in the need to answer the question “what is the best way of doing this?”[1]

The UK national standard body, BSI, identifies three important general drivers for the creation of standards: First, that a standard represents “an agreed way of doing something”. Second, that a standard is “the distilled wisdom of people with expertise in their subject matter and who know the needs of the organisations they represent”. Thirdly, the “point of a standard is to provide a reliable basis for people to share the same expectations about a …” process. See BSI (2018).

We can frame such drivers as the characteristics of a process (or service or product or technology and so on) that has reached sufficient maturity to be standardised. Specifically, we ask whether elicitation processes for the assessment of uncertainty in a quantity of interest have reached such maturity that standardisation would be valuable, and if yes, then how this might be achieved?

We have compiled a set of characteristics of a good elicitation process that are recognised in the literature and correspond to features of an elicitation process that embrace, but also extend beyond, the core scientific principles of Cooke (1991). Further, we have examined the pivotal role of the SRI process in providing a genesis for later, more bespoke elicitation processes. While the latter might emphasise distinct process elements, this might partly be a function of, for example, the distinct purpose of the process in the wider modelling context, the disciplinary bias of the process or method creator and the problem domain in which the process might be applied. By tracing the relationships between features of the leading modern elicitation processes and the SRI in relation to the characteristics of a good elicitation process, we have shown that there is indeed considerable agreement in the way to approach the development of a good elicitation process.

There already exist many guidance documents for elicitation processes and procedures. We have only examined two. Both the EFSA and the IFoA documents are grounded in the wisdom of people with expertise in designing and implementing elicitation exercises as well as experience in understanding the needs of the organisation(s) who will use the elicitation in context. The coverage of such guidance is a function of the scope of the elicitation and the selection of people who have

[1] https://www.iso.org/benefits-of-standards.html [accessed on 20 December 2018].

contributed authorship. The process of creating the guidance documents, of course, will influence the content. Given that there is no guidance yet created by professional standardisation bodies, with all the balances and checks that they deploy in recruiting experts and consensus-forming practices, any elicitation guidance currently in the public domain is subject to the manner in which the commissioning body has procured the guidance, although, having said that, in commissioning guidance there is an implicit intent to provide a reliable basis for people to share the same expectations about the elicitation process.

Following this line of argument, it appears that elicitation processes have reached a state of maturity generally associated with standard creation. Specifically, following Swann and Lambert (2017), we class an elicitation standard as primarily *informative* because they codify process knowledge. This is in contrast to standards that might be classed as primarily *constraining*, such as health and safety. But, even if the intent is to codify and share knowledge to enhance best practice, what are the pros and cons of elicitation process standards? Table 13.1 summarises some key points which we believe are important.

There exist other established standards which achieve the same goals of guiding users in developing, implementing and documenting processes that we might seek to achieve with an elicitation standard. If required by regulation or by contract then such standards can also offer protection to users. Given the recent legal consequences arising in relation to the use of expert assessments of uncertainty as discussed in our Introduction, this aspect might be particularly relevant and novel for elicitation.

There are, of course, mitigations that might help to remove or to reduce the effects of the negative aspects of a standard. For example, much research has been conducted in the relationship between standards and innovation more generally (Blind 2013) with some lessons being applicable for elicitation. Findings show that informative standards can enable, rather than inhibit, innovation within the user organisations given the sharing of codified knowledge. However, the mechanisms for maintaining standards through a formal review and revision process is required to ensure up-to-date guidance. While official standards bodies are empowered to provide such infrastructure, it is not always evident that it occurs within all specific domain bodies. A common concern with standards is that users treat process guidance as a defined procedure rather than think meaningfully about the translation of a process guidance to the specific context. Already elicitation guidance documents have been crafted

Table 13.1 Pros and cons of a standard elicitation process

Pros	Cons
Enhance the craft for those process elements that can be identified a priori	Constrain process innovation to embrace new elicitation knowledge
Increase rigour of process implementation and reduce susceptibility to poor practice	Limit responsibility of the user to think deeply about the specific elicitation
Provide protection to process participants	Make process accessible to poor facilitators

to support better thinking rather than to supplant thinking. However, crafting such guidance is challenging especially if approached at a more general level. As for other process standards, providing guidance on making choices about key activities, such as the selection and definition of the quantities of interest, can be more difficult than giving advice on standard components of documentation, simply because the former is so contingent on the complexity of the modelling problem while the latter is relatively transferrable between applications.

At present, the prevalence of domain elicitation guidance implies that choice of process facilitator is within the control of the commissioning organisations. If there is an elicitation standard, then there may be a growth in the facilitator market meaning less reliance on a smaller pool of knowledgeable facilitators who have earned trust. Creating some form of elicitation facilitator certification might mitigate this risk.

The suggested mitigations tend to rely upon the formalities of a recognised body with responsibility for producing standards as documents established by consensual process. Such bodies already provide standards in other areas of data collection scoped to interface with user needs. There are reduced degrees of standardisation in that recognised bodies also provide technical reports which do allow for sharing of codified knowledge that is informative only. An official standard will contain normative as well as informative text. There is increasing attention to open standards (Maxwell 2006) which are a means to give users permission to use "technology" freely without the involvement of a recognised body in the creation of the standard.

We have established that the practice of elicitation process design and implementation has reached a degree of maturity that allows standard codification of knowledge and we have explored some options regarding creation of a standard for a process for eliciting subjective probability assessments. However, we leave it to the reader to decide whether creating such a standard for an elicitation process would be a valuable endeavour and if, so, in what form.

13.6 Concluding Discussion

We have examined the characteristics of a sound process to elicit judgement to ensure good quality of data to inform decision-making under uncertainty. Even in the contemporary digital world, there is continued need for subjective probability assessments for problems where observed data are non-existent or limited, as well as in situations where observed data are abundant since the relevance of the past to the future need to be assessed with expertise. By exploring the evolution of elicitation processes temporally and across a variety of distinctive problem domains, we have synthesised the characteristics underpinning a good elicitation process—these encompass the probabilistic as well as people aspects of developing a process that aligns with the problem purpose. Such characteristics, and their illustration for elicitation guides produced by two professional organisations, provide a collection of attributes to which a good elicitation should aspire as well as some practical pitfalls to beware. Our goal has been to highlight elicitation process characteristics that are sufficiently general to be widely applicable.

A defensible elicitation process can provide protection to those accountable for the consequences of the determined actions. Appropriate levels of accountability can increase trust in risk information (Frewer et al. 1996). The Cambridge dictionary defines accountability as

> a situation in which someone is responsible for things that happen and can give a satisfactory reason for them

and responsible as

> to have control and authority over something or someone and the duty of taking care of it, him, or her.

As such, a sound elicitation process should produce a satisfactory reason for its results for the person with the duty of care. How satisfactory the reason provided for the subjective probability judgements will depend upon the problem and the associated stakeholders, so guidance will vary in detail across domains. Puig et al. (2018) highlight that a lack of accountability mechanisms are in place to ensure national governments rely on scientifically sound processes for the appropriateness of their forecasting. However, while important, the responsibility does not just rest with the end user. Since some elicitation processes will involve assessing multiple uncertainties by various experts, ensuring each participant is clear about their role in the process should produce accountability; each participant has a responsibility for their contribution. The L'Aquila tragedy, discussed in the Introduction, led to experts being initially held accountable for poor practice and superficial analysis. Of course, a sound process does not guarantee immunity from criticism as other factors will play a role in this social process. For example, Pidgeon (1997) argues that

> despite the inherent complexity and ambiguity of the environments within which large-scale hazards arise and the systemic nature of breakdowns in safety, cultural myths of control over affairs ensure that a culprit must be found after a disaster or crisis has unfolded.

So the responsible should have their reason prepared.

Acknowledgments We would like to thank David Purves and Ian Belton, as well as the editors, for their critical feedback and suggestions.

References

Anderson, G., Walls, L., Revie, M., Fenelon, E., & Storie, C. (2015). Quantifying intra-organisational risks: An analysis of practice-theory tensions in probability elicitation to improve technical risk management in an energy utility. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, *229*(3), 171–180.

Armstrong, J. S., Denniston, W. B, Jr., & Gordon, M. M. (1975). The use of the decomposition principle in making judgments. *Organizational Behavior and Human Performance*, *14*(2), 257–263.

Ashcroft, M., Austin, R., Barnes, K., MacDonald, D., Makin, S., Morgan, S., et al. (2016). Expert judgement. *British Actuarial Journal*, *21*(2), 314–363.

Ashton, A. H., & Ashton, R. H. (1985). Aggregating subjective forecasts: Some empirical results. *Management Science*, *31*(12), 1499–1508.

Aspinall, W., & Cooke, R. (2013). *Quantifying scientific uncertainty from expert judgement elicitation. In: Risk and uncertainty assessment for natural hazards*. Cambridge University Press Cambridge, UK, p 64.

Astfalck, L., Cripps, E., Gosling, J., Hodkiewicz, M., & Milne, I. (2018). Expert elicitation of directional metocean parameters. *Ocean Engineering*, *161*, 268–276.

Ayyub, B. M. (2001). *Elicitation of expert opinions for uncertainty and risks*. CRC Press.

Barons, M. J., Hanea, A. M., Wright, S. K., Baldock, K. C., Wilfert, L., Chandler, D., et al. (2018). Assessment of the response of pollinator abundance to environmental pressures using structured expert elicitation. *Journal of Apicultural Research*, *57*(5), 593–604.

Bedford, T., Quigley, J., & Walls, L. (2006). Expert elicitation for reliable system design. *Statistical Science*, 428–450

Blind, K. (2013). The Impact of Standards on Innovation, NESTA Report. Retrieved 25 Jan, 2019, from http://www.innovation-policy.org.uk/share/14_The%20Impact%20of%20Standardization%20and%20Standards%20on%20Innovation.pdf.

Bolger, F. (2018). *The selection of experts for (probabilistic) expert knowledge elicitation*. In Elicitation, Springer, pp. 393–443.

Bolger, F., & Wright, G. (1992). *Expertise and decision support*. Springer Science & Business Media.

Bonano, E. J., Hora, S., Keeney, R., & Von Winterfeldt, D. (1990). Elicitation and use of expert judgment in performance assessment for high-level radioactive waste repositories. Tech. rep., Nuclear Regulatory Commission, Washington, DC (USA). Div. of High-Level.

Booker, J., & McNamara, L. (2002). Expertise and expert judgment in reliability characterization: A rigorous approach to eliciting, documenting and analyzing expert knowledge.

Brockhoff, K. (2002). The performance of forecasting groups in computer dialogue and face-to-face discussion. In: The Delphi method: Techniques and applications, http://www.is.njit.edu/pubs/delphibook/delphibook.pdf, pp 285–311.

Broomell, S., & Budescu, D. (2009). Why are experts correlated? decomposing correlations between judges. *Psychometrika*, *74*(3), 531–553.

BSI. (2018). Information about standards. Retrieved 20, Dec 2018, from https://www.bsigroup.com/en-GB/standards/Information-about-standards.

Budnitz, R. J., Apostolakis, G., Boore, D. M., Cluff, L. S., Coppersmith, K. J., Cornell, C. A., et al. (1998). Use of technical expert panels: applications to probabilistic seismic hazard analysis. *Risk Analysis*, *18*(4), 463–469.

Burgman, M. (2004). Expert frailties in conservation risk assessment and listing decisions. In: P. A. Hutchings, D. Lunney, C. R. Dickman (Eds.) Threatened Species Legislation: is it just an act?, pp. 20–29.

Chaloner, K., Church, T., Louis, T. A., & Matts, J. P. (1993) Graphical elicitation of a prior distribution for a clinical trial. *The Statistician*. pp. 341–353.

Choy, S. L., O'Leary, R., & Mengersen, K. (2009). Elicitation by design in ecology: using expert opinion to inform priors for Bayesian statistical models. *Ecology*, *90*(1), 265–277.

Christophersen, A., Deligne, N. I., Hanea, A. M., Chardot, L., Fournier, N., & Aspinall, W. P. (2018). Bayesian Network modeling and expert elicitation for probabilistic eruption forecasting: Pilot study for Whakaari/White Island New Zealand. *Frontiers in Earth Science*, *6*, 211.

Clemen, R., & Winkler, R. (1986). Combining economic forecasts. *Journal of Business and Economic Statistics*, *41*(1), 39–46.

Cooke, R. (1991). *Experts in uncertainty: Opinion and subjective probability in science*. Oxford University Press on Demand.

Cooke, R., & Jager, E. (1998). Failure frequency of underground gas pipelines: methods for assessment with structured expert judgment. *Risk Analysis*, *18*(4), 511–527.

Cooke, R. M., & Goossens, L. (1999). *Procedures guide for structured expert judgment* (p. 18820). EUR: Project Report to the European Commission.

Cooke, R. M., & Goossens, L. L. (2008). TU delft expert judgment data base. *Reliability Engineering & System Safety*, *93*(5), 657–674.

Cox, L. A, Jr. (2012). Confronting deep uncertainties in risk analysis. *Risk Analysis: An International Journal*, *32*(10), 1607–1629.

Dalkey, N. (1967). *Delphi (Report P-3704)*. Santa Monica, CA: Rand Corporation.

Dalkey, N., & Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. *Management Science*, *9*(3), 458–467.

Dalkey, N. C. (1969). The Delphi method: An experimental study of group opinion. Tech. rep., The RAND Corporation, Santa Monica, CA (No. RM-5888-PR).

Dias, L. C., Morton, A., & Quigley, J. (2018). Elicitation: State of the Art and Science. In A. Morton & J. Quigley (Eds.), *Dias LC* (pp. 1–14). Springer: Elicitation.

European Food Safety Authority. (2014). Guidance on expert knowledge elicitation in food and feed safety risk assessment. *EFSA Journal*, *12*(6), 3734.

Ferrell, W. R. (1985). *Combining individual judgments*. In: Behavioral decision making, Springer, pp. 111–145.

Ferrell, W. R. (1994). *Discrete subjective probabilities and decision analysis: Elicitation*. Calibration and Combination: Wiley.

Fischhoff, B. (1989). Eliciting knowledge for analytical representation. *IEEE Transactions on Systems, Man, and Cybernetics*, *19*(3), 448–461.

French, S. (2012). Expert judgment, meta-analysis, and participatory risk analysis. *Decision Analysis*, *9*(2), 119–127.

Frewer, L. J., Howard, C., Hedderley, D., & Shepherd, R. (1996). What determines trust in information about food-related risks? Underlying psychological constructs. *Risk Analysis*, *16*(4), 473–486.

Frijters, M., Cooke, R., Slijkuis, K., & van Noortwijk, J. (1999). *Expert judgment uncertainty analysis for inundation probability*. Utrecht: Ministry of Water Management, Bouwdienst, Rijkswaterstaat.

Gigerenzer, G., & Edwards, A. (2003). Simple tools for understanding risks: From innumeracy to insight. *BMJ*, *327*(7417), 741–744.

Gosling, J. P. (2018). SHELF: the Sheffield elicitation framework. In A. Morton & J. Quigley (Eds.), *Dias LC* (pp. 61–93). Springer: Elicitation.

Hampton, J., Moore, P., & Thomas, H. (1973). Subjective probability and its measurement. *Journal of the Royal Statistical Society Series A (General)*, 21–42

Hanea, A. M., Burgman, M., & Hemming, V. (2018). IDEA for uncertainty quantification. In A. Morton & J. Quigley (Eds.), *Dias LC* (pp. 95–117). Springer: Elicitation.

Hartley, D., & French, S. (2018). Elicitation and calibration: A Bayesian perspective. In A. Morton & J. Quigley (Eds.), *Dias LC* (pp. 119–140). Springer: Elicitation.

Hemming, V., Burgman, M. A., Hanea, A. M., McBride, M. F., & Wintle, B. C. (2018). A practical guide to structured expert elicitation using the IDEA protocol. *Methods in Ecology and Evolution*, *9*(1), 169–180.

Hodge, R., Evans, M., Marshall, J., Quigley, J., & Walls, L. (2001). Eliciting engineering knowledge about reliability during design-lessons learnt from implementation. *Quality and Reliability Engineering International*, *17*(3), 169–179.

Hogarth, R. (1978). A note on aggregating opinions. *Organizational Behavior and Human Performance*, *21*, 40–46.

Hora, S. C. (2007). Eliciting probabilities from experts. Advances in decision analysis: From foundations to applications 129.

Hora, S. C., & Von Winterfeldt, D. (1997). Nuclear waste and future societies: A look into the deep future. *Technological Forecasting and Social Change*, *56*(2), 155–170.

Janis, I. L. (1971). Groupthink. *Psychology Today*, *5*(6), 43–46.

Junger, S. (1997). *The perfect storm: A true story of men against the sea*. WW Norton & Company.

Kadane, J., & Wolfson, L. J. (1998). Experiences in elicitation. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *47*(1), 3–19.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases* (pp. 3–20). Judgement under uncertainty: Heuristics and biases.

Kammerer, A. M., & Ake, J. P. (2012). *Practical implementation guidelines for SSHAC Level 3 and 4 hazard studies*. Office of Nuclear Regulatory: United States Nuclear Regulatory Commission.

Keeney, R. L., & Von Winterfeldt, D. (1991). Eliciting probabilities from experts in complex technical problems. *IEEE Transactions on Engineering Management*, *38*(3), 191–201.

Krueger, T., Page, T., Hubacek, K., Smith, L., & Hiscock, K. (2012). The role of expert opinion in environmental modelling. *Environmental Modelling & Software*, *36*, 4–18.

Kuhnert, P. M., Martin, T. G., & Griffiths, S. P. (2010). A guide to eliciting and using expert knowledge in Bayesian ecological models. *Ecology Letters*, *13*(7), 900–914.

Larrick, R. A. E. M., B. S. J. (2011). The social psychology of the wisdom of crowds. In: Ji, K. (Ed.) Frontiers of social psychology: Social psychology and decision making, New York: Psychology Press.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, *108*(22), 9020–9025.

Martelli, A., & Mualchin, L. (2012). Indictment and conviction of members of the Italian "Commissione Grandi Rischi" (open letter to the President of Italy). Retrieved 8, June 2017, from www.cngeologi.it/wp-content/uploads/2012/10/CoverletterandStatementISSO1.pdf.

Mastrandrea, M. D., Field, C. B., Stocker, T. F, Edenhofer, O., Ebi, K. L., Frame, D. J., Held, H., Kriegler, E., Mach, K. J., & Matschoss, P. R. et al. (2010). Guidance note for lead authors of the IPCC fifth assessment report on consistent treatment of uncertainties. Intergovernmental Panel on Climate Change.

Maxwell, E. (2006). Open standards, open source and open innovation; harnessing the benefits of openness. *Innovations: Technology, Governance, Globalization, 1*(3), 119–176.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., et al. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, *25*(5), 1106–1115.

Merkhofer, M. W. (1987). Quantifying judgmental uncertainty: Methodology, experiences, and insights. *IEEE Transactions on Systems, Man, and Cybernetics*, *17*(5), 741–752.

Meyer, M., & Booker, J. (2001). Eliciting and analyzing expert judgment: A practical guide. *American Statistical Association and Society for Industrial and Applied Mathematics*. Philadelphia.

Miller, A. C, I. I. I., & Rice, T. R. (1983). Discrete approximations of probability distributions. *Management Science*, *29*(3), 352–362.

Moore, C. M. (1987). *Group techniques for idea building*. Sage Publications, Inc.

Myers, D. G., & Lamm, H. (1976). The group polarization phenomenon. *Psychological Bulletin*, *83*, 602–627.

Nature. (2011). Scientists on trial: at fault? Retrieved 8 June 2017, from www.nature.com/news/2011/110914/full/477264a.html.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J. et al. (2006). *Uncertain judgements: eliciting experts' probabilities*. Wiley.

Page, S. E. (2007). Making the difference: Applying a logic of diversity. *Academy of Management Perspectives*, *21*(4), 6–20.

Pidgeon, N. (1997). The limits to safety? Culture, politics, learning and man-made disasters. *Journal of Contingencies and Crisis Management*, *5*(1), 1–14.

Puig, D., Morales-Nápoles, O., Bakhtiari, F., & Landa, G. (2018). The accountability imperative for quantifying the uncertainty of emission forecasts: Evidence from Mexico. *Climate Policy*, *18*(6), 742–751.

Quigley, J., & Walls, L. (2010). Reconciling experts opinion concerning the value of testing using Bayesian networks: A bridge too far? In *5th International ASRANet Conference*.

Quigley, J., & Walls, L. (2018). A Methodology for Constructing Subjective Probability Distributions with Data. In A. Morton & J. Quigley (Eds.), *Dias LC* (pp. 141–170). Springer: Elicitation.

Quigley, J., Colson, A., Aspinall, W., & Cooke, R. M. (2018). Elicitation in the classical model. In A. Morton & J. Quigley (Eds.), *Dias LC* (pp. 15–36). Springer: Elicitation.

Regan, T. J., Burgman, M. A., McCarthy, M. A., Master, L. L., Keith, D. A., Mace, G. M., et al. (2005). The consistency of extinction risk classification protocols. *Conservation Biology*, *19*(6), 1969–1977.

Sackman, H. (1975). Delphi critique. Expert Opinion, Forecasting. Group Process NY: Lexington Books, pp. 30–50.

Science. (2012). Earthquake experts convicted of manslaughter. Retrieved 8 June 2017, from www.sciencemag.org/news/2012/10/earthquake-experts-convicted-manslaughter.

Science. (2014). Updated: Appeals court overturns manslaughter convictions of six earthquake scientists. Retrieved 8 June 2017, from www.sciencemag.org/news/2014/11/updated-appeals-court-overturns-manslaughterconvictions-six-earthquake-scientists.

Seaver, D. A., Von Winterfeldt, D., & Edwards, W. (1978). Eliciting subjective probability distributions on continuous variables. *Organizational Behavior and Human Performance*, *21*(3), 379–391.

Shanteau, J., Weiss, D. J., Thomas, R. P., & Pounds, J. C. (2002). Performance-based assessment of expertise: How to decide if someone is an expert or not. *European Journal of Operational Research*, *136*(2), 253–263.

Shephard, G. G., & Kirkwood, C. W. (1994). Managing the judgmental probability elicitation process: a case study of analyst/manager interaction. *IEEE Transactions on Engineering Management*, *41*(4), 414–425.

Siu, N., Xing, J., & Taylor, G. (2015). Eliciting expert judgment—peer review observations from a recent exercise and future plans. www.nrc.gov/docs/ML1502/ML15028A183.pdf.

Soll, J. B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 299.

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 780.

Spetzler, C. S., & Stael von Holstein, C. A. S. (1975). Exceptional paper-Probability encoding in decision analysis. *Management Science*, *22*(3), 340–358.

Swann, G., & Lambert, R. (2017). Standards and innovation: A brief survey of empirical evidence and transmission mechanisms. In R. Hawkins, K. Blind, R. P. (Eds.) Handbook of innovation and standards, Edward Elgar Publishing.

Taleb, N. N. (2007). *The black swan: The impact of the highly improbable*, (Vol. 2). Random House.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgement. *Psychological Review*, *90*(4), 293–315.

Tversky, A., & Koehler, D. (1994). Support theory: A non-extensional representation of subjective probability. *Psychological Review*, *101*(4), 547–567.

United States Nuclear Regulatory Commission (1975) Reactor Safety Study, An Assessment of Accident Risks in U.S. Commercial Nuclear Power Plants, WASH-1400. NUREG-75014 [commonly referred to as the Rasmussen Report].

United States Nuclear Regulatory Commission and others. (1990). Severe accident risks: An assessment for five US nuclear power plants. NUREG-1150.

US Nuclear Regulatory Commission. (1997). (NUREG/CR-6372) US Nuclear Regulatory Commission, Guidance on Uncertainty and Use of Experts. www.nrc.gov/reading-rm/doc-collections/nuregs/contract/cr6372/.

Walls, L., & Quigley, J. (2001). Building prior distributions to support Bayesian reliability growth modelling using expert judgement. *Reliability Engineering & System Safety*, *74*(2), 117–128.

Wallsten, T. S., Budescu, D. V., Rapoport, A., Zwick, R., & Forsyth, B. (1986). Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General*, *115*(4), 348.

Wilson, K. J. (2017). An investigation of dependence in expert judgement studies with multiple experts. *International Journal of Forecasting*, *33*(1), 325–336.

Winkler, R. L. (1967). The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical Association*, *62*(319), 776–800.

Wisse, B., Bedford, T., & Quigley, J. (2008). Expert judgement combination using moment methods. *Reliability Engineering & System Safety*, *93*(5), 675–686.

Chapter 14
Developing a Training Course in Structured Expert Judgement

Philip Bonanno, Abigail Colson, and Simon French

Abstract The chapter discusses the design and development of a training course in structured expert judgement (SEJ). We begin by setting the course in the context of previous experiences in training SEJ to postgraduates, early career researchers and consultants. We motivate our content, discussing the theoretical framework that guides the design of such a course. We describe our experiences in presenting the course on two occasions. Detailed analysis of the different course components—the learners/participants, the content, the context and the method, was carried out through surveys given to participants. This helped identify the successful course characteristics, which were then summarised in a customised design template that can be used to guide its conceptual, structural and navigation design.

14.1 Introduction

Structured expert judgement (SEJ) is not an easy topic to teach in a formal course. Many of the processes and techniques used in interacting with experts require tacit skills that are not easily conveyed in a 'chalk-and-talk' environment. So, the objective of the COST Structured Expert Judgement Network,[1] to help creating a new generation of scientists who are confident and able to bridge the gaps between science and policy through the use of SEJ, was a challenging one.

[1] COST Action IS1304: Expert Judgement Network—Bridging the Gap Between Scientific Uncertainty and Evidence-Based Decision Making: see https://expertsinuncertainty.net/ and http://www.cost.eu/COST_Actions/isch/IS1304. The Network ran from 2013 to 2017.

P. Bonanno (✉)
University of Malta, Msida, Malta
e-mail: philip.bonanno@um.edu.mt

A. Colson
University of Strathclyde, Glasgow, Scotland, UK

S. French
University of Warwick, Coventry, UK
e-mail: simon.french.50@gmail.com

A. M. Hanea et al. (eds.), *Expert Judgement in Risk and Decision Analysis*,
International Series in Operations Research & Management Science 293,
https://doi.org/10.1007/978-3-030-46474-5_14

Over the years, several members of the network had accumulated extensive experience in mentoring and training postgraduate students and early career researchers in SEJ methodologies. Much of this had been through one-to-one mentoring, e.g. in supervising and mentoring research students. Some, however, had an experience of training SEJ in various courses. In the mid-1990s, Tim Bedford, Roger Cooke, Simon French and Jim Smith had run a number of courses in Cambridge on *Dependence Modelling and Risk Management* (see Chap. 17 for some reminiscences from an alumni). These covered many topics in addition to SEJ and the Classical Model, e.g. belief nets and decision analysis. There were some practicals but these involved training in the Excalibur software and understanding the analysis rather than the process of designing an SEJ study, interacting with experts, and eliciting judgements. More recently, many of the network members have been involved in designing and giving several different courses on SEJ at the European Food Safety Authority (EFSA) to help instil SEJ into their toolkits (see EFSA (2014) for a description of their processes). Again, however, the EFSA courses do not really seek to develop the tacit skills used in interacting with experts and eliciting their judgements. Their aim is to provide EFSA staff with an understanding of the entire SEJ process, and how it fits into a wider risk assessment. It trains them to commission and manage EFSA scientific SEJ studies rather than conduct them.

Away from the field of SEJ, some of the network members had run a variety of action-learning programmes to develop tacit skills. In particular, one of us (SF) had been involved in developing training programmes in risk communication for the UK Department of Health (French and Maule 1999, 2010). Public risk communication is again a tacit skill. These courses had used hypothetical but realistic scenarios to focus discussion and provide a context in which the participants could develop a risk communication strategy.

A key challenge would be the length of the course. COST funding limited us to about two and a half days, but the volume and range of material that we wished to convey would require far more than that. We realised that we would need to draw on e-learning tools within a modern web-based learning environment. Many within the network were experienced in teaching in such environments; and we were fortunate in that one of us (PB) had specific expertise in designing blended learning courses (Bonanno 2011).

Together these factors led us to adopt a mentored action-learning approach, based around an extended hypothetical scenario, and supported by e-learning tools.

14.2 Some Theoretical Underpinnings

Both the structure and methodology of this blended learning (face-to-face and online) course were inspired by key epistemologies, pedagogical models and learning theories. In today's technology-infused society, expert knowledge does not reside only in the mind of experts but distributed between experts and digital systems (Siemens 2005, 2006; Downes 2012). Expertise is increasingly becoming hybridised comprising human and digital components. The traditional knowledge forming part of the mind of the expert, which is communicated and shared through different forms of interpersonal communication, merges with the network of knowledge residing in digital systems; in this case, the Moodle course management system is hosted by the University of Strathclyde, Glasgow, UK. This digital system comprises crystalised forms of expert knowledge in the form of digital resources and data, together with all the interactions between experts and novice scientists or interactions between course participants in the form of written reflections, comments and enriching contributions involving the addition of other resources or examples.

The integration of technologies in formal and non-formal training contexts creates the need for a different frame-of-mind to conceptualise the learning experience. It demands objective analysis of models that emphasise learning as a process of content transmission and move on to a process-oriented methodology that considers learning and knowledge building in terms of dimensions and levels of interaction with the physical world, with conceptual artefacts (Bereiter 2002) and in between members of a learning group.

Developments in various fields of research point to the importance of adopting process-oriented approaches in analysing such contexts. Cognitive neuroscience (Eg. Frith 2007; Frith and Frith 2003; Frith and Wolpert 2003) emphasise the importance of interpreting human behaviour in whatever context from a social perspective, focusing on the dynamics of interactions that each member of a group triggers or reacts to. Connectivist (Siemens 2004) and Constructionist (Kafai and Resnik 1996; Sabelli 2008) epistemologies advocate a process-oriented methodology that considers learning and knowledge building as a process of interactions with the external world, with the intra-individual reality, with conceptual artefacts and with communities of learning in different domains of expertise. Gaining competence and expertise in any field imply a continual process of establishing and elaborating interactions with the physical and conceptual artefacts of that field and with knowledgeable persons in that area.

Learning in blended methodologies is driven by key intrapersonal and psychosocial processes, which give rise to various dimensions of interactions. The underlying processes of skill imitation (Frith and Wolpert 2003), negotiation and argumentation (Dillenbourg et al. 1996) generate task-oriented interactions related to competence development along the domain and technology dimensions. On the other hand, the psychosocial processes of impression formation (Kreijns et al. 2003), mentalising (Frith and Frith 2003), social[2] monitoring (Jost et al. 2002) and interpersonal communication generate categories of person-oriented interactions that characterise technology-mediated group dynamics. In this way, the online learning community is capable of promoting reflection about the content of the field of expertise (SEJ), the individual developmental journey of expertise acquisition of each member of the learning group and the distributed knowledge, experience and expertise characterising the learning group.

[2] https://lighttwist-software.com/excalibur/.

Using this theoretical framework, Bonanno (2011) developed a model to design blended and ubiquitous learning considering different dimensions and levels of interactions. A technology-enhanced course, such as the one developed for the training school in SEJ, comprises interactions along with the domain (SEJ), technology (Moodle, EXCALIBUR software) and the community (experts, tutors, participants). It also organises interactions along with three increasingly engaging levels of learning—Acquisition (understanding new ideas about expert judgement), Participation (in thematic or case-based discussions) and Contributions (generating and sharing of ideas). All the stages of course development, including front end analysis, conceptual design, structural design, navigational design and evaluation were guided by this process-oriented interaction-based model. The following section describes in detail the key stages in the design process.

14.3 Design of the Course Content

Although we were committed to instilling many tacit skills into the participants, we were also aware that there was much technical background on SEJ, its techniques and processes that would be needed. So, in planning the course, we split the time roughly 50:50 between

- lectures, case studies, research seminars and discussions;
- a group exercise based around a hypothetical scenario.

We also decided that many of the lectures would contain short experiential exercises so that the students were aware of the cognitive issues that experts face in making probability judgements.

The first course was held in Madrid in April 2015, immediately before a COST workshop. The tutors were David Rios Insua (local organiser), Eva Chen, Jesus Rios Alaga, Oswaldo Morales, Philip Bonanno, Eva Chen, Roger Cooke and Simon French. The participants gathered on a Sunday evening and were together until Wednesday lunchtime. In order to sensitise the participants to issues of calibration, heuristics and biases in judging uncertainties (see Chap. 1), we used the Sunday evening to run a short elicitation exercise which could be explored during the opening session on Monday.

The hypothetical example was built around what was then a current possibility, namely that the University of Warwick was planning to open a campus in California. The participants were asked to forecast the number of students recruited onto a Masters course in Expert Judgement in Risk Analysis during the 3 years after the campus' opening. Details of the task given to the students are given in Box 14.1. We chose this example because all of the course tutors had an experience of running master programmes generally and so could act as experts in the exercises.

Box 14.1: The group exercise for the Madrid course. For explanations of terms such as seed variable see Chaps. 1 and 10.

Group Expert Judgement Exercise

The University of Warwick is establishing a new Californian University near Sacramento. It will concentrate on graduate programmes. One option that it is exploring is that of a new Masters programme on *Expert Judgement in Risk Analysis*. The course team preparing the programme are none other than David, Eva, Jesus, Oswaldo, Philip, Roger, and Simon. The programme will spend about 60% of its time on the principles and practice of risk analysis and 40% on those of Expert Judgement. It would be a year-long programme with heavy practical emphasis, including a 3-month project usually involving secondment to a government agency or utility to work on a major risk analysis.

To make a business case for such a Masters programme, there is a need to forecast the number of students to be recruited in each of the first three academic years, 2020, 2021 and 2022. Moreover, the numbers should be broken into students from the US, Europe, Asia and the rest of the World. Your task is to elicit distributions for these numbers from as many of us you can interview on Tuesday afternoon.

This afternoon your task is to develop a set of appropriate seed variables for this task. You will also need to develop elicitation sheets, a protocol for interviewing us individually. You need to set up all the tools for the study this afternoon. Your mentor is there to facilitate your discussion and offer methodological advice, *but not to suggest seed variables.* Those are for you to develop using the web or whatever. Later this afternoon, you may if you wish find one of David, Eva, Jesus, Oswaldo, Philip, Roger and Simon (not your mentor) to test your elicitation on. You should also spend some of the time this afternoon ensuring that at least one of your expert groups can use Excalibur fully.

Tomorrow between 3.00 and 7.00 you will need to elicit judgements from as many of the remaining five of us (not your mentor, nor the person you trialled your elicitation on). You need to feed your results into Excalibur and analyse the results.

On Wednesday morning you should present your results describing how you developed your seed variables, etc., as well as giving your distributions. You will have about 10 min for the presentation (7 min talking and 3 min for questions and comments).

The course was supported by a Moodle learning environment with introductory notes, lecture slides, other course materials and discussion areas, and the students would be asked to use this in the weeks immediately before the course to gain some basic appreciation of the issues and topics that we would discuss.

The outline and structure of the course being settled there was the question of what topics, theories and applications to cover in the lectures, seminars and discussion sessions. Some were obvious. The classical method (see Chap. 10) has a long history of application and would be well fitted to the needs of the group exercise. So, we would include lectures and short experiential exercises to train the participants in that. But what else? Something on behavioural issues, heuristics and biases surely: understanding those were essential to understanding the processes of elicitation. For the rest of the content, we asked the participants. In the weeks before the course, we sent the participants a short survey with ten general, mainly open-ended, questions to gather their expectations. Their thoughtful responses did much to shape the final details of the course. A detailed analysis of the survey is given in Appendix 14.7. In the end, we included a lecture on behavioural issues and training in the implication of these for how we should elicit probabilities. Discussion of the Good Judgement Project and the Wisdom of Crowds added further material (Tetlock and Gardiner 2016). Two case studies both explained how the tools could be used and discussed the particular tools that had been used on those. A general lecture gave an overview of topics such as aggregation of several experts' judgements, Bayesian approaches, group decision-making and meta-analysis.

The outline of the course as it was delivered in Madrid is given in Box 14.2.

The course was supported through a Moodle managed by Strathclyde University. The major sections in this dedicated Moodle instance include:

- An introductory section comprising a welcome note and a list of the learning outcomes for the course.
- A Moodle Book including:
  - Expert Judgement, Risk and Decision-Making
  - The 'Three' Contexts of Expert Judgement.
- A resource section including all readings, presentations and documents used during the course.
- A forum to host discussions during and after the event to continue sharing experiences in applying the ideas after returning to their research institutes and universities.
- A forum/discussion space for practical exercises including:
  - Individual judgements of the uncertainty of a set of calibration variables
  - Identify the training that they would need to provide as analysts to their experts
  - Structured elicitation of their individual judgements on further seed variables and quantities of interest.
- A forum to record and pool group discussion with mentors including:
  - Identified key uncertainties needing expert judgement and potential seed variables for these.
  - Description of their final expert judgement.
  - Personal reflections summarising their feelings and learning during mentoring.
  - Feedback and further personal reflections on the programme activities and results of the previous day.

– Analysis, reports and presentation on the results of the group exercise about the elicitation of judgements from ‘experts’.
– Personal reflection on how participants will continue to promote the acquired skills in their research and professional activity.

The following year a second course was run during March in Warsaw. Course tutors were Philip Bonanno, Tim Bedford, Abigail Colson, Rene von Dorp, Tina Nane and Michał Zdziarski (local organiser). Participants gathered on Sunday evening and were together until 3:00 p.m. Wednesday. Based on the experience of the first training course, we extended the final day a couple of additional hours to allow further reflection on the group experiential exercise.

Box 14.2: Outline of Madrid Course given in April 2015

Evening of Arrival Day

Welcome Reception, Introduction, a brief overview of the course, a short elicitation exercise and Dinner.

First Day of the Course

Morning: Welcome, Introduction, management issues, familiarisation with available facilities, followed by lectures and discussions on:

- the Classical Method,
- behavioural issues including heuristics and biases,
- training in making probability judgements and how to train others in the same
- a case study on financial forecasting.
- the Classical Method,
- behavioural issues including heuristics and biases,
- training in making probability judgements and how to train others in the same
- a case study on financial forecasting.

Evening: Reception and a lecture on various experiences in use of SEJ over several case studies.

Second Day of the Course

Morning: Lectures and discussions on:

- expert judgement theory, Bayesian approaches, group decision making, meta-analysis and others.
- case study on aviation safety
- design, reporting and peer review of expert judgement studies, including an exercise critiquing a report of an expert judgement study.
- seminar on the good judgement project and the wisdom of crowds

Afternoon: Second session of group work on the hypothetical example

Third Day of the Course

Morning:

- Third session of group work on the hypothetical
- Group presentations of the results of the elicitation exercise and discussion.
- General discussion on how participants will take their skills forward in their research and professional activity.

Close of course.

The Moodle learning environment again supported the course with the introductory material mostly remaining unchanged from the original Madrid course. An additional document demonstrating the use of Excalibur and a sample Excalibur data file was added, providing the participants with more opportunities to learn how to use the program. As the feedback from the first course was positive, the general scope of the topics covered in the second course remained the same. As new tutors were involved, however, case studies and other extensions of the core material were changed to reflect their interests and experience. An outline of the course is given in Box 14.3.

For the Warsaw course, the experiential example was built around the impact of a hypothetical sugar-sweetened beverage tax in the UK (See Box 14.4). Participants were asked to forecast the impact of the tax on demand for sugar-sweetened beverages across different income groups. Incidentally, the UK's current sugar-sweetened beverage tax was announced on the first day of the training course, so it was a very timely example. Participants found data for seed questions from historical data on consumer behaviour in the UK and the experience of Mexico, California and other areas that previously implemented similar taxes.

Box 14.3: Outline of Warsaw course given in March 2016

Evening of Arrival Day

Welcome reception at the Invisible Exhibition Warsaw, Introduction, Dinner.

First Day of the Course

Welcome, Introduction, and lectures and discussions on:
Uncertainty, probability, and decision making
Theory and application of the Classical Model
Introduction to paired comparison methods
Overview of the group exercise

Participants had several hours in the afternoon to work on the group exercise, including a group discussion to see what, if any, questions and challenges were emerging before breaking for the day.

Second Day of the Course

An optional help session on Excalibur was followed by feedback on an elicitation exercise the participants did during Day 1. Several case studies were presented, and the participants had all afternoon to work on the group exercises, including eliciting assessments from the "experts".

The evening included a visit to the Warsaw Uprising Museum and a discussion of the decision-making under uncertainty that led to that event. This was followed by a group dinner.

Third Day of the Course

Lectures included problem structuring with expert judgement, additional case studies and an introduction to the Classical Model database, which was used to answer many of the questions the participants raised over Day 1 and Day 2. Participants had a bit of time for final group work before presenting their work.

After lunch, Prof Tim Bedford gave a keynote lecture open to the university on validating expert judgements, which was followed by a final Q&A and reflection session.

Box 14.4: The group exercise used in the Warsaw Course

Group Expert Judgement Exercise

Hypothetical Scenario:

The United Kingdom is considering introducing a new "sugar tax" on soda drinks. The primary purpose of the tax is to improve public health by reducing sugar consumption, a large proportion of which happens through soda drink consumption. The revenue raised by the tax is a secondary issue for the government. The envisaged tax would be a 20% tax on sugar-sweetened beverages, as recommended by the British Medical Association. This would cover sugary sodas as well as sugar-sweetened juices, sports drinks, and other non-alcoholic drinks. It would not cover drinks with artificial sweeteners (i.e. diet sodas). If approved, the tax would go into effect on January 1, 2017.

As a first step toward understanding the public health impact of such a tax, you have been asked to forecast the change in consumption of sugar-sweetened beverages over each of its first three years (2017, 2018 and 2019), relative to the baseline consumption of 2015. As the Department of Health is particularly interested in the impact of the tax in different socioeconomic groups, the forecasts should be broken down by household income quintiles. The department is interested not just in the point estimates of impact, but also in the uncertainty distribution surrounding those estimates.

Your Challenge:

With your group, you need develop appropriate seed questions and an elicitation protocol that includes the variables of interest. Each group will have a mentor to facilitate discussion and offer methodological advice, *but not to suggest seed variables*. You need to develop those using the information on the web or whatever else you have at your disposal.

Over the next few days, time will be set aside for this group work. A recommended plan of attack would be:

- Wednesday afternoon: Identify seed questions and write elicitation protocol.
- Thursday morning: Finalise the elicitation protocol and test it on someone other than your group's mentor.
- Thursday afternoon: Elicit judgements from as many of Abby, Michał, Philip, Rene, Tim and Tina as possible (aside from your mentor and whomever you tested the protocol on).
- Friday morning: Analyse results in Excalibur and prepare a presentation.

On Friday, each group will have 15 min to present their seed questions, protocol and results.

References

British Soft Drinks Association. 2015. "Changing tastes: The UK soft drinks annual report 2015." http://www.bma.org.uk/foodforthought.

British Medical Association. 2015. "Food for thought: Promoting healthy diets among children and young people." http://www.bma.org.uk/foodforthought.

Public Health England. 2015. "Sugar reduction: From evidence to action." https://www.gov.uk/government/publications/sugar-reduction-from-evidence-into-action.

14.4 Evaluation of the Courses

Both courses were evaluated through a post-course survey. Participants were asked to score the different components of the training course using a 5-point Likert scale. Responses were very positive with everyone rating all components as good or excellent. The results are summarised in Fig. 1. More details on the responses are provided in Appendix 14.8.

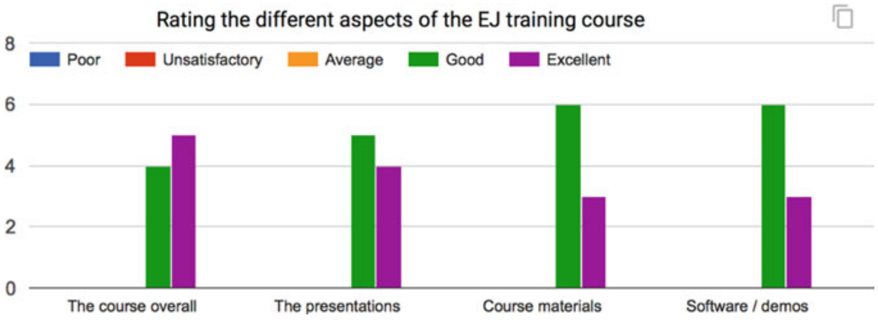


Fig. 1 Evaluation of course components

14.5 A Course Design Template

Drawing on our experiences of running these courses, we have developed a checklist for planning future courses, given below. It identifies the Role of Course Administrators and describes the Learning Outcomes, Course Content, Course Structure, Course Delivery, Assessment of Course Content, Course Evaluations and Resources (for knowledge communication, collaborators, reflection):

Roles of Course Administrators	✓
Processes course applications	
Sets up competence-related working groups	
Organises online course (in Moodle)	
Organises face-to-face events	
Coordinates the design and development of course units (possible by providing a unit design template)	
Coordinates the customisation of course schedules	
Provides short introductory unit about online learning	
Provides identified methodological tools	
Coordinates course evaluation	
Provides mentors for participants	
Provides a certificate of attendance for course-related events	
Provides an ‘End of course certification’	
Learning Outcomes	
Identify different schools of thought/areas of expertise in SEJ	
Understand the process for SEJ	
Can develop techniques for eliciting information from experts	
Define proper seed questions for experts	
Process experts’ answers and translate them into numerical outcomes	
Apply SEJ to identified area of expertise	

(continued)

(continued)

Roles of Course Administrators	✓
Apply SEJ to support decision making	
Apply SEJ to identified real case studies	
Use identified software tool for processing information	
Build own tools to process information related to SEJ	
Use Excalibur for post-processing of the results and comparing aggregations of elicited distributions	
Course Content	
Introduction to online learning	
Introduction to underlying course philosophy, methodology and organisation	
Development of the EJ concept	
Schools of thought/areas of expertise in SEJ	
SEJ Methodology	
Comparing methods in SEJ	
Mathematical Principles behind the SEJ method	
SEJ applied to Natural and Biological systems	
SEJ applied to Non-natural catastrophes	
SEJ applied to Security	
Presentation skills	
Course Structure	
Includes a preliminary introduction to online learning	
Includes a range of graded units	
Includes cases from a range of areas involving risk situations	
Includes interdisciplinary units that merge a wide range of issues and perspectives from different areas of expertise	
Interlinks units into a system of topics and learning activities	
Adopts a range of pedagogical strategies—instructional, participative and contributory	
includes an online version of the whole course	
Preferably adopts a blended approach that combines face-to-face with online sessions	
Shows flexibility in combination of units and assessment modes	
Links theory with practice	
Includes continual formative and summative evaluation	
Course Delivery	
Includes face-to-face lectures	
Includes online lectures	
Includes face-to-face seminars/discussions	
Includes online peer discussions	

(continued)

(continued)

Roles of Course Administrators	✓
Includes online tutoring/discussions with experts and course lecturers	
Includes activities for small groups	
Includes collaborative projects	
Follows a customised course schedule	
Provides formative and summative evaluation exercises	
Includes activities to practice ‘Presentation skills’	
Assessment of Course Content	
Practice Exercises	
Short Questions (Quizzes)	
Multiple Choice	
Individual presentation about topic/technique	
Collaborative presentation about topic/technique	
Case-based problem analysis	
Seminars for collaborative assessment	
Practical assignment	
Exam/test with use of books	
Course Evaluation	
Questions to improve course content and delivery	
Questions to grade parts of and the whole course	
Written comments about course method	
Resources for Knowledge Communication	
Lecture notes	
Videoed lectures	
Instructional videos	
Reading list	
Resources for Collaboration	
List of discussion papers/articles	
Online forum	
Webinar App	
Resources for Reflection	
Compendium of Case-studies	
Case-based videos	
Videoed tutorial/Screencast for worked examples	
Online reading list	
Online environment dedicated for SEJ researchers	
Software Tools for Data Analysis and knowledge Creation	
MS Excel	

(continued)

(continued)

Roles of Course Administrators	✓
Vose Model Risk	
Pertmaster	
MathCad	
Mathematica	
R system/programming	
EXCALIBUR software	
Bayesian Belief Network Software	

14.6 Conclusion

Designing and developing this technology-enhanced course in SEJ involved not only the assembly and ordering of course content but it also attempted to add value and meaning to this field of knowledge. It aimed at simplifying and clarifying the theory and practice underlying SEJ. This could possibly entice and motivate participants to get involved in this professional field of practice or incorporate SEJ into their areas of application. The course was developed as an outcome of a design process that was informed by relevant theories of learning and guided by design principles using a recursive methodology of analysis followed by design and refinement. This ensures that teaching and learning are organised most effectively and engagingly.

To ensure an attractive and effective course design, the learning processes of novice learners were taken into account by analysing the characteristics of targeted learners, the nature of the content, the role of the community in shaping learning and integration of ongoing feedback and assessment. Ample opportunities for practice and tutor feedback were provided to guide the development of knowledge in action. Besides this, the course addressed the needs of participants as adult learners by recognising their professional experiences, which were integrated into the various course activities that addressed real-life challenges. The course was balanced in providing self-directed learning, giving an opportunity to participants to reflect on and analyse their own practice, and yet providing collaborative activities sharing knowledge and experience with experts and getting support from peers.

Taking into consideration a recurring suggestion in the front-end analysis, special attention was given to the design and delivery of the course in linking theory to practice. The integration of case studies from different fields of expert judgement emphasised moving away from 'knowledge about practice' and reflecting more about 'knowledge in practice'.

To accommodate for a range of learning styles the course included various forms of assignments, activities and assessments that allowed learners to interact and practice with content in multiple ways, on multiple cognitive levels and using different methods to assess learning. Within the limitations of the Moodle environment, the course was designed to be followed in a flexible way. Content, activities and experiences were organised in the *e*-learning environment in a sequential, cumulative and coherent format ensuring as much as possible the sequential movement from simple to complex, from concrete to abstract and from general to specific. Activities were designed to be as interactive as possible allowing for a range of levels of learning, learner entry points and experiences requesting the completion of a range of tasks (finding information, communicating, writing, reflecting, organising information, etc.).

The Moodle environment was not used just as a medium of delivery, but as a learning aid that provides opportunities for concrete, contextually meaningful experiences. Course participants could search for patterns, raise their own questions and construct their own models, concepts and strategies and share these with others.

Moodle was also used to provide flexibility in the delivery of course integrating face-to-face with distance aspects, providing as much as possible options and choice in terms of time, place and technology. Participants could access and use materials during the course but also when they get back to their own institutions or job situations. Most of the participants were determined to revisit the course content developed as a Moodle book with notes about all the topics covered during the course. They would also go again through the individual cases considered during the course and uploaded in Moodle. Since for most participants, it was the first time they did an elicitation exercise, many were looking forward to go through again the exercises recorded in Moodle. They applauded the fact that they could access the dedicated Moodle space at their own convenience when they had time in between study or work.

It is hoped that this course would help young scientists to engage more with the field of SEJ and become active and empowered learners in this field. On the other hand, it should provide field experts with a methodology to communicate and share their expertise in a more efficient and effective way.

14.7 Appendix 1: Detailed Analysis of the Pre-course Survey

The pre-course survey comprised ten questions targeting aspects of subject content and course method. The responses were analysed and categorised around key aspects of the course. Below we summarise the responses according to the different questions.

1. *Which Expert Judgement area/theme are you interested in?*

Participants expressed interest in the following topics organised under overarching themes:

- **Natural and Biological systems**
  - Natural disasters: volcanology including volcanic hazard/risk assessment; mitigation of Natural disasters
  - Natural effects (e.g. seismic, weather, corrosion, etc.)

  - Health sector
  - Medicine
  - Eco and Biological Systems
  - Veterinary Science

- **Non-natural catastrophes**
  - Nuclear hazards including reprocessing and clean-up/decommissioning areas and deliberate attacks on site impacting old containment-related structures
  - Fracking
  - Offshore structure maintenance
  - National Annexes/Regulations
  - Risk and reliability modelling in engineering, nuclear and other power systems.

- **Security**
  - National Security against Natural hazard
  - Information security
  - Reliability of Electronic Devices
  - Insurance

- **Methodology**
  - Mathematical frameworks behind the elicitation procedures
  - Dependence elicitation

2. *What is your aim/purpose for participating in the proposed SEJ course?*

Respondents mentioned both general aims and specific objectives for participating in the course. Some respondents commented that through the training course they would like to learn about SEJ and its application. Others were more specific referring to the development of particular skills and their application. For example, some considered learning techniques in SEJ to gain a broader range of different methods with associated strengths and shortcomings. Most of the respondents named elicitation techniques as specific skills to be developed by the training course. They were also specific about the application of these techniques in their field of specialisation or practice:

- To help reduce the high-hazard risks that nuclear sites face
- To apply them in my Ph.D. project about volcanic hazard assessment
- Different examples/case-studies in which expert elicitation has been used
- Apply process and procedures of structured expert judgement in projects and good practice
- In relation to the development of decision-making skills.

Some respondents indicated networking as an important outcome of the training course. Besides providing training and experience with regards to content and techniques, they expected that the training event would help them build contacts with field experts and with other early career researchers.

3. *What do you expect to learn/acquire by the end of the course?*

Respondents identified four categories of learning outcomes from this training course:

- theoretical discussion,
- learn about techniques in SEJ,
- application of such techniques,
- relevant tools in SEJ.

Regarding the theoretical dimension, respondents expected that the training course would give them a better idea of the main areas of academic dispute in this field of research, and which experts and groups belong to the different schools-of-thought. Besides that, they were interested in familiarising themselves with the theoretical underpinnings of SEJ and how to evaluate expert judgement elicitation.

Respondents expected to acquire a range of techniques including the understanding of the process for SEJ, basic and essential techniques for eliciting information from experts, how to define proper seed questions for experts and how to process experts' answers and translate them into numerical outcomes. Besides the acquisition of techniques respondents also consider the application of techniques as an important outcome of the training course. This includes understanding the applicability of SEJ in medicine and health, the application of SEJ together with other methods to support decision making, use of evidence in different aspects of decision-making including policy, clinical decisions and economic decisions. They also included the use of SEJ in real examples and case studies.

Respondents also consider the use of relevant tools as an important learning outcome. Reference was made to software tools for processing information. Some even suggested that the course should train them to build their own tools for processing information related to SEJ. The tool Excalibur was suggested to be used for facilitating expert elicitation and to process the results.

4. *Which are the main strengths which you would like to see in the proposed course?*

Respondents identified eight key features of a training course in SEJ. These are:

- A strong theoretical framework that describes the development of the SEJ concept and motivation, gives an overview and compares current methods for SEJ and explains the mathematical principles behind the SEJ method.
- SEJ Techniques including practice on elicitation preparation, elaboration of outcome data from SEJ session and checklists for the SEJ processes.
- Application mainly providing the possibility for participants to apply the SEJ principles to their area of specialisation.
- Customisation especially providing course participants with the possibility of choosing from different units according to the learner's interest.
- Diversity in experience by providing participants the opportunity to learn from what is happening in different areas involving risk situations.

- Inter-disciplinarity in course units built around a wide range of issues and perspectives from different areas of expertise.
- Interaction by providing the opportunity to communicate and get feedback (face to face and online) from different participants, ESRS and also experts/course lecturers.
- Collaboration with field specialists within the COST Action involving the acquisition of expert knowledge and guidance in applying it to specific contexts; it also involves establishing strong inter-participant cooperation.

5. *Which are the main weaknesses of the proposed course which you consider should be avoided?*

The following potential weaknesses were identified by respondents with regards to the content and process of the course:

- The course content should be comprehensible to non-experts and decision-makers in explaining and justifying the use of SEJ techniques by adjusting the level of difficulty.
- The course should avoid sophisticated mathematical procedures, for example, by breaking down the classical method and explaining its formulas in a guide for non-experts.
- Course content should be relevant avoiding situations where participants have to do a lot of work which is not applicable to them.

With regards to the process, the proposed training course

- Should adopt a systems approach in SEJ controlling for unstructured or fragmented methodologies.
- The adopted pedagogical approach should integrate theory with practice and avoid over-reliance on theoretical lecturing and readings.
- The training course should adopt a flexible approach giving participants the possibility to follow the course at their convenience and not be bound by a particular period.

6. *Think about your learning characteristics. What is your prior experience in SEJ? What are your course expectations? What is your experience in online learning?*

Prior experience in SEJ:
Four categories of prior experience in SEJ were identified amongst respondents.

- The first category declared that they have very little experience in SEJ: ‘I have no specific experience in SEJ, just scattered information gathered from sparse papers’.
- The second category is those that came across the use of SEJ in their job thus compelled to adopt a more hands-on approach that could be lacking in theoretical underpinnings. Comments included: ‘Basic on-the-job training with no reference to SEJ models or procedures’; ‘Involved in SEJ but never facilitated on my own, or used the software to analyse the results’.

- The third category includes those who learned about SEJ as part of their academic course in their professional development: 'Experienced SEJ during my academic study as part of the PhD as recommended by tutors' and 'Studied for PhD in future volcanic activity and worked with different methods (Cooke Classical Model, the Expected Relative Frequency model of Flandoli et al. (2011), and also a Simple Equal Weights model); worked with different scoring rules and uncertainty distributions'. Through their studies others feel more competent in this area: 'Have quite a profound knowledge about SEJ, even though elicitations were mainly made within my own university so that I am missing some actual experience'.
- The final category comprises those with a relevant theoretical background which can facilitate SEJ yet they did not have the opportunity to practise SEJ: 'I have a maths/statistics background so am comfortable with the theory of the different approaches to SEJ. I have never carried out a practical elicitation session.'

This shows the heterogeneity in prior experience of potential course participants in SEJ. Consequently, a training course in SEJ should adopt a differentiated approach providing graded course content, together with customised and flexible learning approaches.

Course expectations of respondents:

With regard to content, respondents expected the training course to offer them a good grounding in theoretical underpinnings making specific reference to the probabilistic approach to Expert Judgement. They also expressed the need to ground theoretical frameworks within authentic real-life situations.

With regards to process, reference was made to the course assessment procedure. Respondents expected that course assessment would be based on short questions and/or multiple-choice (that are not tricky), avoiding assessment by assignments.

Respondents experience in online learning

Respondents demonstrated a continuum of experience in online learning. Some declared that they had never experienced online learning: 'I never had an experience of on-line learning'. Others declared that they took some online courses:

- 'Completed some online courses such as Coursera etc';
- 'Currently following an online course on Moodle';
- 'I am using an on-line learning platform for learning Dutch: it works fine for doing examples and exercises and have an on-line literature reference';
- 'I have done many online courses as part of my day-to-day work—some were excellent and very well balanced, others could have been improved in various ways (getting this right appears to be something of an art-form)'.

Some respondents were very competent in online learning and conversant with online course design:

- 'I teach on Distance and Flexible learning courses using Moodle based sites.'
- 'Course should avoid audio lecturing as this limits customisation'.

This varied experience in online learning necessitates the adoption of a differentiated approach in dealing with the online learning experience. Some course participants may need a gradual introduction and mentoring into online learning. Others need to be assured about the design of the course describing the underlying philosophy, methodology and strengths. Yet with others, an evaluative approach should be adopted asking them to evaluate its content and design while going through the course.

7. *Which teaching and learning methods would you suggest to be employed in the course?*

Respondents suggested the following pedagogical approaches:

- All respondents (N = 20) showed keen interest in 'Case-based investigations' that link theory to practice.
- 75% (N = 15) of respondents suggested the use of 'Online lectures' to teach basic principles.
- 75% suggested that 'On-line and Offline discussion' should be ongoing throughout the course and for following specific topics.
- 75% recommended 'Small group work' possibly at workshops during COST Action's meetings;
- 75% commented that a 'Project-based approach' may prove quite difficult to realise unless a multidisciplinary approach is adopted.
- 50% suggested that the course should provide for 'Individualised self-paced learning' since some participants will be working practitioners;

Respondents also commented on other pedagogical strategies. For some course participants, *e*-Portfolios may prove to be a difficult approach to follow. Other suggestions include the use of Blogs for posting questions and the possibility of having 'Personal tutoring' during the course. Ideally, the course should adopt a 'Multi-method distance-learning approach' that combines online with face-to-face interaction, and digital resources with real-life authentic contexts.

8. *Which assessment methods do you suggest to be employed in the course?*

Respondents suggested the following assessment methods to be included in a training course:

- 60% of the respondents suggested 'Quizzes' for assessing course content and process
- 60% proposed the use of 'Exercises' both for consolidating the acquisition of knowledge and practicing the application of content and skill
- 60% considered individual and collaborative 'Presentations' as a good means for assessing both content and process
- 50% consider 'Problem analysis' as an important assessment component
- 50% suggested the use of 'Seminars' as a tool for collaborative assessment
- 40% referred to 'Assignments' as a practical form of assessment

- 25% advocated a strategy that embeds assessment within 'Coursework'
- Only one respondent (5%) suggested:
  - the use of *e*-Portfolios for assessment
  - Assessment through tests/exams. A number of respondents were adamant that the unit/course assessment should not be through exams. In case this mode of assessment is adopted, exams should be done with access to books.

Besides the listed assessment modes, respondents were asked to suggest any other relevant form of assessment. Practical sessions emerged the most significant mode of assessment, especially considering the repeated plea that the course should link theory to practice.

9. *Which resources do you suggest should be used/included in the course?*

The resources mentioned by respondents can be grouped into the following four categories:

- Knowledge communication
  - Lecture notes
  - Videoed lectures
  - Instructional videos
  - An 'accessibility-based' website aimed at a general audience (i.e. decision-makers as potential customers) providing easy-to-understand materials/activities about SEJ. This should be either the COST website or a dedicated website linked to the COST website.
- Collaboration
  - Reading of identified papers about EJ topics followed by discussion in the group.
  - Discussion fora about identified themes/documents.
  - Live online discussions with "lecturers" including mentoring.
  - Webinars
- Reflection
  - Presentations of case-studies
  - Case-based videos
  - Video tutorial for worked examples
  - Research agenda
  - W-based reading and research
  - Sharing specialised material for in-depth analysis in a dedicated website for researchers.
- Data Analysis and knowledge creation tools—Software tools used to analyse data, identifying trends and quantifying parameters that help decision-making:

- Software for SEJ data analysis
- MS Excel (for various statistical analysis)
- Vose Model Risk (an Excel add-on that enables Monte-Carlo analysis. This can be considered as a more advanced version of @risk)
- Pertmaster (a Primavera-based risk-analysis program for Monte-Carlo analyses on detailed schedules).
- MathCad
- Mathematica
- R system/programming
- EXCALIBUR software
- Bayesian Belief Network Software (precision tree or similar).

The training course should help participants identify open-access software and facilitate access to expensive commercially available software. Some very expensive software programs can be made available through collaborative programmes/universities joint ventures.

10. *Other suggestions for the training course proposed by survey respondents*

Respondents to the survey made the following suggestions:

- Since young scientists will have to communicate their ideas, findings and proposed judgements to various audiences, it is important that the course provides the opportunities to develop presentation skills and communication techniques.
- Respondents also suggested the development of a mentoring scheme by which novice scientists will be guided in applying SEJ in their respective fields.
- Participants should be awarded a certificate following the training course.
- Specialised training tools and resources to communicate complex information (case-based videos, comic-books and video-games) should be made available.
- Already-existing resources that relate in some way to EJ or that EJ could possibly enhance in some way should be organised and made available.

14.8 Appendix 2: Evaluation of the Course

An evaluation exercise was carried out after running the course on two separate occasions. During the last session of each course, a questionnaire was made available on Moodle, and participants were asked to complete it. The following is the feedback obtained for each item of the questionnaire.

Did the course meet your expectations? Explain.

Most agreed that the course met their expectations for the following reasons:

- I needed a lot of practice exercises and the group exercise in particular was extremely useful.
- I felt a real improvement on my understanding of expert judgement analysis, and the available software tools.

- I was expecting to get: some background in EJ, an introduction into the classical method, some explanation of how to apply it, some experience in applying it, and an overview of other possible methods.
- As I expected, the course has been very interesting both in terms of the theory and practical cases. I had little experience dealing with expert judgement however I had not dug as deep as the course did.
- The training course met my expectations, primarily due to: a) Meeting other researchers. b) Topic selection and discussions around EJ. c) Very good lectures and follow-up. d) The inter-disciplinary nature of the participants.
- Prior to the course I was concerned that the course would be too focused on mathematical methodology, but this was not the case. Ideally before the course the details of the course content are communicated so that the applicants are clearly guided. I felt the interaction with the other participants very useful.
- The course was well structured and there were enough examples supporting the methodology.

Which activity did you like most?

The most liked activity was the Group Elicitation Exercise (78%), followed by lectures, Case studies and discussions. Related comments include:

- While the case studies were informative and the working exercise gave a sense of what Expert Judgement is, the lectures discussing the theory were the most I liked.
- The group exercise, where we had to practically set up an elicitation with seed questions, was extremely useful.
- Must be seen as a whole. Lectures, discussions and the exercise all together made it a very interesting, useful and motivating course.

Name any topic/s covered in this course that you recommend should be included in future courses?

Participants proposed the following topics to be included in future courses: all theory lectures, using the EXCALIBUR software, real-life cases from different fields, Elicitation exercise, the applicability of SEJ in different fields/problem situations, current status and challenges within EJ.

In your opinion, what are the main strengths and weaknesses of the training course?

Participants pointed out the following strengths:

- The course involved authoritative scholars/researchers and the experts proposing the theoretical framework were members of the organising team and available to participants.
- The course included high-quality lectures and offered a good compromise between theoretical and practical part. ‘The quality of presentations and lecturers really pushed me to dig further into EJ.’

- The practical aspect of the course providing both examples of practice and discussion with people who apply them in various situations, thus combining others' practice with own experience.
- The interaction with the tutors was well planned, organised and executed.
- The complete overview of SEJ including different types of methods, the pros and cons of each method, the background mathematics of the classical method and how to use it.

With regard to weakness, many participants declared that they could not identify any weaknesses.

One participant pointed out that it would have been ideal to have the presentations beforehand.

Some examples should be added where the classical method was used to estimate failure probabilities, showing the example seed questions.

Name topics and activities that in your opinion should NOT be included in future courses.

- Two participants agreed that all proposed topics should be included in future courses.
- One claimed that they could not name any one topic to exclude.
- Reference to specific projects can be changed each time to ensure discussion of recent work and application.
- Examples involving a lot of mathematical derivation (which was not the case in this course) should ideally be left separate and optional.

What are your suggestions to improve the programme (course schedule)?

Two participants declared that they like it as is. Other comments were: 'It was really intense, but I can't see how to improve this…. maybe an extra day.' This was confirmed by another participant: 'Extend the course by at least one day. This would enable more discussions, and maybe insight into where EJ goes, challenges and research gaps etc.'

Some comments were task-specific:

'Maybe present an actual elicitation in more detail. The choices made in designing the elicitation, the process followed, the steps taken from start of the project until finish, lessons learned, etc.'

'Maybe more details on the applicability of Bayesian methods in order for the naïve participants to understand their applicability better.'

'The afternoon time on the last day of the course could be used to finish the group activity, this way there would be more time in executing the activity and learning from it.'

References

Bereiter, C. (2002). *Education and mind in the knowledge age*. New Jersey: Lawrence Erlbaum Associates.

Bonanno, Ph. (2011). A process-oriented pedagogy for ubiquitous learning. In T. Kidd, & I. Chen, (Eds): *Ubiquitous Learning: A Survey of Applications, Research, and Trends.* Information Age Publishing. pp. 17–35.

Dillenbourg, P., Baker, M., Blaye, A., & O'Malley, C. (1996). The evolution of research on collaborative learning. In E. Spada & P. Reiman (Eds.) *Learning in Humans and Machine: Towards an interdisciplinary learning science*, pp. 189–211. Oxford: Elsevier.

Downes, S. (2012). Connectivism and connective knowledge—essays on meaning and learning networks. eBook published under a Creative Commons License. ISBN: 978-1-105-77846-9.

EFSA. (2014). Guidance on expert knowledge elicitation in food and feed safety risk assessment. *EFSA Journal, 12*(6), 3734–4012.

Flandoli, F., Giorgi, E., Aspinall, P., & Neri, A. (October 2011). Comparison of a new expert elicitation model with the classical model, equal weights and single experts, using a cross-validation technique. *Reliability Engineering & System Safety, 96*(10), 1292–1310.

French, S., & Maule A. J. (1999). Improving risk communication: scenario based workshops. Risk Communication and Public Health: Policy Science and Participation. P. G. Bennett and K. C. Calman. Oxford, Oxford University Press, pp. 241–253.

Frith, C. D. (2007). The social brain? *Philosophical Transactions: Biological Sciences, 362*, 671–678. The Royal Society. Published online 24th Jan 2007.

French, S., & Maule, A. J. (2010). Exploring and communicating risk: scenario-based workshops. Risk Communication and Public Health. 2nd Edn. P. G. Bennett, K. C. Calman, S. Curtis and D. Fischbacher-Smith. Oxford, Oxford University Press, 299–316.

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions: Biological Sciences*, *358*(1431), 459–473. The Royal Society.

Frith C. D., & Wolpert D. M. (2003). Decoding, imitating and influencing the actions of others: the mechanisms of social interaction. *Philosophical Transactions: Biological Sciences*, *358*(1431), 431–434. The Royal Society.

Jost, J. T., Kruglanski, A. W., & Nelson, T. O. (2002). Social metacognition: An expansionist review. Article prepared for a special issue of *Personality and* Social Psychology Review. Source: http://gobi.stanford.edu/researchpapers/detail1.asp?Paper_No=1464.

Kafai, Y., & Resnick, M. (Eds.). (1996). *Constructionism in practice: Designing, thinking, and learning in a digital world*. Mahway, New Jersey: Lawrence Erlbaum Associates, Publishers.

Kreijns, K., Kirschner, P. A., & Jochems, W. (2003). Identifying the pitfalls for social interaction in computer-supported collaborative learning environments: a review of the research. *Computers in Human Behavior, 19*(2003), 335–353.

Sabelli, N. (2008). *Constructionism: A New Opportunity for Elementary Science Education.* DRL Division of Research on Learning.

Siemens, G. (2004). Connectivism: A learning theory for a digital age. Elearningspace.org 12 December, 2004 URL: http://www.elearnspace.org/Articles/connectivism.htm. Retrieved January 2007.

Siemens, G. (2005). Connectivism: A learning theory for the digital age. *International Journal of Instructional Technology and Distance Learning*, *2*(1), 3–10.

Siemens, G. (2006). Knowing knowledge. Available http://www.elearnspace.org/KnowingKnowledge_LowRes.pdf.

Tetlock, P. E., Gardner, D. (2016). *Superforecasting: The art and science of prediction.* Random House.

Chapter 15
Expert Judgement for Geological Hazards in New Zealand

Annemarie Christophersen and Matthew C. Gerstenberger

Abstract Expert judgement is important for the short- and long-term assessments of natural hazards in New Zealand, contributing to their risk analyses and informing decision-making. The problems are complex and usually require input from experts from different sub-disciplines. Expert judgement, like all human cognitive processes, is prone to biases. Therefore, we aim to use methods that are robust, transparent, reproducible and help reduce biases. The Classical Model treats expert opinion as scientific data and its performance-based weighting of experts allows us to measure the uncertainty of a quantifiable problem. We have developed a protocol for risk assessment, including structured expert judgement, which is centred around workshop-style interactions between experts to share knowledge. The protocol borrows heavily from the framework for the risk management process of the International Organization for Standardization. We outline seven recent applications of structured judgement, mostly in seismology and volcanology. Most of them use the Classical Model to aggregate the expert judgement. We discuss challenges and insights, concluding that developing an optimal protocol for expert judgement is a continuing journey.

15.1 Introduction

New Zealand lies in the south-west Pacific Ocean, along the junction between the Pacific and Australian tectonic plates (Fig. 15.1). The collision of the tectonic plates causes rugged mountains, active volcanoes and frequent earthquakes. Secondary geological hazards arise from landslides, tsunamis and flooding. A damaging earthquake can occur anywhere in New Zealand and a volcanic eruption can cause ash fall over most of the North Island. Given the small size of the country and the interdependencies of infrastructure, logistics and business, a major earthquake or volcanic eruption can affect the whole society. Assessing these hazards, either as immediate

A. Christophersen (✉) · M. C. Gerstenberger
GNS Science, 1 Fairway Drive, Avalon, New Zealand
e-mail: A.christophersen@gns.cri.nz

A. M. Hanea et al. (eds.), *Expert Judgement in Risk and Decision Analysis*,
International Series in Operations Research & Management Science 293,
https://doi.org/10.1007/978-3-030-46474-5_15

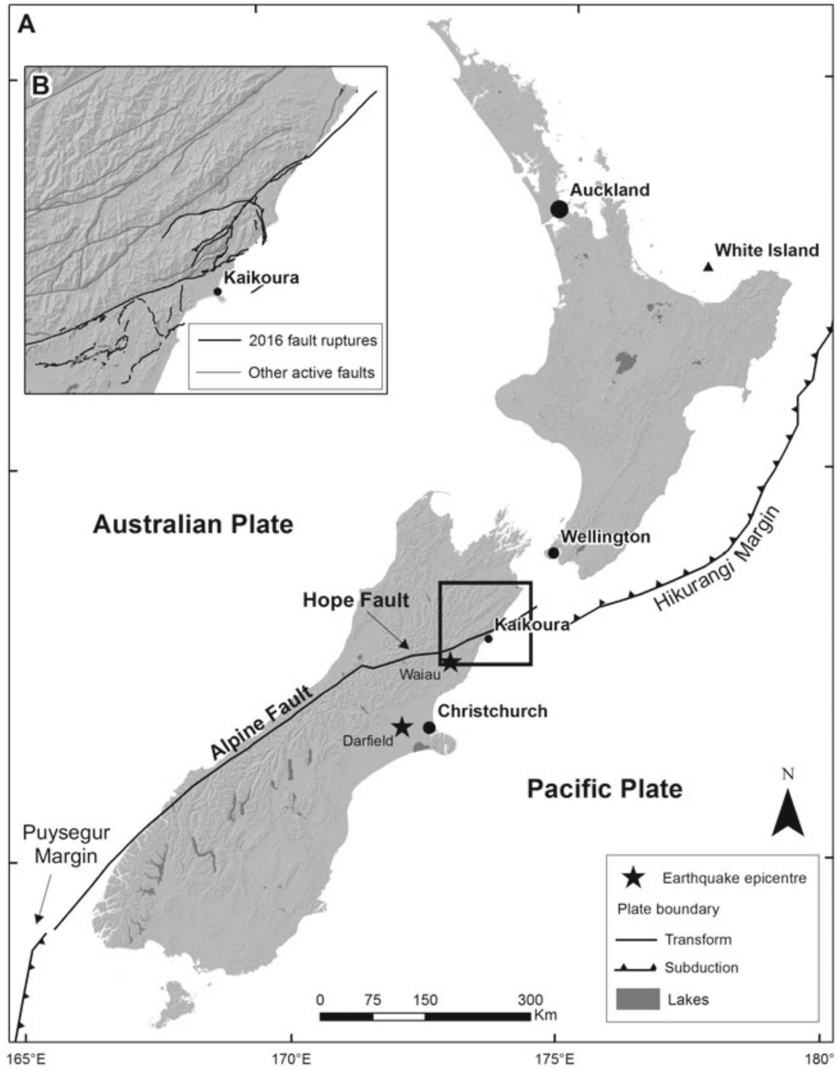


Fig. 15.1 Map of New Zealand; (A) showing the position on the plate boundary, with the Puysegur Margin in the south-west, the Alpine and Hope Faults in the South Island and the Hikurangi Margin in the east of the North Island. The stars indicate the locations of the two major earthquakes that initiated project 2–4 in Table 15.1. Also show is White Island volcano (project 6)

threats or in the long term, typically requires expert judgement; in part, this requirement is due to the low probability of major events and the limited data available for model building.

GNS Science advises the New Zealand government on geological hazards and contributes to the management of public information on geological hazards and associated emergencies (New Zealand Ministry of Civil Defence and Emergency Management 2015). It has similar functions to geological survey institutions in other countries. GNS Science manages the GeoNet system for the detection of earthquakes, land movement, volcanic activity and the potential for local-source tsunamis. GeoNet coordinates responses to natural hazard events.

Whenever the earth rumbles, rolls or fumes, scientists gather at the GeoNet offices to work out what has happened, is happening and might happen next. Scientists from different sub-disciplines share their data and knowledge to interpret what is going on. This informal expert judgement, for example, when complemented by rigorous statistical models for earthquake (Christophersen et al. 2017 for an overview), has been very effective in providing scientific advice to New Zealand government agencies, the media, public and other stakeholders. In contrast to understanding what is going on during an event response, long-term hazard models estimate the probability of occurrence of a specific hazard, in a specific future time period, as well as its intensity and area of impact. These models provide a basis for decision-making aimed at reducing the impacts of geological hazards to society. The development of long-term hazard models also involves elements of expert judgement.

Expert judgement, like most human thinking and judgement processes, is prone to biases that are often hidden from awareness (Bang and Frith 2017). Kahneman (2011), who jointly with Tversky pioneered the study of biases (Tversky and Kahneman 1974), describes the brain as consisting of two systems. System 1 is almost automatic and instinctive, while System 2 deals with rational thought and conscious decision-making. Working with System 2 requires energy and focus; this is mentally draining. The brain aims to preserve energy and preferably uses System 1 that takes many short-cuts, called heuristics, to process information and reach conclusions. Heuristics allow for faster processing of information but can cause biases and flawed decision-making.

For the development of robust geological hazard models and to be able to give the best possible scientific advice, we are interested in structured expert judgement (SEJ). The purpose of SEJ, as defined by Hanea et al. (2018), is to (1) address questions that theoretically could be measured or calculated if there was sufficient time and enough data, (2) follow reproducible and transparent rules, (3) anticipate and aim to mitigate biases, (4) be thoroughly documented and (5) provide opportunities for empirical evaluation and validation. Given the complexity of the problems that we address in geological hazards, we do not expect experts to reach consensus on any given question. Quite the contrary, we are keen to explore the uncertainty of a question of interest. In many cases, we need to estimate the likely occurrence of low-probability events. This makes it challenging to measure the success of any protocol and test for reproducibility. Therefore, we are looking for a method that has robust foundations and has been well scrutinized with evidence of skill in other applications.

The Classical Model treats expert judgement as scientific data and follows scientific principles from probability and statistics (Cooke 1991). It is built on rational consensus, in which experts agree on the method of aggregating individual judgement rather than seeking consensus on any specific problem. The method weights experts' judgement based on the experts' ability to estimate uncertainty for questions

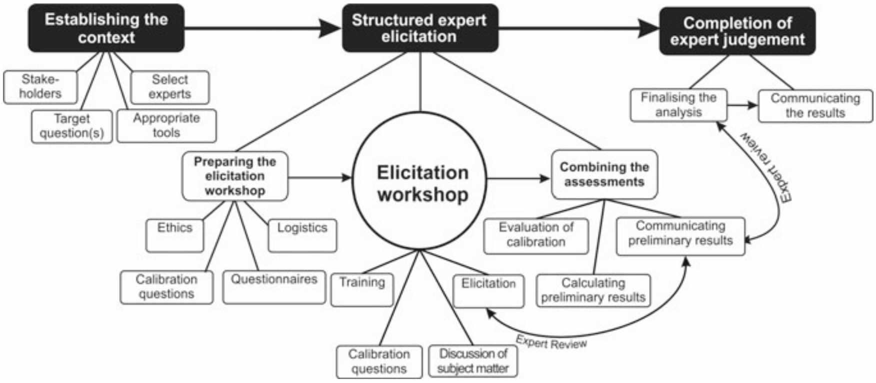

Establishing the context
Structured expert elicitation
Completion of expert judgement
Stake-holders
Select experts
Target question(s)
Appropriate tools
Finalising the analysis
Communicating the results
Expert review
Preparing the elicitation workshop
Elicitation workshop
Combining the assessments
Ethics
Logistics
Evaluation of calibration
Communicating preliminary results
Calibration questions
Questionnaires
Training
Elicitation
Calculating preliminary results
Expert Review
Calibration questions
Discussion of subject matter

Fig. 15.2 Our suggested protocol of a structured expert judgement with an elicitation workshop

with known answers, so-called calibration or seed questions (Cooke 1991; Quigley et al. 2018). The Classical Model suits our requirements well. We have developed a protocol for applying the Classical Model in workshop-style sessions for experts to share their knowledge and understanding of the problem so that they can best estimate the answer, including the uncertainty, to the problem at hand (Fig. 15.2).

In this chapter, we provide an overview of the biases that we try to mitigate. We introduce the protocol that we have used for multiple elicitations in the last few years, in which the Classical Model is ideally applied, and which is centred around workshop-style interactions. The main part of this chapter introduces seven recent examples of expert judgement applied to seismic and volcanological hazards. We discuss some of the challenges encountered as well as the benefits of using SEJ.

15.2 Developing a Protocol for SEJ

We began developing our procedures for SEJ within the context of risk assessment. Between 2010 and 2013, GNS Science led the development of risk assessment methods for CO2CRC (Gerstenberger et al. 2012). CO2CRC is Australia's leading carbon capture and storage research organization (CO2CRC 2011) and operates a study site in the onshore Otway Basin in south-western Victoria, Australia, for injection experiments (Jenkins et al. 2012). As part of the risk method development, we investigated Bayesian networks as tools for modelling complex problems (Gerstenberger et al. 2015) and explored SEJ methods for working with experts when data are unavailable or sparse. In Sect. 15.2.1, we provide an overview of common biases to be avoided, followed by a rational for the workshop-style expert interaction in 15.2.2, and a section on the Classical Model for assessing the risk and quantifying uncertainty in 15.2.3. Our expert judgement protocol is described in Sect. 15.3.

15.2.1 Common Biases

There is a large body of literature investigating biases, their causes and possible ways of mitigating them. Broadly speaking, biases fall into three categories with some overlap between them. Cognitive biases are mistakes in reasoning, evaluating, remembering, or other cognitive process. Motivational biases occur when the judgement is influenced by the expectation of the results and outcomes. Group biases may occur due to group dynamics. Montibeller and von Winterfeldt (2015) provide a recent review on cognitive and motivational biases and their mitigation in decision and risk analysis. More recently they have extended their analysis to include group biases (Montibeller and von Winterfeldt 2018).

The boundaries between different categories of individual biases are not always clear cut. For example, confirmation bias, "the seeking or interpreting of evidence in ways that is partial to existing beliefs, expectations, or a hypothesis at hand" (Nickerson 1998), is classified as motivational bias by Montibeller and von Winterfeldt (2015) while Kunda (1990) and Westen et al. (2006) discuss the cognitive aspects of confirmation bias. Nickerson (1998) discusses how confirmation bias results from not considering alternative hypotheses and that in turn can be associated with overestimating the accuracy of one's judgement. A narrow range of variation on estimated values (over-precision) is associated with overconfidence bias (Montibeller and von Winterfeldt 2015). Overconfidence bias is also used to describe the observation that people overestimate their own skill (overestimation) and that they believe they are better than others (over-placement). Over-precision, i.e. not appreciating the uncertainty of one's knowledge, is more prevailing than either overestimation or over-placement (Moore and Healy 2008), and is referred to as overconfidence in this chapter.

Anchoring is a bias that occurs when the assessment of a numerical value is based on an initial estimate and is not sufficiently adjusted to accommodate other information (Tversky and Kahneman 1974). This bias also applies when assessing confidence intervals and thus links with overconfidence. In short judgement can go wrong in many ways.

Montibeller and von Winterfeldt (2015, 2018) provide extensive lists of biases in the above-mentioned categories, and mitigation options. One bias missing from their compilation of cognitive biases is authority bias (Milgram 1963, 1974), which refers to the inclination to follow the lead of an authority figure. However, once the authority is challenged (by other group members or the facilitator, if in a workshop-style format), it is easier for individuals to disobey the authority (Milgram 1974). Groups can reinforce individual biases; in particular, if all experts view a problem from a similar perspective, flaws can be enhanced (Kerr and Tindale 2011). However, group processes can also have advantages in surmounting biases (Bang and Frith 2017).

Careful facilitation, good elicitation design and training of the experts can help to mitigate some cognitive biases. Motivational biases are challenging to mitigate in an individual. The best approach to achieving an unbiased final judgement is to include a number of experts with different viewpoints, challenge viewpoints in discussions and encourage alternative opinions. It is also useful to let experts provide their judgement confidentially to avoid peer pressure.

It is noteworthy that individuals generally only consider one hypothesis at a time and tend to assume that this hypothesis is true (Nickerson 1998). Consequently they look for evidence to confirm this hypothesis. Nickerson (1998) suggests that this form of confirmation bias can be mitigated by training experts to think of alternative hypotheses early in the elicitation process. This supports workshop-style sessions similar to our response to major earthquakes, where all streams of evidence, be it in the form of data or models, are presented and discussed prior to eliciting judgement.

15.2.2 Workshop-Style Expert Interaction

There are a number of advantages in group processes: they allow for the pooling of relevant information and for error checking, and can enhance individual task motivation (Kerr and Tindale 2011). Recent research confirms that groups tend to perform better than most individuals (Hemming et al. 2018). A recent literature review on common problems of decision-making in individuals and groups found that group processes have advantages in surmounting biases, exploring good models of the world and finding good solutions to problems (Bang and Frith 2017). In particular, discussions in small groups and without time pressure benefit from the knowledge held by individuals (Bang and Frith 2017 and references therein). This is consistent with our observations from the GeoNet-led earthquake responses, where experts from different sub-disciplines come together, unfortunately under time pressure, and share their knowledge to understand a complex problem. In workshop-style sessions, each expert represents the key findings from their sub-discipline. This allows for informed discussion and sharing of all relevant information. In such situations, experts can assess the arguments and form opinions. Research shows that individuals are more likely to change their mind for a well-argued opinion than for one stated with high confidence (Trouche et al. 2014). Other advantages of workshop-type interactions, going beyond accuracy of the final result, include that individual group members can voice their opinions, which helps to foster feelings of fairness/justice and inclusiveness, and increased legitimacy of and willingness to rely on the results.

Disadvantages of group interaction can be the pressure to conform to a majority view, the risk of being led astray by a dominant leader and the inattention to novel and unshared information (Kerr and Tindale 2011). The first two concerns may be mitigated by encouraging open discussion, in which the facilitator challenges dominant experts and thus makes it easier for the experts to disagree with the dominant person. Encouraging different viewpoints and exploring alternative hypotheses may also mitigate confirmation bias.

15.2.3 The Classical Model to Quantify Uncertainty

The Classical Model is a method for SEJ that mathematically aggregates expert judgements, based on the experts' ability to assess uncertainty. Experts provide their uncertainty for two types of questions: target questions and calibration questions. Target questions are the variables that cannot be adequately answered with other methods and thus require expert judgement. Calibration questions are similar in nature to the target questions and have values that are not known to the experts during the elicitation but become known during the analysis or are known to the analyst. Experts provide their uncertainty as percentiles, typically the fifth, fiftieth and ninety-fifth. Thus, they are asked for their best estimate and the 90% credible range for the true value to lie within. We tend to ask for an 80% credible range, i.e. for the tenth, fiftieth and ninetieth percentile, in an attempt to counterbalance the experts' overconfidence.

There are two measures to evaluate the experts' performance: statistical accuracy, also referred to as calibration, and informativeness (Cooke 1991). The statistical accuracy is the probability with which one would falsely reject the hypothesis that the experts answer according to the multinomial theoretical distribution determined by the inter-quantile intervals. Theoretically, calibration can take values between 0 and 1 but in practice they hardly ever get close to one, and most individual experts achieve a calibration below 0.05, see Chap. 10, this volume. Cooke (1991) defines a quantity that is based on how an expert estimates uncertainty over the number of calibration questions in relation to the percentiles of the credible range. For example, with the credible range of 80% mentioned above and ten calibration question, the true answer to the calibration question is expected to fall below the tenth percentile for one question, between the tenth and the fiftieth for four questions, between the fiftieth and the ninetieth for another four questions and above the ninetieth for one question. A transformation of this quantity is distributed like a chi-square random variable with three degrees of freedom. The calibration measures how this quantity diverges from the theoretical distribution. However, Chap. 10, this volume, illustrates that calibration does not clearly distinguish between well-calibrated experts. For example, two experts with nearly identical assessments on ten calibration questions can have a 0.44 difference in calibration score. On the other hand, experts, who are not well calibrated, can have a very low calibration score. Cooke (1991) argues that ten calibration questions and a significance level of 0.05 are sufficient to distinguish whether an expert is well calibrated or not.

The second measure of performance is informativeness. For example, an expert might provide very wide uncertainty intervals and by this potentially achieve good calibration but be not very informative. To calculate informativeness, an intrinsic range is determined for each calibration and target question. This covers the lowest and highest uncertainty estimates of all experts, and the true answer for each individual question plus an overshoot of each interval to capture the possible minimum and maximum of the interval. The informativeness of an expert is measured by comparing the estimated uncertainty widths with the intrinsic range and scaling the divergence using either a uniform or log-uniform distribution that covers the intrinsic range. Details and illustration of the methods are given by (Cooke 1991; Chap. 10, this volume; Quigley et al. 2018). Informativeness is a strictly positive function; the higher the score, the more informative an expert is. Typical values for informativeness can be found in the TU Delft expert judgement data base (Cooke and Goossens 2008). For 322 experts across the pre-2006 study the informativeness ranged from 0.25 to 3.81, with half of the experts scoring above 1.47, Chap. 10, this volume.

The experts' calibration and informativeness can be combined in different ways to derive weights to apply to the target questions (Cooke 1991). The combination of experts' weight is called the decision-maker. Different types of weights are available: global, itemized and optimized. Global weights average each expert's informativeness across all calibration questions. Raw weights are then calculated for each expert. Experts with a calibration score below a selected level of, for example 0.01, may be given a weight of zero, if a cut-off is chosen. The weights are then normalized across all experts with non-zero weights.

Itemized weights take advantage of the fact that informativeness for any expert can vary across questions while calibration is usually calculated over all calibration questions. Itemized weights are calculated for each question and each expert separately as the product of the informativeness on that question and the calibration score over all calibration questions. Again, experts with a very low calibration score may be given a weight of zero and therefore be excluded from the normalization of weights.

Optimized weights are calculated by varying the level of the calibration cut-off to maximize the score of the decision-maker. This may lead to some experts getting zero weights. However, zero weight does not mean zero value because all experts contribute to the intrinsic range. Figure 10.12 in Chap. 10, this volume, gives an example, in which the optimized decision-maker uses only two of ten experts, but the exclusion of one particular zero-weighted expert would lead to a significant reduction in the performance of the decision-maker.

The weighted combination of the experts' judgements is applied to the calibration and the target questions. This way, the Classical Model validates both individual expert assessments and the performance-based combinations against observed data.

As further discussed in Sect. 15.3.2.2, we usually administer the calibration questions in the early stages of the workshop to be able to show the initial results to experts before they finalize their answers to the target questions. This is against the standard recommendations to make the calibration questions as indistinguishable from the target questions as possible to be unbiased performance measures (Cooke 1991; Quigley et al. 2018). However, there are two advantages in showing experts the calibration results. While individual experts tend to be overconfident, i.e. they provide too narrow uncertainty intervals and therefore are not well calibrated, the decision-maker tends to find the true value of the calibration question. Seeing that the decision-maker of the Classical Model finds the answers that the individuals struggled with builds confidence in the method. Secondly, as a consequence of realizing their own overconfidence, we find that experts widen their confidence intervals when answering the target questions. This way we are likely to better measure the uncertainty of the target questions, because the experts have learned to counter-bias their overconfidence. On the down-side, the performance on the calibration questions may not then be a true reflection of the performance on the target question(s).

15.3 A Risk-Based Protocol

The International Organization for Standardization's principles on risk management (ISO 2009) provides a useful framework to adapt to an expert elicitation protocol. The risk management process has three main components: (1) establishing the context, (2) risk assessment and (3) risk treatment. "Communication and consultation" and "monitoring and review" inform each step of the process. The risk assessment is split into the sub-components of risk identification, risk analysis and risk evaluation. We have modified the ISO framework for risk assessment in carbon capture and storage (Gerstenberger and Christophersen, 2016, project 1, Table 15.1) and volcanic eruption forecasting (Christophersen et al. 2018, project 6, Table 15.1). Here we adapt the same framework to a protocol for structured expert judgement (Fig. 15.2). There

Table 15.1 An overview of recent expert elicitations, the methods used and the roles of the authors. MG stands for Matt Gerstenberger and AC for Annemarie Christophersen

	Project	Method(s) used	Roles
1	Risk assessment in carbon, capture and storage	Classical Model in workshop-style setting	Project leader, workshop facilitator, analyst (MG) coordinator of calibration questions, analyst (AC)
2	Time-dependent seismic hazard model for the recovery of Christchurch 2a source model 2b GMPE model	Classical Model in workshop-style setting	Project leader, workshop facilitator, analyst, coordinator of calibration questions (MG) Contributor to calibration questions (AC)
3	Probability of large earthquake following Kaikōura earthquake	Informal elicitation of probabilities and uncertainties in workshop-style setting	Project leader, facilitator, analyst and expert (MG) and expert (AC)
4	Probability of large earthquake following Kaikōura earthquake	Classical Model in workshop-style setting	Project leader, facilitator, analyst (MG), coordinator of calibration questions, analyst (AC)
5	Australian national seismic hazard model 5a source model 5b GMPE model	Classical Model in workshop-style setting	Facilitator, coordinator of calibration questions, analyst (MG); contributor to calibration questions (AC)
6	Development of eruption forecasting tool	Individual probability estimates in workshop-style setting	Project leader, workshop facilitator, analyst (AC)
7	National-level long-term eruption forecasts	Classical Model in workshop-style setting	Control expert (AC)

are three main components: establishing the context is the starting point as in the ISO framework; however, risk assessment is replaced with structured expert elicitation and the risk treatment with the completion of the structured expert judgement. We describe the three different components and their building blocks below.

15.3.1 Establishing the Context

Establishing the context includes four main components: (1) identifying the stakeholders and their roles, (2) defining the target question(s), (3) selecting appropriate tools and (4) selecting the experts.

Stakeholders can include a wide range of people, who may or may not be involved directly with the elicitation. For geological hazards in New Zealand, the public are also stakeholders and are usually informed about the outcome. There are several roles within an SEJ project; the problem owner, the coordinator, the facilitator and the analyst (e.g. Hemming et al. 2017). The problem owner is often the person who initiated the elicitation, or, who has been delegated the task of being responsible for the SEJ. The coordinator manages the process, including time lines and collection of responses. The facilitator handles the interactions between experts and needs to be diplomatic, and in our case, able to facilitate group processes with a wide range of different personalities. The facilitator needs to be aware of biases and how to mitigate them. The role requires a good understanding of the problem to be addressed and neutrality with respect to the outcome. The analyst is responsible for processing and analysing the responses and providing feedback to the experts. Applying the Classical Model further requires someone to coordinate the calibration questions. Depending on the scope of the project, the roles can be undertaken by one person, if no conflict of interest exists, or shared by many.

The target question(s) need(s) to be unambiguous, clear and well defined. For example, when asking for the probability of a large earthquake in central New Zealand, it is important to define the magnitude threshold, the region and the timeframe. Experts might want to know whether the earthquake has to be nucleating within the defined region or whether an earthquake that occurs at the boundary of the region and only partially within the region is seen as occurring within the region. It is helpful to write down the target question(s) early in the process and get feedback from various stakeholders whether the problem is appropriately addressed by the target question(s). We find that in discussions with experts during the elicitation workshop that the target question(s) may be slightly modified for clarity.

Appropriate tools include any material, methods or models that can help the experts evaluate the problem. For the risk assessment in carbon, capture and storage (Sect. 15.4.1), the tool was a Bayesian network model. For the time-dependent seismic hazard model for the recovery of Christchurch (Sect. 15.4.2), the tool was the hazard model, the various earthquake forecast models and the ground motion prediction equations. Appropriate tools can include all the background information that can be useful for the experts to make their assessment. It may take some time to prepare the material for the elicitation process and to decide on the most appropriate method of presenting the material.

Selecting appropriate experts is a key component of any SEJ. Good judgement does not depend only on substantive expertise, i.e. knowledge of the domain in question but also on the ability to adapt one's knowledge to novel events, and the ability to communicate one's knowledge and the limitations of one's knowledge in terms of quantiles and probabilities (Hemming et al. 2018). Traditionally, an expert has been defined by qualification, track record and experiences. More experienced experts have been expected to give better advice (Burgman et al. 2011). However, expert status defined by the citations (Cooke et al. 2008) or ranking on an 11-point scale (0 = 'no expertise', 5 = 'moderate expertise', 10 = 'highly expert') in the areas of training, professional experience and current role (Burgman et al. 2011) are a poor guide to actual performance. For geological hazards, we usually select a combination of experts with local and site-specific knowledge and general experts with subject-related experience from elsewhere. These are usually the primary drivers to illuminate the problem. In addition we include challengers, who are related domain experts, who can bring a different perspective to addressing the target question(s), and overall questioners, who also have specific sub-discipline knowledge but can look at the overall system and ask big picture questions. The use of students or early career scientists, who start the process without already having an answer and therefore have the ability to take in information from all sources and draw informed conclusions, can also help to minimize bias (Gerstenberger and Christophersen 2016 and references therein). We refer to these experts as "assimilators". For workshop-style sessions with the experts, we find that eight to 15 experts is a good number and allows for the different expert types to be included, as well as for free discussion with a manageable facilitation burden.

15.3.2 Structured Expert Elicitation

A well-facilitated elicitation workshop is central to our protocol. The workshop needs to be well prepared, including considering ethics requirements, preparing the questionnaires for the calibration and target questions and testing their utility by having colleagues and/or other stakeholders, who are not involved in the elicitation workshop, to answer them ahead of time. The preparation also includes logistics, such as travel arrangements, arranging a meeting facility, catering and preparing all necessary material for the workshop. The elicitation workshop itself includes several elements such as training, calibration questions, discussion of the subject matter and of course the elicitation itself. We consider the combining of the assessments to be part of the structured expert elicitation, and again there are several components including the processing of the questionnaires, evaluation of the calibration and communication of the preliminary results for the experts to review and provide feedback on. In the following we describe each component in more detail.

15.3.2.1 Preparing the Elicitation Workshop

Human ethics approval is required for all research conducted about people. For geological hazards the subject is usually the earth and working with experts does not necessarily require an ethics procedure. It is still important to follow ethical principles such as respecting people, minimizing harm to participants and researchers, ensuring informed and voluntary consent to participate in the research, respecting privacy and confidentiality, avoiding conflict of interest and being socially and culturally sensitive. Research that asks experts about their personal experiences and their thoughts will require an ethics procedure to ensure the research does not cause harm to the participants. Procedures for human ethics approval vary in different countries and local practices will need to be followed.

Calibration questions are central to the Classical Model to allow for performance-based weighting of the experts' judgement. Ideally the calibration questions are close in nature to the target questions so that the experts use similar thinking processes and so that performance on the calibration questions is relevant for the target question(s). Calibration questions have been classified into predictions, where answers are not known during the time of the elicitation but will become known during the analysis, and retrodictions, where the answers are known already but not to the expert during the elicitation (Cooke and Goossens 1999). Calibration questions are further distinguished by whether they are from within the domain of the target question or from an adjacent domain. Domain predictions are ideal, followed by domain retrodictions or adjacent predictions; less ideal is adjacent retrodictions (Cooke and Goossens 1999; Quigley et al. 2018).

Finding suitable calibration questions is not an aspect of the Classical Model that is widely discussed in the literature, even though it can be challenging, in particular when the target questions are small probabilities or parameters for models. Recent work proposes some strategies to finding suitable calibration questions (Quigley et al. 2018). Among them are using results from future measurements that are performed before the analysis is complete; unpublished measurements and mining data for relevant but unusual features. Given that we elicit the calibration question in a workshop-style setting, where experts do not have access to the Internet or their computers, we can use published data, as well as data sets that the experts are very familiar with but cannot access and query at the time. Questions about the experts' own datasets can be useful to highlight the overconfidence bias; experts tend to think that they know simple summary statistics much better, particularly from their own data, than they can recall. Seeing the results of the calibration questions and realizing that the true values are often outside their confidence ranges, gives experts a whole new appreciation of the limitation of their knowledge and consequently experts tend to increase their uncertainty bounds. It can be useful to work with colleagues of the experts to identify calibration questions under the premise that the questions will be kept confidential until after the elicitation. For our recent applications of SEJ in geological hazards in New Zealand (Sect. 15.4), our target questions were mostly about probabilities, weighting models or conditional probabilities for discrete Bayesian network models (Table 15.2). While we could find calibration questions in the same

Table 15.2 Details on the methods for the recent expert elicitations listed in Table 15.1

Project	Number of experts	Number of calibration questions	Number of target questions	Type of target questions	Workshop duration	Time for experts to review their estimates	Aggregation
1	10	10	335	Conditional probabilities	2 half-days	About 1 month	Classical model
2a	12	14	14	Weights of models	3 days	2 weeks	Classical model
2b	5	11	12	Weights of models	1 day	Only on the day	Classical model
3	11	None	1	Probability	2 h	2 days	Average weights
4	14	16	4	Probabilities	2 days	Extra time available but not taken	Classical model
5a	15	17	84	Weights of models	1 day	Only on the day	Classical model
5b	10	16	77	Weights of models	1 day	Only on the day	Classical Model
6	11	None		Conditional probabilities	2 half-days	1 week	Average weights
7	28	24	133	Probabilities, time to eruption, durations of next eruption, vent location	1 day	A couple of months	Classical model

domain, they were not of the same nature as the target questions. In such cases, there is always an element of doubt about whether the expert performance on the calibration questions is valid for the target questions. We aim to include more calibration questions than the recommended number of eight to ten for eliciting three quantiles (Cooke 1991) to be able to test the sensitivity of the performance weights to including different calibration questions. However, the number of calibration questions needs to be balanced with the time required for the experts to answer them and the mental energy required that takes the focus away from the target question(s).

Preparing the elicitation workshop also includes preparing questionnaires for both the calibration and target questions. One aspect of this is the wording of the questions to remove any ambiguities and make them as clear as possible. Another aspect is what medium to use. Paper and pencil seem to work best in workshop-style settings, so that experts can scribble notes at the sides. If the target questions are many conditional probabilities such as in examples 15.4.1 and 15.4.2, it is useful to collect the data in an electronic format, such as a spreadsheet or an online questionnaire. Having the answers electronically circumvents tedious data entry and possible challenges in deciphering handwritten notes. We aim to make the process as convenient as possible for the experts and sometimes offer different options for providing the answers. If we use the Classical Model to aggregate the expert judgement, we include some basic information about the method on the questionnaire. The information contains a figure of a Gaussian distribution function with the percentiles that we elicit marked and the expected location of the answers to the seed questions with respect to the percentiles.

The logistics of the workshop depend on where the workshop is held and include organizing travel, a suitable venue, and catering, including meeting special dietary requirements, to ensure that the experts feel well taken care of and can concentrate on the elicitation exercise. It is also important to plan breaks and opportunities to refresh, to avoid fatigue and allow System 2 of the brain to be engaged.

15.3.2.2 Elicitation Workshop

The elicitation workshop has four important components: training, administrating the calibration questions, discussion of the subject matter, and the elicitation.

The training aims to make experts aware of biases and encourages them to question their knowledge and to facilitate thorough estimates of uncertainty. It includes an introduction to the Classical Model to explain the method and illustrate the question format. It is useful to discuss one or more calibration questions in detail to demonstrate how to think about the percentiles that are elicited. It is good practice to encourage experts to think about the extreme values first to counterbalance the anchoring bias. Administering the calibration questions within the first part of the workshop, following the initial introduction and training, allows for an analyst to process the results and to show them to the experts during the workshop. We find that, despite training, experts are overconfident in their knowledge. Once they have gone through an hour of answering calibration questions and have been presented with the results, they appreciate their overconfidence and tend to widen their confidence intervals. We have not yet mixed further calibration questions in with the target question to formally test this general observation. We are aware that this might influence and change the way the experts answer the target questions compared to the calibration questions. Therefore, the calibration questions become less relevant for performance weighting but are more important as a training tool for estimating uncertainties, as further discussed in Sect. 15.5.1.

Showing the results of the calibration question during the workshop demonstrates that the combined results (see Sect. 15.3.2.3) usually find the correct answer for the calibration questions, despite most individual experts being overconfident in their knowledge. This observation builds the experts' confidence in the method. Giving experts immediate and definitive experience in answering challenging and complex questions of similar type to the target questions also builds the confidence in their own capability. We find that it is not uncommon for some experts to initially think they are unable to estimate any useful answer (despite their inherent overconfidence in any single answer). The feedback on the calibration questions tends to alley these initial concerns.

While most experts seem to enjoy the learning opportunity provided in the way we administer and discuss calibration questions, some experts feel a bit apprehensive and put on the spot, similar to taking an undesired examination. This apprehension seems to be of particular concern when the facilitator and/or analysts are close colleagues. We aim to process the questionnaires so that individual answers, also for the target question, are not even known to the analyst and to ensure confidentiality of individual estimates. This cannot be fully guaranteed because sometimes handwritten or illogical responses need to be confirmed with the experts.

Most of the time during the workshop is spent on discussing the subject matter. This usually includes presentations by domain experts with plenty of time for questions and discussion. The presentation of material requires careful facilitation to avoid anchoring. We encourage experts to think broadly and to consider what might be missing from the presented material and how they can account for unknowns in their uncertainties.

The elicitation of the target questions begins during the workshop. We generally hand out the questionnaires with the target questions before discussions on the subject matter starts, so that experts have the target questions in front of them and can take relevant notes during the discussion. Experts fill in their questionnaire individually, usually within the room. If the target questions fall into different topic areas, we discuss the particular topic area and ensure everyone has the same understanding of the questions being asked, and then allow time for experts to fill in their estimates without interruption. If, during the discussion there are any dominant views, the facilitators try to challenge them by making counterarguments so that experts do not fall for dominance bias. The facilitators encourage experts throughout to answer question to the best of their own knowledge and understanding, and to consider the limits of the knowledge and how best to reflect that in their uncertainty estimates.

Experts usually have extra time beyond the completion of the workshop to review and finalize their answers to the target questions, as indicated in Table 15.2 for our different example applications.

15.3.2.3 Combining the Assessments

Combining the assessments includes a more thorough evaluation of the calibration questions than during the elicitation workshop, processing of the questionnaires, calculating the preliminary results and communicating these to the experts for review.

For the evaluation of the calibration questions, according to the Classical Model, there is software called Excalibur (Cooke and Solomantine 1992), which is freely available (Lighttwist Software 2008) and runs on a Windows operating system. Notes on expert elicitation with Excalibur and a tutorial are also available online (Aspinall 2008; Colson 2016).

For the initial analysis of the calibration questions we usually use global weights without optimization (see Sect. 15.2.3). There are two parameters that can modify the decision-maker and the weights between experts; these are the calibration power and the intrinsic range. The calibration power allows us to compare the calibration of experts between studies with different numbers of calibration questions, and is defined as the ratio of the number of calibration questions used in two different studies, see Chap. 10, this volume. It can vary between 0.1 and 1, with 1 for the studies having the same number of calibration question and 0.1 for one having ten times as many questions as the other. A calibration power of 0.5 reduces the resolution of the significance test to that of one with half as many questions. In practice, reducing the calibration power distributes the weights more equally between experts. The recommendation is to only use a power of less than 1 if all experts have calibration scores less than 0.05 and to avoid giving all the weight to one badly calibrated expert, see Chap. 10, this volume. We often reduce the calibration power because we want to equalize the weights between experts. In our first application (Sect. 15.4.2), one expert got nearly all the weight. However, there was doubt about the calibration questions perfectly representing the target questions. We have reduced the calibration power in subsequent applications with similar motivation, while ensuring not to reduce the overall performance of the decision-maker.

The intrinsic range defines by how much the support of the variable is extended beyond the minimum and maximum percentile of any experts (see Sect. 15.2.3). In Excalibur this value can vary from 0.01 to 100, where 0.10 is the default and corresponds to 10% extension of the overall range on either side. The intrinsic range is important for determining informativeness. A larger support will result in higher informativeness of experts whose quantiles are more widely spread.

The processing of the questionnaire depends on the extent of the target questions; if only a small number of variables are elicited this can be fast and straight-forward. If model parameters or model weights are elicited this might involve lengthy calculations. For large numbers of target questions having the experts fill in their answers in some electronic form can help to reduce the burden of data processing.

We always communicate the initial results to the experts for them to provide feedback on the outcome (Fig. 15.2). This is particularly important when eliciting model parameters and/or model weights. The overall result can be surprising and counter-intuitive. We want to hear experts' thoughts on the overall results. There may be a possibility that experts use this opportunity to sway results in a way they would like to see them go (motivational bias). However, in our applications, we have not observed any evidence for this.

15.3.3 Completion of the Expert Judgement

Completion of the expert judgement involves finalizing the analysis and communicating the results. The final results take experts' feedback into account. For geological hazards, it is important to communicate the results to a wide range of stakeholders, including the public. GNS Science has a social science team that conducts research into how messages are best understood and communicated to reach the relevant stakeholders.

15.4 Application of SEJ for Natural Hazards in New Zealand

Here we introduce seven recent projects that include elements of expert judgement. Table 15.1 provides an overview of the projects, the methods used and our roles. Two of the projects do not strictly fit within the umbrella of geological hazards in New Zealand. Project 1 is about a Bayesian network model for the detection of injected CO_2 in a saline aquifer and sums up the development of our risk assessment method for carbon, capture and storage that led us to introduce structured expert judgement and Bayesian network modelling to geological hazards projects. Project 5 is about the recent update of the Australian national seismic model, which is exemplary for involving the wider research community in seismic hazard assessment.

15.4.1 Risk Assessment in Carbon, Capture and Storage

The Otway Stage 2C project of the CO2CRC involved a test injection of 15,000 tons of supercritical gas mixture at the CO2CRC Otway site in the Australian state of Victoria. The objective was to examine the limits of detecting the gas plume with seismic surveying on the surface and to conduct detailed pressure monitoring of the injection (Pevzner et al. 2015). The risk register for the Otway injection site identified the risk of not being able to detect the injection plume with seismic surveying and not being able to demonstrate stabilization of the plume. We had the opportunity to apply the risk assessment method that we had developed during our CO2CRC involvement, in particular Bayesian networks and structured expert elicitation with the Classical Model, to address these risks. The development of the Bayesian network model structure was an informal and iterative process through remote interaction between GNS staff and CO2CRC. The conditional probabilities for the Bayesian network were elicited in a workshop over two half-days in March 2013, in an SEJ process including the application of the Classical Model. We had the opportunity to investigate possible calibration questions ahead of time (Christophersen et al. 2011). Since we administered the calibration questions during the face-to-face workshop, we could ask questions from the published and grey literature as well as about specific data from the Otway basin. Asking the experts about their own data was particularly useful to understand overconfidence. Experts were critical about the calibration questions during the workshop. One expert questioned the quality of the work chosen from the grey literature and was encouraged to consider that in the uncertainty estimate. The critique allowed for a solid discussion on the purpose of the calibration questions.

The result of the Bayesian network was a 74% probability of detecting the plume, and a 57% probability that there will be consistency between the model-predicted plume behaviour and the observations. The plume detection has been successful (Pevzner et al. 2017).

15.4.2 A Time-Dependent Seismic Hazard Model for the Recovery of Christchurch

The New Zealand National Seismic Hazard model (NSHM; Stirling et al. 2012) estimates earthquake ground shaking and forms the basis for structural design in New Zealand. The NSHM applies the well-established practice of probabilistic seismic hazard analysis, which has three key components: the fault source model, the distributed source model and ground motion prediction equations. The NSHM is regularly updated to include the latest science.

The Canterbury earthquake sequence increased the rate of seismicity in the Canterbury region well above the long-term rates and the seismicity is expected to stay elevated for years, if not decades (Gerstenberger et al. 2014, 2016). The elevated seismicity warranted the development of a new time-varying seismic hazard model for the Canterbury region because the NSHM was expected to underestimate the seismic hazard due to ongoing aftershocks and the possibility of further triggered earthquakes. The new seismic hazard model has the same components as the NSHM: a fault source model, a distributed source model and ground motion prediction equations. The fault model was extended from the 2010 NSHM update but was not subject to SEJ. The distributed source model is the dominant contributor in this case and is a combination of earthquake-clustering models of three timescales (short-term, medium-term and long-term). Weights for the models were elicited in a two-day workshop including the application of the Classical Model.

The ground motion prediction equation component of the model was extended to include a new Christchurch-specific model (Bradley 2010, 2013). A one-day workshop was held to elicit the necessary parameters and weights for the ground motion prediction equations, again including the application of the Classical Model.

The resulting hazard model represents the seismic hazard for the Canterbury region for the next 50 years. The model has been used to provide earthquake probabilities to a range of end users on timescales from 1 day to 50 years. The 50-year hazard forecast has informed the revision of the New Zealand building design guidelines and other aspects of the rebuilt of Christchurch.

15.4.3 *Informal Elicitation of the Probability of Large Earthquakes in Central New Zealand Impacted by Slow Slip and the Kaikōura Earthquake*

The 14 November 2016 Kaikōura earthquake with magnitude M = 7.8 triggered wide-spread silent and slow movement along the plate boundary, also called slow slip events (SSE); these events can take weeks to months to occur but are not felt by people. By 25 November, observations from global positioning system (GPS) stations indicated that three regional SSE were occurring. While SSE in these regions have been observed numerous times in the past 20 years, they had never occurred simultaneously before and one of them appeared to have a larger slip rate than previously observed. These observations raised concerns about the impact of the SSE on future earthquake occurrence. On 25 November, the New Zealand Ministry of Civil Defense and Emergency Management (MCDEM) was briefed about the concerns, and consequently expected formal advice from GNS Science on the likelihood of future M $\geq$ 7.8 events in central New Zealand, including any potential impact of the ongoing SSE on this likelihood.

While GNS Science has provided earthquake forecasts in response to large earthquakes since the September 2010 M = 7.1 Darfield earthquake (Christophersen et al. 2017), no earthquake forecasting model implicitly considers SSE. To fulfil MCDEM's expectation, GNS Science used expert elicitation. We had about a week to pull together different strands of evidence including the forecasts from the statistical model, results from synthetic earthquake data (Robinson et al. 2011) and the NSHM (Stirling et al. 2012). We analysed the effect of SSE on seismicity, calculated Coulomb stresses and consulted with international experts (Gerstenberger et al. 2017). It was not possible to develop subject-appropriate calibration questions within that short time period and with an active response to the mainshock still ongoing. On 1 December 2016, we held a two-hour workshop with 11 New Zealand experts, who were mostly GNS Science staff. We presented and discussed all information available at that time. Experts then individually estimated the probability of an M $\geq$ 7.8 events in central New Zealand within the next year. Everyone provided their best estimate and a 90% confidence interval. The results were communicated to MCDEM and to the public via the GeoNet website.

15.4.4 *SEJ and the Classical Model to Assess the Probability of Large Earthquakes in Central New Zealand Impacted by Slow Slip*

In the year following the Kaikōura earthquake, GNS Science conducted further research on the effect of SSE on earthquakes (Kaneko et al. 2018; Wallace et al. 2017) and continued to consult with international colleagues two workshops were held, including an initial one at the annual meeting of the Southern California Earthquake Center, in California to discuss initial model developments. Subsequently, we conducted a second SEJ on the one-year anniversary of the Kaikōura earthquake to estimate the probability of large earthquakes in central New Zealand within the subsequent one and ten years. The second elicitation workshop was held over two days at GNS Science and was attended by fourteen experts from four different countries and nine different organizations. We applied the Classical Model with calibration questions that were again derived from the published literature and relevant publicly available data sets that the experts could not access during the workshop.

The most striking observation when comparing the results of the expert elicitation in December 2016 and November 2017 is an increase of the uncertainty estimates in 2017, even though the 2016 estimates were 90% confidence intervals versus 80% in 2017. Although a direct comparison is difficult, this observation is consistent with our expectation that through training and a much more thorough process the experts increase their uncertainty once they have seen the results from the calibration question. It also seems that the experts' answers were more anchored on the results from the statistical model in the 2016 December when experts had not gone through the SEJ process.

15.4.5 Australian National Seismic Hazard Model

Geoscience Australia is an agency of the Australian government and is responsible for the Australian national seismic hazard model. In the 2018 update of the model, NSHA18, Geoscience Australia undertook a new, and so far unique for seismic hazard, approach: it invited the Australian earthquake hazard community to submit peer-reviewed seismic source and ground motion models for consideration (Allen et al. 2018; Griffin et al. 2018). This resulted in 16 seismic source models and 20 ground motion models being proposed and contributing to NSHA18, demonstrating the range of expert opinions on characterizing seismic hazard for a low seismicity region like Australia. Following similar methods as described above for the Canterbury hazard model, Geoscience Australia held two expert elicitation workshops in March 2017 to weight different seismic source models and ground motion models. The workshop applied the Classical Model and GNS Science assisted with the calibration questions and workshop facilitation. The 17 workshop participants represented the collective expertise of the Australian earthquake hazard community. Feedback from the workshop participants was positive, with experts reporting being challenged by, but enjoying, the calibration and elicitation process.

The NSHA18 yields much lower hazard estimates than previous assessments (Allen et al. 2018). This is due to a number of factors, including the revision of earthquake magnitudes and the use of more modern ground motion models than previously available. Given tight timelines, there was no chance for the experts to review their contribution once the hazard was calculated. For future studies, Geoscience Australia recommends to re-engage with the experts to allow them to review and reassess their choices, despite concerns that experts may be motivated to tweak answers to move results closer to their expectation (Allen et al. 2018). Such a review process would be consistent with our protocol (Fig. 15.2).

15.4.6 Development of an Eruption Forecasting Tool

Volcanic eruptions are usually preceded by a period of unrest, during which small earthquakes occur around the volcano; the volcano can emit increasing amounts of gas, and ground deformation may be observed. GeoNet coordinates the volcano monitoring team that consists of GNS staff based at three sites. The team meets regularly (partly remotely) to review the status of all 12 monitored New Zealand volcanic centres. It sets the Volcano Alert Levels (Potter et al. 2014) and the Colour Codes of the International Civil Aviation Organization and regularly estimates the probability of forthcoming eruptions for internal health and safety policy requirements (Deligne et al. 2018; Jolly et al. 2014). In recent years, there have been small volcanic eruptions, including the fatal December 2019 Whakaari/White Island eruption. New Zealand has the potential for much more disruptive volcanic eruption.

There are limited quantitative tools in eruption forecasting (Sparks et al. 2012) that can help the volcano monitoring team to assess the probability of upcoming eruptions. Given the success of Bayesian networks in the CO2CRC-project to model complex problems, we proposed to trial Bayesian networks as decision-support tool in volcano monitoring (Christophersen et al. 2018). We started with a small team with wide-spread expertise. In an informal process, the team adapted a published Bayesian network model for eruption forecasting (Hincks et al. 2014), which was reviewed by some members of the volcano monitoring team. In a structured process, we elicited the conditional probabilities for the Bayesian network in a workshop over two half-days in early December 2015. The workshop included a presentation on the Classical Model and some example calibration questions to introduce the method. We did not have the time and resources to develop appropriate calibration question for the conditional probabilities of the Bayesian network. Given the previous experience with experts' unease about the calibration questions, we decided against using the Classical Model so as to not distract from the main objective of exploring the potential use of Bayesian networks in volcano monitoring and eruption forecasting. In feedback questionnaires, the experts were supportive of applying the Classical Model in future elicitations. The finding of the project was that Bayesian networks are promising tools for volcano monitoring with many recommendations for future work, mainly focussing on developing Bayesian networks with continuous variables and exploring dynamic Bayesian networks but also including SEJ for parameterising the model.

15.4.7 National-Level Long-Term Eruption Forecasts

Volcanoes cause many different hazards, including ash fall, pyroclastic density flows, lava flow and lahars. These hazards can impact near and far from the volcano, before, during and after an eruption (National Academies of Sciences 2017). Many volcanic hazards depend on the weather conditions like wind direction and rain, the presence of snow and ice, and the local topography (Stirling et al. 2017). Thus, the development of a comprehensive volcano hazard model is a complex task. The first step involves quantifying the frequency, size and location of eruptions for each volcano. A recent project led by Massey University with broad collaboration across other New Zealand organizations including GNS Science, conducted an SEJ to estimate the timing and sizes of the next eruption for 12 volcanoes (Bebbington et al. 2018). A total of 28 experts including volcanologists, statisticians, and hazards scientists, provided estimates that were combined using the Classical Model to arrive at hazard estimates. The same experts contributed to an informal expert elicitation to outline the next steps for developing a national probabilistic volcanic hazard model for New Zealand (Stirling et al. 2017). Given the wealth of material to elicit, the discussion during the workshop was kept relatively short. There was ample opportunity for experts to revise their answers. The results and challenges of the study have been well documented (Bebbington et al. 2018).

15.5 Discussion and Conclusion

There are many applications for expert judgement in geological hazards in New Zealand. We have introduced seven recent applications that we have been involved within different roles. The problems are often complex and require input from multiple sub-disciplines. Being aware of the human brain's preference to take shortcuts, potentially causing biases, we are interested in robust expert elicitation protocols that minimize biases and quantify uncertainty. We have introduced a protocol for expert judgement that is based on risk assessment methods and has workshop-style interaction at its heart, so that experts can share all evidence and reach a good understanding of the problem. The Classical Model is well suited to explore the uncertainty around the complex issues that we are addressing. Here we discuss how our application of the Classical Model differs from the standard recommendations, some of the challenges we have encountered, and the benefits of our protocol.

15.5.1 Tweaks in Applying the Classical Model

The way that we apply the Classical Model differs in two ways from the standard recommendations (e.g. Cooke 1991; Quigley et al. 2018). Firstly, we clearly set apart the calibration questions from the target questions and use them as a training tool to improve uncertainty estimates for the target questions. As a consequence, experts may potentially have a different philosophy when assessing confidence bounds for target and calibration questions. Ideally, experts become better in assessing uncertainty, which may mean they are more likely to increase their uncertainty bounds. Alternatively, because the calibration exercise is separate, it allows experts to reduce their uncertainty bounds on the target questions to obtain desired results. Thus, if expert behaviour is inconsistent between the two question sets, our approach may reduce the value of the calibration questions for absolute performance weighting. However, in our opinion the potential for improved quantification of uncertainty through training outweighs the potential for inconsistent expert behaviour between calibration and target questions.

Secondly, we reduce the power of the calibration when aggregating the expert judgements. Using full power when determining the weights assumes the list of calibration questions is exhaustive and fairly represents the knowledge required for the elicitation; we do not feel this is a reasonable expectation and allow for probable inadequacy of the selected calibration questions by reducing the power. The effect of reducing the power is to distribute the weights more evenly across the experts. Reducing the power can be carefully balanced to not significantly reduce the overall performance of the decision-maker.

15.5.2 Challenges

Over the past few years, we have moved from relatively informal elicitation processes to a more structured protocol. We have encountered various challenges along the way, often associated with stakeholders' unfamiliarity with SEJ. For example, several times stakeholders welcomed the use of the Classical Model because they thought that the performance-based weighting would allow them to know who to ask in the future. We explained that the weights are specific to a particular calibration exercise with questions designed for the target questions, and that the discussion in the workshop-style sessions is essential for the experts to gain a comprehensive understanding of the problem and arrive at their estimates. Also, following ethics protocols, the weights are anonymous and not shared.

Another challenge has been the small size of the team at GNS Science that is involved in SEJ; for some applications, team members were also domain experts. It can be difficult to keep role separation and avoid real or perceived conflicts of interest.

Sections 15.3.2.1 and 15.3.2.2 include a description of the challenges in developing calibrations questions. In particular, when the target questions are weights and probabilities there may be extra difficulty in finding suitable calibration questions. These challenges can be overcome with experience and allowance for extra time, both for the project team and the experts.

15.5.3 Benefits of Our Protocol

Our protocol has solid foundations, being built on the principles of risk management (ISO 2009) and using the Classical Model for combining expert judgement (Cooke 1991). It aims to provide training for experts to recognize their own biases and limitations in their knowledge. The discussions during the workshop ensure the representation of varied opinions and experiences. One key aspect of the protocol is that the uncertainty estimates can be propagated through to the results (Gerstenberger and Christophersen 2016).

Given that it is challenging to quantify the success of any protocol, in particular when estimating low-probability events, we measure the success of our protocol by its acceptance by the stakeholder, its ability to produce results in a timely manner and its solid foundations. We find that experts enjoy the experiences, in particular the thorough discussions of the subject matter in the workshop-style setting. Therefore, they contribute their thoughts and understanding of the problem, which leads to the development of new knowledge and advanced understanding.

15.5.4 Outlook

Developing measures for evaluating different SEJ methods continues to be an important topic for further research. Given that our focus is on applications of SEJ, we do not have much of an opportunity to conduct methodological research into SEJ. Hence it is important for us to stay involved with the international community on SEJ to have the opportunity to present and discuss our work with the experts in the field. As a consequence of these interactions, and further experiences in future applications, our protocol will continue to evolve.

Acknowledgments We would like to acknowledgement Roger Cooke and Willy Aspinall for their support and encouragement in applying the Classical Model. We thank all of the experts who took part in different projects. David Rhoades and Anca Hanea reviewed the manuscript before submission. Annemarie Christophersen thanks Anca Hanea for many lively and educational discussions, and acknowledges a travel grant by the Royal Society of New Zealand to attend the workshop on "Science, uncertainty and decision-making in the mitigation of natural risk" of the European Union funded COST Action "Expert Judgment Network: Bridging the Gap Between Scientific Uncertainty and Evidence-Based Decision Making".

References

Allen, T. I., Griffin, J., Leonard. M., Clark, D., & Ghasemi, H. (2018) The 2018 National Seismic Hazard Assessment for Australia: model overview. *Geoscience Australia, Canberra, Australia.* https://doi.org/10.11636/Record.2018.027.

Aspinall, W. (2008) Expert judgment elicitation using the classical model and EXCALIBUR. Retrieved 8 June 2018, from http://dutiosc.twi.tudelft.nl/~risk/extrafiles/EJcourse/Sheets/Aspinall%20Briefing%20Notes.pdf.

Bang, D., & Frith, C. D. (2017). Making better decisions in groups Royal Society Open Science 4 https://doi.org/10.1098/rsos.170193.

Bebbington, M. S., Stirling, M. W., Cronin, S., Wang, T., & Jolly, G. (2018). National-level long-term eruption forecasts by expert elicitation. *Bulletin of Volcanology, 80,* 56. https://doi.org/10.1007/s00445-018-1230-4.

Bradley, B. A. (2010). NZ-specific pseudo-spectral acceleration ground motion prediction equations based on foreign models. University of Canterbury.

Bradley, B. A. (2013). A New Zealand-specific Pseudospectral acceleration Ground-Motion prediction equation for active shallow crustal earthquakes based on foreign models. *Bulletin of the Seismological Society of America, 103,* 1801–1822. https://doi.org/10.1785/0120120021.

Burgman, M. A., et al. (2011). *Expert Status and Performance PLOS ONE, 6,* e22998. https://doi.org/10.1371/journal.pone.0022998.

Christophersen, A., Deligne, N. I., Hanea, A. M., Chardot, L., Fournier, N., & Aspinall, W. P. (2018). Bayesian Network modeling and expert elicitation for probabilistic eruption forecasting: pilot study for Whakaari/White Island. *New Zealand Frontiers in Earth Science, 6,* 23.

Christophersen, A., Gerstenberger, M., & Nicol, A. (2011) The feasibility of using seed questions for weighting expert opinion in CCS risk assessment. CO2CRC.

Christophersen, A., Rhoades, D. A., Gerstenberger, M. C., Bannister, S., Becker, J., Potter, S. H., & McBride, S. (2017). Progress and challenges in operational earthquake forecasting in New Zealand. *Paper presented at the New Zealand Society for Earthquake Engineering Technical Conference, Michael Fowler Centre, Wellington*, 27–29 April 2017.

CO2CRC. (2011). CRC for Greenhouse Gas Technology. http://www.co2crc.com.au/ (2018).

Colson, A. R. (2016). Excalibur tutorial. Retrieved 8 June 2018, from https://www.expertsinuncertainty.net/Portals/60/ESR%20Warsaw/Excalibur%20tutorial.pdf?ver=2017-11-27-124915-117.

Cooke, R. M. (1991). *Experts in uncertainty: Opinion and subjective probability in science.* USA: Oxford University Press.

Cooke, R. M., ElSaadany, S., & Huang, X. (2008). On the performance of social network and likelihood-based expert weighting schemes. *Reliability Engineering & System Safety, 93,* 745–756.

Cooke, R. M., & Goossens, L. H. J. (1999). Procedures guide for structured expert judgment. Luxemburg.

Cooke, R. M., & Goossens, L. L. H. J. (2008). *TU Delft Expert Judgment Data Base Reliability Engineering & System Safety, 93,* 657–674.

Cooke, R. M., & Solomantine, D. (1992). EXCALIBUR—integrated system for processing expert judgements. Delft, The Netherlands.

Deligne, N. I., Jolly, G. E., & Taig, T. (2018). Evaluating life-safety risk for field work on active volcanoes: VoLiST, a volcano observatory's decision-support tool Journal of Applied Volcanology in review.

Gerstenberger, M., Christophersen, A., Buxton, R., Allinson, G., Hou, W., Leamon, G., et al. (2012). *Integrated risk assessment for CCS Energy Procedia, 37,* 2775–2782.

Gerstenberger, M. C., & Christophersen, A. (2016). A Bayesian network and structured expert elicitation for Otway Stage 2C: Detection of injected CO2 in a saline aquifer *International Journal of Greenhouse Gas Control*, *51*, 317–329. https://doi.org/10.1016/j.ijggc.2016.05.011.

Gerstenberger, M. C., Christophersen, A., Buxton, R., & Nicol, A. (2015). Bi-directional risk assessment in carbon capture and storage with Bayesian Networks. *International Journal of Greenhouse Gas Control, 35,* 150–159. https://doi.org/10.1016/j.ijggc.2015.01.010.

Gerstenberger, M. C., Kaneko, Y., Fry, B., Wallace, L., Rhoades, D., Christophersen, A., & Williams, C. (2017). Probabilities of earthquakes in central New Zealand. Lower Hutt (NZ). https://doi.org/10.21420/g2fp7p.

Gerstenberger, M. C., McVerry, G. H., Rhoades, D. A., & Stirling, M. (2014). Seismic hazard modelling for the recovery of Christchurch. *New Zealand Earthquake Spectra, 30,* 17–29. https://doi.org/10.1193/021913EQS037M.

Gerstenberger, M. C., Rhoades, D. A., & McVerry, G. H. (2016). A Hybrid time-dependent probabilistic seismic-hazard model for canterbury. *New Zealand Seismological Research Letters, 87,* 1311–1318. https://doi.org/10.1785/0220160084.

Griffin, J., et al. (2018). *Expert elicitation of model parameters for the 2018 National Seismic Hazard Assessment*. Canberra, Australia: Geoscience Australia.

Hanea, A. M., McBride, M. F., Burgman, M. A., & Wintle, B. C. (2018). The value of performance weights and discussion in aggregated expert judgments. *Risk Analysis, 38,* 1781–1794. https://doi.org/10.1111/risa.12992.

Hemming, V., Burgman, M. A., Hanea, A. M., McBride, M. F., & Wintle, B. C. (2017). *A practical guide to structured expert elicitation using the IDEA protocol Methods in Ecology and Evolution, 9,* 169–180. https://doi.org/10.1111/2041-210X.12857.

Hemming, V., Walshe, T. V., Hanea, A. M., Fidler, F., & Burgman, M. A. (2018). Eliciting improved quantitative judgements using the IDEA protocol: A case study in natural resource management PLoS ONE, *13* https://doi.org/10.1371/journal.pone.0198468.

Hincks, T. K., Komorowski, J. C., Sparks, S. R., & Aspinall, W. P. (2014) Retrospective analysis of uncertain eruption precursors at La Soufrière volcano, Guadeloupe, 1975–77: volcanic hazard assessment using a Bayesian Belief Network approach. *Journal of Applied Volcanology*, *3*, 1–26.

ISO. (2009). *ISO 31000:2009 Risk management—Principles and guidelines*. Geneva, Switzerland: International Organization for Standardization.

Jenkins, C. R. et al. (2012). Safe storage and effective monitoring of CO2 in depleted gas fields Proceedings of the National Academy of Sciences, *109*, E35-E41 https://doi.org/10.1073/pnas.1107255108.

Jolly, G. E., Keys, H. J. R., Procter, J. N., & Deligne, N. I. (2014). Overview of the co-ordinated risk-based approach to science and management response and recovery for the 2012 eruptions of Tongariro volcano. *New Zealand Journal of Volcanology and Geothermal Research*, *286*, 184–207 https://doi.org/10.1016/j.jvolgeores.2014.08.028.

Kahneman, D. (2011). *Thinking, fast and slow*. Straus and Giroux, New York: Farrar.

Kaneko, Y., Wallace Laura, M., Hamling Ian, J., & Gerstenberger Matthew, C. (2018). Simple physical model for the probability of a subduction-zone earthquake following slow slip events and earthquakes: application to the Hikurangi Megathrust. *New Zealand Geophysical Research Letters, 45,* 3932–3941. https://doi.org/10.1029/2018GL077641.

Kerr, N. L., & Tindale, R. S. (2011). Group-based forecasting?: A social psychological analysis. *International Journal of Forecasting, 27*, 14–40 https://doi.org/10.1016/j.ijforecast.2010.02.001.

Kunda, Z. (1990). *The Case for Motivated Reason Psychological Bulletin, 108,* 480–498. https://doi.org/10.1037/0033-2909.108.3.480.

Lighttwist Software. (2008). Excalibur. Retrieved 8 June 2018, from http://www.lighttwist.net/wp/excalibur.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology, 67,* 371–378.

Milgram, S. (1974). *Obedience to authority: An experimental view*. Taylor & Francis.

Montibeller, G., & von Winterfeldt, D. (2015). Cognitive and motivational biases in decision and risk analysis risk. *Analysis, 35,* 1230–1251. https://doi.org/10.1111/risa.12360.

Montibeller, G., & von Winterfeldt, D. (2018). Individual and group biases in value and uncertainty judgments. In L. M. A. Dias, J. Quigley (Eds.) Elicitation, vol 261. International Series in Operations Research & Management Science. Springer, Cham. https://doi.org/10.1007/978-3-319-65052-4_15.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115,* 502–517. https://doi.org/10.1037/0033-295X.115.2.502.

National Academies of Sciences, Engineering, and Medicine. (2017). Volcanic Eruptions and Their Repose, Unrest, Precursors, and Timing. The National Academies Press, Washington, DC. https://doi.org/10.17226/24650.

New Zealand Ministry of Civil Defence and Emergency Management. (2015). The Guide to the National Civil Defence Emergency Management Plan 2015.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2,* 175–220. https://doi.org/10.1037/1089-2680.2.2.175.

Pevzner, R., Caspari, E., Gurevich, B., Dance, T., & Cinar, Y. (2015). Feasibility of CO2 plume detection using 4D seismic: CO2CRC Otway Project case study—Part 2: Detectability analysis GEOPHYSICS 80, B105-B114 https://doi.org/10.1190/geo2014-0460.1.

Pevzner, R. et al. (2017). Stage 2C of the CO2CRC Otway Project: Seismic Monitoring Operations and Preliminary Results Energy Procedia, *114*, 3997–4007 https://doi.org/10.1016/j.egypro.2017.03.1540.

Potter, S. H., Jolly, G. E., Neall, V. E., Johnston, D. M., & Scott, B. J. (2014). Communicating the status of volcanic activity: Revising New Zealand's volcanic alert level system. *Journal of Applied Volcanology*, 3.

Quigley, J., Colson, A., Aspinall, W., Cooke, R. M. (2018) Elicitation in the classical model. In L. C. Dias, A. Morton, J. Quigley (Eds.) *Elicitation: The science and art of structuring judgement*. Springer International Publishing, Cham, pp. 15–36. https://doi.org/10.1007/978-3-319-65052-4_2.

Robinson, R., Van Dissen, R., & Litchfield, N. (2011). Using synthetic seismicity to evaluate seismic hazard in the Wellington region. *New Zealand Geophysical Journal International, 187,* 510–528. https://doi.org/10.1111/j.1365-246X.2011.05161.x.

Sparks, R. S. J., Biggs, J., & Neuberg, J. W. (2012). *Monitoring Volcanoes Science, 335,* 1310–1311. https://doi.org/10.1126/science.1219485.

Stirling, M., et al. (2017). Conceptual development of a national volcanic hazard model for New Zealand frontiers in Earth. *Science, 5,* 51.

Stirling, M., McVerry, G., Gerstenberger, M., Litchfield, N., Van Dissen, R., Berryman, K., Barnes, P., Wallace, L., Villamor, P., Langridge, R., Lamarche, G., Nodder, S., Reyners, M., Bradley, B., Rhoades, D., Smith, W., Nicol, A., Pettinga, J., Clark, K., & Jacobs, K. (2012). National seismic hazard model for New Zealand: 2010 update. *Bulletin of the Seismological Society of America*, *102*, 1514–1542. https://doi.org/10.1785/0120110170.

Trouche, E., Sander, E., & Mercier, H. (2014). Arguments, more than confidence, explain the good performance of reasoning groups. *Journal of Experimental Psychology: General*, *143*, 1958–1971. https://doi.org/10.1037/a0037099.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science, 185,* 1124–1131.

Wallace, L. M. et al. (2017). Large-scale dynamic triggering of shallow slow slip enhanced by overlying sedimentary wedge. *Nature Geoscience*, *10*, 765 https://doi.org/10.1038/ngeo3021.

Westen, D., Blagov, P. S., Harenski, K., Kilts, C., & Hamann, S. (2006). Neural bases of motivated reasoning: An FMRI study of emotional constraints on partisan political judgment in the 2004 U.S. *Presidential election Journal of cognitive neuroscience, 18,* 1947–1958. https://doi.org/10.1162/jocn.2006.18.11.1947.

Chapter 16
Using the Classical Model for Source Attribution of Pathogen-Caused Illnesses

Lessons from Conducting an Ample Structured Expert Judgment Study

Elizabeth Beshearse, Gabriela F. Nane, and Arie H. Havelaar

A recent ample Structured Expert Judgment (SEJ) study quantified the source attribution of 33 distinct pathogens in the United States. The source attribution for five transmission pathways: food, water, animal contact, person-to-person, and environment has been considered. This chapter will detail how SEJ has been applied to answer questions of interest by discussing the process used, strengths identified, and lessons learned from designing a large SEJ study. The focus will be on the undertaken steps that have prepared the expert elicitation.

16.1 Introduction

Source attribution is the process by which illnesses caused by specific pathogens are attributed to sources of infections. Illnesses transmitted by food and water result in a major disease burden worldwide. The World Health Organization (WHO) estimated that in 2010, 31 known hazards resulted in 600 million foodborne illnesses and 420,000 deaths globally (Asratian et al. 1998). A separate study in the United States

E. Beshearse
University of Florida, 2055 Mowry Road, Gainesville, FL 32610, USA

Present Address:
Centers for Disease Control and Prevention, 1600 Clifton Road, Atlanta, GA 30329, USA

G. F. Nane (✉)
Delft University of Technology, van Mourik Broekmanweg 6, 2628XE Delft, The Netherlands
e-mail: g.f.nane@tudelf.nl

A. H. Havelaar
University of Florida, 2055 Mowry Road, Gainesville, FL 32610, USA

A. M. Hanea et al. (eds.), *Expert Judgement in Risk and Decision Analysis*,
International Series in Operations Research & Management Science 293,
https://doi.org/10.1007/978-3-030-46474-5_16

estimated that approximately 9.4 million illnesses, 56,000 hospitalizations, and 1,351 deaths are caused by 31 known pathogens transmitted through food each year. Despite recognition of the high disease burden caused by food- and waterborne illnesses, gaps in data remain a barrier to producing fully data-based source attribution and burden estimates. Methods to produce such estimates have historically used outbreak analysis, epidemiologic studies, and other surveillance-based data.

However, these methods are limited due to scarce and incomplete data. Structured expert judgment (SEJ) methods have been increasingly applied to address the lack of data. SEJ has been applied to estimate the source attribution globally and at the national level in many countries, including Australia, Canada, and the Netherlands. To aid with addressing the ongoing efforts for the prevention and control of foodborne and waterborne diseases in the United States, a national SEJ study using Cooke's Classical Model (Cooke 1991) was undertaken to attribute domestically acquired illnesses to a comprehensive set of transmission pathways representing 100% of transmission for the 33 selected pathogens. These pathways included foodborne, waterborne, person-to-person, animal contact, and environmental transmission.

This chapter will explore how SEJ has be applied to answer the questions of interest, by discussing the process used, strengths identified, and lessons learned from designing a large SEJ study. A procedure guide was followed to ensure completeness and comprehensiveness (Cooke 1991). The guide divides the study into three primary stages: preparation, elicitation, and post-elicitation. This chapter focuses primarily on the preparation and elicitation stages, as the data analysis completed during the post-elicitation is covered in greater detail in Chap. 10, this volume, for example. Furthermore, the expert data analysis and results from this SEJ study will be reported in a separate manuscript.

The first steps in following the procedures guide are the preparation for the study. The importance of adequate time and review for this step of an SEJ study cannot be overstated. Without proper preparation, it is difficult to execute subsequent elicitation. We will further discuss each of the steps in the order and how they were addressed for this study. Nonetheless, these steps do not necessarily have to be followed in the exact stepwise order, as some may need to occur simultaneously.

16.2 Definition of the Case Structure and Questions of Interest

Firstly, exactly what will be elicited must be determined, that is identifying the questions of interest, also referred to as target variables. The questions of interest or target variables should include uncertain quantities for which there are no (easily) available data. For the US SEJ study discussed in this chapter, this means the identification of pathogens that are transmitted via the five main pathways, but for which the proportion of illnesses that occur through each transmission pathway is uncertain. Pathogens, that were known to have greater than 95% transmission through a single pathway, were not elicited (e.g., *Staphylococcus aureus* toxin was considered primarily foodborne transmission). Input from both the problem owner and the research team has been used to guide and make decisions about the variables to include. Consideration has been given to the number of target questions to be elicited, as expert fatigue can occur if too many are assigned to each individual. Some studies address this issue by focusing their questions on target variables that cause the highest burden as opposed to eliciting a more comprehensive selection of causes (Vally et al. 2014). Initially, 33 distinct pathogens were considered for this study. After extensive discussions with the problem owner, it was considered relevant to elicit separate estimations of transmission pathways for multiple clinical manifestations or subtypes for certain pathogens. Hence, the final total number of target questions included in the elicitation was 47.

Next, clear definitions for the items being elicited needed to be established. Several challenges can exist when attempting to write clear definitions. Established definitions that are used broadly may not exist. This may mean that different institutions, and even individual expert participants, might have disparate views on the elicited quantities. For the estimates elicited from the experts to be meaningful, they must all apply the same definitions in the same way for all target questions. Without this, experts will be providing estimates with differing understandings and the combined assessments will produce inaccurate uncertainty quantifications for the variables of interest. To highlight an example of some of these challenges, consider how an individual can become ill from *Salmonella.*

Example: Suppose an individual owns chickens, goes out to feed them, and his hands become contaminated with *Salmonella.* He subsequently comes inside and washes his hands, but the sink has now been contaminated. He prepares his lunch by rinsing lettuce and placing it in the sink. Water droplets containing *Salmonella* bacteria from when he washed his hands contaminate the lettuce. He then eats this lettuce and becomes ill. What would you say is the transmission route for this? Is it animal contact because the bacteria originated from his chickens, or is it foodborne because it is ultimately the lettuce he ate was the vehicle that led to ingestion of the bacteria?

Without clear definitions of exactly what constitutes animal contact transmission and foodborne transmission, expert participants might answer this question differently. Numerous other examples can be given, in which there might not be a clear definition of the pathway transmission. Thinking through challenging examples to test the study definitions can help strengthen the study design and to identify gaps that may not have been considered beforehand.

As surprisingly as it may seem, there are no existing, broadly agreed upon, definitions for attribution of foodborne and waterborne illness transmission. Based on our experience, even within a single institution, different groups use different definitions for the pathway transmissions. Consequently, several months of iteration were needed to achieve clear transmission pathway definitions, that were comprehensive and mutually exclusive for this study. To accomplish this, multiple meetings, discussions, and testing of definitions against difficult scenarios that could cause infection were necessary.

In addition, expert participants received, approximately two months before the elicitation, both a training webinar on pathway definitions followed a quiz with 20 challenging exposure scenarios to verify there was a common understanding of the definitions used in the study. Documents with the definitions were provided to the experts in advance of the webinar. This provides the experts with the opportunity to ask questions and clarify any concerns with the definitions. In addition, the questionnaire responses allow study designers to ensure that a common understanding was met, and if not, address this prior to the elicitation. It can also be helpful to think through and prepare responses as to why definitions were formatted and designed in the way they were. For this study, the definitions were aligned with how they would be applied and used by the stakeholders and this was explained to the expert participants. So, while they might use slightly different definitions in their work, they would be applying the study definitions when providing estimates.

An example of a problem given during the webinar is presented below.

Example Please choose the transmission pathway that best fits each scenario described

1. Norovirus illness among attendees of a banquet linked to carpet and indoor environment that had been contaminated with vomit the day before the banquet and subsequently cleaned

   a. Foodborne transmission
   b. Waterborne transmission
   c. Person-to-person transmission
   d. Animal contact transmission
   e. Environmental transmission

Finally, for some target variables, transmission pathways have been blocked by the research team. The decision has been based on the well-known microbiology and ecology of certain pathogens. If the blocked pathways were not in line with experts' beliefs, they had the opportunity to provide estimates for transmission by one or more blocked pathways, along with providing motivation for their assessments. During the elicitation, some experts used this opportunity.

All in all, careful attention has been given to this step in preparing for the expert elicitation and this focus was a strength of the US SEJ study.

16.3 Identification of Calibration Questions

The next step in the study was the selection of calibration questions, also known as seed variables. Calibration questions are designed to assess an expert's ability to provide valid estimates under uncertainty and are used to weight the responses to the target questions. In order to accomplish this, answers to these questions should be known to the study design team, but not to the expert participants. The 14 calibration questions have been selected within the domain of the target variables, and they can be classified as retrodictions. For an overview of the types of calibration questions by domain, see Chap. 10, this volume. Since the US SEJ study had target questions addressing proportions of transmission by food, water, and other pathways, the 14 calibration questions focused on public health surveillance data for food- and waterborne diseases, frequency of exposure to hazards, and food consumption patterns within the United States. All topics are within the domain of food- and waterborne illness, and which have an impact on disease transmission.

As with target questions, clearly defining the calibration questions, as well as an explicit reference to the data on which the answers are based are critical in designing the calibration questions. As can be seen in the example below, it is important to first provide background about where the data are derived from, so experts provide their responses based on this. Multiple questions can be based on a single data source, so this background could apply to more than one question. Including clearly defined formats for how experts should provide their answers is important as well. If needed, giving an example of the format may be warranted, as seen in the example below. There should be some consideration for how much detail to be included, as these questions should still probe the expert's ability to provide estimates under uncertainty. In the included example, the previous year's incidence was provided to show how to calculate the requested estimate. This still requires experts to consider in their response whether the previous year's data were typical or unusual if there have been changes to the trend seen, and predict what might have been seen in the unpublished data. Testing of the calibration questions to ensure clarity should be included in the dry-run exercise prior to the elicitation.

Example

Background: The US Foodborne Diseases Active Surveillance Network, or FoodNet, has been tracking trends for infections commonly transmitted through food. This is done through active surveillance in the following ten states: Connecticut, Georgia, Maryland, Minnesota, New Mexico, Oregon, Tennessee, and parts of California, Colorado, and New York. Preliminary data from the previous year are released annually, usually in the spring. The most recently available data are for 2015. Data for 2016 are expected to be published in the spring of 2017.

Question: Based on active surveillance data from FoodNet, what was the incidence (per 100,000 population) of laboratory-confirmed human *Cyclospora cayetanensis* infections for the year 2016?

For example, in the year 2015, a total of 65 cases of *Cyclospora cayetanensis* were reported in the FoodNet database. This represents an incidence of 0.13 per 100,000 population.

When experts provide answers to these questions, they should rely only on their own knowledge. Therefore, they should not have access to additional resources while answering the calibration questions. This can be challenging when performing the elicitation remotely as opposed to doing an in-person workshop. After the calibration questions are completed, the experts should not be able to return to them to revise the answers.

It is often the case that experts will want feedback on how "well" they performed on the calibration questions. The standard approach is that experts should not be told how much weighting they received for the target questions, based on their answers to the calibration questions. Nevertheless, if the experts insist to know their scores, then they will be informed about the performance of their assessments. The experts agree with the method that is being used to evaluate their assessments and aggregate them based on the objective scores. This is referred to as rational consensus. It is up to the study administrators, if they have the proper approvals and wish, to provide the answers to the calibration questions.

16.4 Identification and Selection of Experts

No formal definition for what constitutes an expert exists, but appropriate identification and recruitment of expert participants is an important factor in the success of an SEJ study. A sufficient pool of potential experts is needed, as response rates can sometimes be low. The number of experts needed will depend, in part, on the number of target questions being elicited. For the US SEJ study, the initial list of experts was compiled by both the problem owner and the research team together. One method to identify additional experts not previously identified that was used in this study was snowball recruitment. This was done by asking experts to provide the names of other experts they know of in their field that may be qualified to participate, then inviting those experts if they have not been previously identified. In the US SEJ study, a total of 182 potential experts were identified based on previous work experience, topical expertise in food- and waterborne illnesses, or previous participation in SEJ studies.

Experts were invited to apply for participation via a formal email that included details about the study and application process. The application included questions on areas of expertise (e.g., microbiology, epidemiology, public health, virology, etc.), job history, education history, conflicts of interest, and self-ranking questions for individual pathogens. This self-ranking for specific pathogens was an area that introduced challenges in assigning experts to target questions. Individuals often struggle to accurately measure their own expertise, so the question was framed as "professional interest," "knowledge," and "experience" for individual pathogens using a Likert scale of high, medium, low, or none. However, this did not overcome the inherent problems with the use of self-ranking. Nonetheless, the categories of "high, medium, low, or none" were not adequately defined and consequently, the applicants interpreted their meaning differently. For example, some applicants provided a 'high' ranking due to their extensive work with certain pathogens in the past but lacked, however, any recent experience. This, therefore, led to some experts declining to provide estimates for assigned pathogens during the elicitation because they did not feel sufficiently knowledgeable. Nonetheless, the number of these cases was extremely low.

Fifty-eight experts replied to the invitation and sent their CV, along with information about their professional interests, knowledge, and experience for each of the 33 pathogens, which has been quantified using a 4-point Likert-type scale (high, medium, low, or none). The applications of experts have then been evaluated with respect to area expertise, education, work history, professional interest, and experience. Furthermore, the publication record has not been used to determine eligibility, since domain experts who might have not published frequently might have been excluded. After the selection process and some dropouts due to, e.g., unavailability on the date of the physical meeting, 48 experts participated in the elicitation. Around 44% were female experts and 56% were male experts.

No established, uniform way to assign experts to target questions exists, to the best of our knowledge. Examination for how this can be accomplished in the most scientific way should be considered early in the study design phase. The number and breadth of the target questions, as well as the number of experts, and a range of expert backgrounds should be taken into account.

Due to the large number of target questions in the US SEJ study, 15 panels consisting of related pathogens were created and experts were assigned to these panels instead of to individual pathogens. For example, one panel consisted of three different protozoa that are thought to be transmitted primarily through water, while another covered multiple serotypes of the same pathogen. Maximum bipartite matching (Asratian et al. 1998) using the *igraph* package in R[1] has been used to assign experts to the panels in the study. The parameters used ensured experts were not assigned to panels with pathogens for which they reported ‘none’ or ‘low’ experience and were assigned to provide estimates for no more than 15 pathogens. While this method is an useful tool for ensuring the highest expertise ranking to all panels, it heavily relies on the high quality of the input data. The self-ratings have been quantified by using 0 = none, 1 = low, 2 = medium, and 3 = high. The study team had to add additional points based on the review of the expert’s curriculum vitae, which was a lengthy and time-intensive process. This emphasizes, once more, the importance of careful preparation for an SEJ study.

The minimum number of experts assigned to a panel was 9, whereas the maximum number of experts assigned to a panel was 21.

16.5 Dry-Run Exercise

The dry-run exercise is an important step in determining that all documents, instructions, and questions are clear and easy to use. Documents can be provided to the dry-run participants in advance to ensure adequate time to review and formulate comments. Depending on the selected participants, this can be done in-person or

[1] https://igraph.org/r/doc/igraph.pdf.

remotely. In order to gain insight from multiple perspectives, the US SEJ study had six individuals with expertise in food- and waterborne pathogens who were not participating in the elicitation to provide feedback during the dry-run exercise. While all participant's primary work focused on food and waterborne illness, they had a variety of backgrounds including public health, government, and academia. The dry-run was conducted in a webinar format that included a review of the expert training materials, calibration questions, target questions, and fillable answer forms to be used for providing estimates. During our study, we were aware of how important it is to ensure adequate time between the dry-run exercise and the elicitation, in order to incorporate feedback and make recommended changes. We believe this was another strength of our study.

16.6 Elicitation

The formal elicitation session can be conducted in a variety of ways and this will impact some of the study design decisions. For the US SEJ study, a 2-day, in-person workshop design was chosen to standardize the process for a large number of expert participants. Individual phone calls and discussions that have been utilized in other studies would not have been feasible.

The agenda included a project introduction, a tutorial on the Classical Model for structured expert judgment along with a probabilistic training of the experts. It should be expected that most experts will be unfamiliar with providing estimates under uncertainty and specific training on probabilistic methodology is highly recommended. For this, three-domain questions have been used to train experts in reasoning with uncertainty.

The following session has been devoted to responding to the calibration questions, which was followed by one which introduced and plenary discussed the target variables and the elicitation protocol. The remaining time was dedicated to responding the target questions. During the second day, preliminary results from analyzing the calibration questions were presented and the experts were given an opportunity to revise their answers for the target questions.

Experts received a number of documents necessary for the elicitation, including the definitions of pathways and a background document with detailed epidemiological information about all pathogens, along with an extensive list of references. It has been emphasized that the background document was not meant to be exhaustive but to provide guidance and points for consideration. The document of 122 pages included a standardized table for each pathogen, with clinical and elicitation-specific information, surveillance, and outbreak data in the period 2009–2015, as well as data from case-control studies and other epidemiologic information. Available literature has been gathered in an extensive reference list. Finally, the document also included statistics on the US population.

Moreover, each expert has received an Excel file containing the elicitation instrument, along with information about panel and pathogen assignment. An example of one sheet from the elicitation instrument is provided in Appendix 16.8. Another document provided detailed instructions on completing the elicitation instrument (the Excel file), and is included in Appendix 16.9. The instructions covered specific steps and timelines, along with guidelines for providing estimates. These included specific requirements, such as the requirement that the elicited quantiles should be distinct values in ascending order, that values outside the assessed 90% confidence intervals would still be possible, but the expert would be surprised to see them and guidelines on how experts should provide estimates for very unlikely and very likely events, etc. These requirements have been detailed and exemplified in Chap. 10, this volume. Measures to reduce entry errors were included, such as error flags if the 5th, 50th, and 95th percentile estimates were not in ascending order.

As mentioned beforehand, for some pathogens, the research team in consultation with the problem owner has concluded that one or more pathways were very unlikely. Consequently, very low values have been assigned a priori and the pathway has been regarded as "blocked." The experts have been asked to indicate if they did not agree with these assumptions.

Strict timelines have been used for answering the calibration questions, that is, experts needed to complete the calibration questions at the end of the dedicated session during the first day of the workshop. Answers to the calibration questions were recorded and stored electronically for all the experts by the end of the first day and experts could not make changes to these. Experts were provided time on both the first and second days of the workshop to complete their target questions. While the experts were not allowed to access any resources for answering the calibration questions, they were encouraged to access resources provided in the background document and other available materials or to engage in discussions with colleagues. This ensured that the experts had access to as much information as possible when answering the target questions. While this might raise a question on differences between the elicitation of the calibration and target questions, we will not address this matter in this chapter.

A number of experts were able to finish answering all the target questions during the workshop. Others requested more time and they sent, via email, their assessments within a week after the workshop.

16.7 Discussion and Conclusion

The US SEJ study has been an ample study, which involved an impressive number of experts and elicited variables. Consequently, a considerable amount of time, of roughly 14 months, was required from the research team to carry out all necessary steps for careful preparation of the expert elicitation. The substantial allocated time rendered a smooth elicitation process and high expert data quality.

Despite the extensive preparation that went into the US SEJ study, re-elicitation for some variables after the in-person meeting was required. The problem owner requested further division and inclusion of other clinical manifestations for several pathogens. These re-elicitations were conducted through video webinars but finding time for all experts to participate was extremely challenging. This highlighted the importance of verifying all final materials with all involved stakeholders, especially if multiple teams are involved.

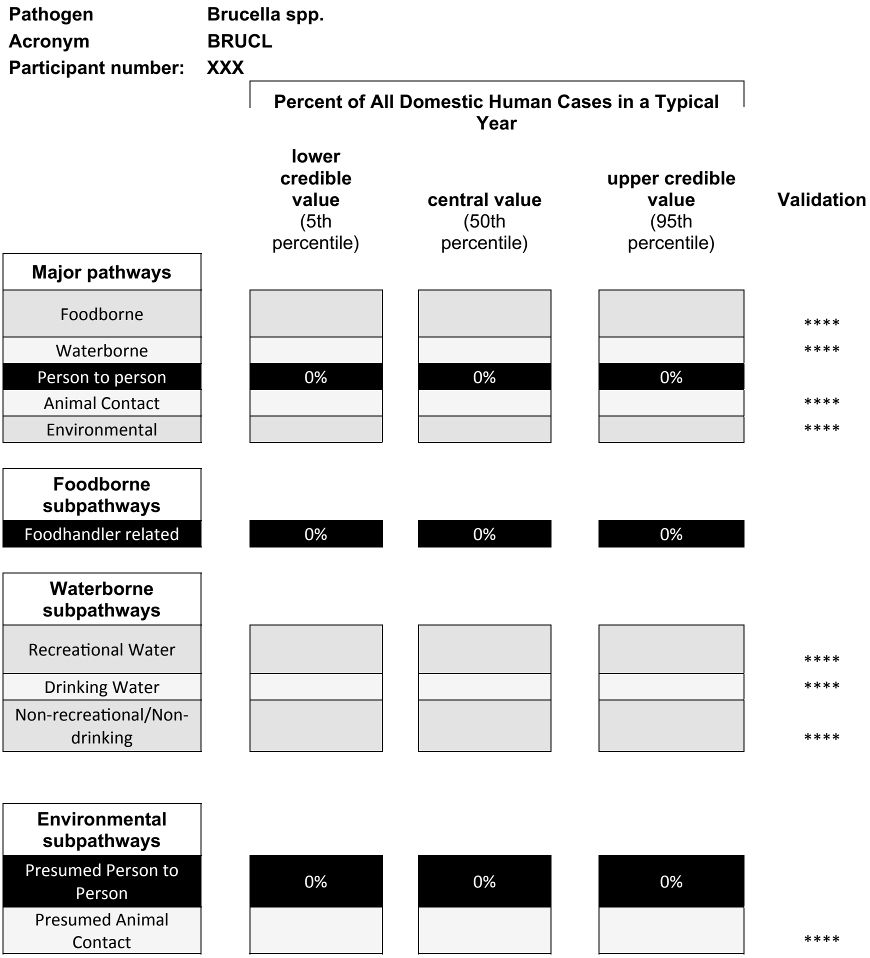
Furthermore, during the elicitation, some experts expressed their desire to discuss their estimates of the target questions together with other participants. Unfortunately, there was no time during the 2-day workshop to have discussions over the 517 total of target questions that were distributed over the 15 panels. It is worthwhile mentioning that the IDEA protocol (Hanea et al. 2018) allows for the discussion of experts' estimates in between the two rounds of individual assessments. The discussion is meant to clarify ambiguities and to allow motivation of individual assessments, and it is not mean to influence one's opinion.

A challenging aspect of the study has been the expert assignment to panels. We encountered situations when an expert's self-assessment differed from the evaluated experience from the CV or publication list. Subjective evaluations of experience can thus lead to a different interpretation of what defines a "non-relevant" experience, for example.

Overall, the US study is an example of a robust, well-executed structured expert judgment study. Attention to detail in the preparation and study design ensured the results would be high-quality and meaningful. Hopefully, this chapter provides some insight into the practical application and decisions that can occur when designing a structured expert judgment study.

16.8 Appendix 1

Elicitation instrument for Brucella bacteria and the 11 major pathways and sub-pathways. The black shaded boxes indicate the blocked (sub)pathways. Note that the corresponding percentages are not, in fact, 0%, but actually 0.000001% (for the 5% quantile), 0.0001% (for the 50% quantile) and 0.01% (for the 95% quantile). The validation flag ensured that experts would provide strictly increasing quantiles.



Pathogen	**Brucella spp.**
Acronym	**BRUCL**
Participant number:	**XXX**

	Percent of All Domestic Human Cases in a Typical Year			
	lower credible value (5th percentile)	**central value** (50th percentile)	**upper credible value** (95th percentile)	**Validation**
Major pathways				
Foodborne				****
Waterborne				****
Person to person	0%	0%	0%	
Animal Contact				****
Environmental				****
Foodborne subpathways				
Foodhandler related	0%	0%	0%	
Waterborne subpathways				
Recreational Water				****
Drinking Water				****
Non-recreational/Non-drinking				****
Environmental subpathways				
Presumed Person to Person	0%	0%	0%	
Presumed Animal Contact				****

16.9 Appendix 2

Completing the elicitation instrument

- Save the Excel file as “DATE Expert name completed.xlsx.” The original file can be used as a backup.
- Check your name in the sheet “ID.”
- You have been assigned a random participant number that will be used for data analysis and anonymous presentation. The key to link your name and the random number is known only to the elicitation team and will not be revealed to any other individual or organization.

- The sheet “CALIBRATION” will be completed in the morning session.
- For each calibration question, please provide your 5-percentile, median (50-percentile) and 95-percentile values in the grey shaded cells. Refer to the document “Calibration questions.docx” for full details. You will also receive a hard copy.
- Remember that values outside your 90% interval are still possible but you would be surprised to see them. We are not asking for a full range of possible values. By definition, a value outside your 90% interval might be observed one out of ten times.
- The cells have been preformatted for the appropriate number of significant digits or as a percentage, where applicable.
- Your percentiles should be distinct values in ascending order, with 5-percentile <50-percentile <95-percentile. If values have been entered correctly, the **** flag to the right of the entries should disappear.
- The sheet “Pathogens” contains a complete list of pathogens included in the elicitation session and the codes used in the data analysis. For each pathogen, you will find a standardized set of epidemiological data and other information in the file “Expert background document.docx.” Supplemental information can be found in pdf files that have been organized by the pathogen.
- Your personalized instrument includes target questions for only those pathogens that have been assigned to you. Please complete estimates for each and every pathogen in the afternoon session.
- For some pathogens, the elicitation team in consultation with the problem owner has concluded that one or more pathways are very unlikely. These are indicated by white font against a black background and have been assigned very low probabilities a priori. Please do not overwrite these cells.
- If you consider additional pathways very unlikely, you can also assign very low probabilities to these pathways. The data analysis program does not accept “0” probability. In such cases please enter 0.000001%, 0.0001% and 0.01% for your 5-, 50-, and 95-percentiles, respectively.
- If you consider one particular pathway very likely (virtually 100%), you can assign very high probabilities to this pathway. The data analysis program requires ascending values for the percentiles and does not accept 100%. Therefore, enter 99%, 99.9%, and 99.99% for your 5-, 50-, and 95-percentiles, respectively.
- All pathways and subpathways are mutually exclusive. Please refer to the pathway definitions when needed. The major pathways and the subpathways for the waterborne route are also comprehensive, i.e., they cover all possible transmission pathways. The medians of these pathways should sum up to approximately but not necessarily exactly 100%. To assist you in evaluating this, the sum of the medians is calculated directly below the cells where you enter your estimates.
- The worksheets have not been protected against accidental overwriting. Please enter your data carefully. In case of accidental overwriting, we will use the original file for reconstruction. Please consult the elicitation team if necessary.

- Please complete all sheets, indicated by the disappearance of the **** flag, and return your completed worksheet to the elicitation team. Also, store a copy on your computer for backup.

  Thank you!

References

Asratian, A. S., Denley, T. M., & Häggkvist, R. (1998). *Bipartite graphs and their applications* (Vol. 131). Cambridge university press.

Cooke, R. (1991). *Experts in uncertainty: Opinion and subjective probability in science*. Oxford University Press on Demand.

Cooke, R. M., & Goossens, L. J. H. (1999). *Procedures guide for structured expert judgment* (p. 18820). EUR: Project report to the European Commission.

Hald, T., et al. (2016). World Health organization estimates of the relative contributions of food to the burden of disease due to selected Foodborne hazards: A structured expert elicitation. *PLoS ONE, 11*(1), e0145839.

Hanea, A. M., McBride, M. F., Burgman, M. A., & Wintle, B. C. (2018). Classical meets modern in the IDEA protocol for structured expert judgement. *Journal of Risk Research, 21*(4), 417–433.

Quigley, J., Colson, A., Aspinall, W., & Cooke, R. M. (2018). Elicitation in the classical model. In *Elicitation* (pp. 15–36). Springer, Cham.

Scallan, E., et al. (2011). Foodborne illness acquired in the United States–major pathogens. *Emerging Infectious Diseases, 17*(1), 7–15.

Vally, H., et al. (2014). Proportion of illness acquired by foodborne transmission for nine enteric pathogens in Australia: an expert elicitation. *Foodborne Pathog Dis, 11*(9), 727–733.

https://igraph.org/r/doc/igraph.pdf.

Part IV
Applications

The chapters in this part illustrate the variety of application domains in which structured expert judgement has been used and might profitably be used.

Chapter 17
Reminiscences of a Classical Model Expert Elicitation Facilitator

Willy Aspinall

Abstract In this chapter, I trace my introduction to the Classical Model and to the thoughts and philosophy of Roger Cooke, and then go on to recount some experiences of acting as a facilitator in many real world and sometimes crucial expert elicitations. The essence of my own history is that it took me a very long time to start to understand, and appreciate, the elegance of Roger's Classical Model (Cooke 1991), its mathematical probity and how it is best deployed in application. I am sure I still haven't fully mastered the probability calculus entirely, and don't doubt others might quarrel with my preferred way of conducting elicitations that rely on the Classical Model, using a plenary workshop approach. This said, I can't find, devise or even imagine, a better alternative to the Classical Model. And, on top of this extraordinary intellectual achievement, Roger Cooke has been a beneficent collaborator *nonpareil*, always willing to help me, and anyone else, avoid self-inflicted elicitation and probabilistic infelicities.

17.1 First Encounters with Roger and His Classical Model

I was first introduced to the Classical Model by my long-term colleague and friend, Dr Gordon Woo. In the 1980s and 1990s, Gordon and I were working on a major seismic hazard assessment project in the British nuclear power industry. Gordon encountered the Classical Model, and its author, Roger Cooke, thirty years ago at PSA '89 in Pittsburgh. At the time, Gordon was involved in the elicitation of expert judgement for the disposal of radioactive waste, being undertaken for the UK Department of the Environment. With his customary perceptiveness, Gordon had instantly appreciated the mathematical sophistication of the model, and enjoyed Roger's dynamic presentational style—neither, it must be said, matched at the podium by any sartorial elegance.

W. Aspinall (✉)
University of Bristol and Aspinall & Associates, Bristol, UK
e-mail: Willy.Aspinall@bristol.ac.uk; Willy@aspinallassociates.com

A. M. Hanea et al. (eds.), *Expert Judgement in Risk and Decision Analysis*,
International Series in Operations Research & Management Science 293,
https://doi.org/10.1007/978-3-030-46474-5_17

Gordon also recognised the potential of the method for formalised uncertainty quantification and the enumeration of tricky variables, on a rational and transparent basis by expert judgement, when data are sparse or non-existent; the scope of subjects to which he has introduced Cooke's method subsequently is wide and varied (for examples and references, see Chap. 22, this volume). When Gordon suggested we could use the approach in one of our seismic hazard assessments, my interest was immediately piqued because, for a long time, I had been concerned about the way volcanologists were handling uncertainties and judgements in eruption crises.

17.1.1 Expert Judgement and Volcanic Eruptions

Many years earlier, in 1976, there had been a major volcanic crisis on the island of Guadeloupe, in the West Indies and, at the time, considerable disagreement and contention developed among scientists monitoring the volcano (I was there and observed this evolving débâcle, but did not participate in top-level scientific discussions). Senior scientists indulged in public disputes about the state of the volcano and its potential for a dangerous eruption, making individual and selective inferences from imprecise, uncertain and unreliable monitoring data. These contending views were aired, with characteristic Gallic *férocité*, in public and before the media, and the upshot was an unfortunate breakdown in the credibility of scientific advice. The political and social ramifications of this particular episode afflicted applied volcanology for many years thereafter, particularly in the Eastern Caribbean region. (For a description of the context and circumstances of the 1976 crisis, see a retrospective analysis by Hincks et al. (2014); these authors formulated a probabilistic re-interpretation of what were, then, the key volcanological issues, using modern judgements elicited with the Classical Model to quantify relevant uncertainties).

Gordon introduced the concept and principles of the Classical Model into our work for seismic hazard (noting *en passant* that his initiative in this domain was killed at birth by Thatcherism, the privatisation of the UK nuclear power industry and a consequent discharging of forward-thinking consultants). I immediately felt that it offered a viable way for volcanologists in crisis conditions to handle the sort of uncertainties that had blighted the scientific inputs to the Guadeloupe episode.

Gordon and I, therefore, wrote a short paper for a conference in Rome on large explosive eruptions (Aspinall and Woo 1994), in which we outlined the way the Classical Model might be used for managing volcanological judgements, illustrating the conceptual procedure with some imagined crisis examples. When I presented the paper at the conference it was received with interest, by some, but dismissed out-of-hand by the senior leading Italian volcanologist of the day. While I don't now recall his exact criticism of our proposition, his opinion was indicative of a deterministic unwillingness to consider probabilistic theories or Bayesian thinking, or to cede personal scientific authority to the wider—likely wiser—collective judgements of other colleagues, pooled via an impartial numerical algorithm.

Parenthetically, one should explain, for the benefit of younger colleagues brought up in the twenty-first century pre-Brexit era of easy international scientific exchanges, that the academic world of Italy in the 1970s was very hierarchical and deferential. In Italian volcanology, this has completely changed now, and several major projects have embraced expert elicitation. Moreover, this author has recently witnessed a similar sea change in Japan, having conducted an expert elicitation there in relation to natural hazards aspects of radwaste facility siting.

Going back to 1994, when Gordon and I made our tentative suggestion to use a structured expert elicitation approach for volcanic hazard assessment, little did I anticipate that the Classical Model would be soon utilised in earnest. But, before that, in April of the following year, Roger, with other luminaries of the probability analytics firmament (viz. Simon French; Jim Q Smith; Tim Bedford) presented a University of Cambridge Programme for Industry (CPI) course on data acquisition and dependence modelling for safety and risk assessment, which I signed up for. My aim—for a modest outlay it must be said—was to learn more about the Classical Model at the feet of the Master (I did gain much, too, from the other tutors). Then, literally within a few weeks, the volcano on the island of Montserrat, a British Overseas Territory, suddenly and unexpectedly awoke, after being quiescent for more than 350 years.

In the ensuing public alarm and administrative concern, I became involved in providing scientific advice to the British Government and the local government, sitting betwixt two sets of scientists whose perspectives on what might happen were becoming increasingly polarised. Thus, just as in Guadeloupe twenty years earlier, once again seemingly irreconcilable differences in opinion and judgement emerged between two scientific groups. This dichotomy of advice provided a headache for the authorities, who wouldn't and couldn't know whose line of advice they should follow.

My role was to act as an intermediary and to attempt to broker a common, agreed view for the authorities so that they could make informed decisions about hazard mitigation and the protection of life and limb. In order to move these often fractious—and very time consuming—scientific disputations away from quasi-deterministic dogmatic deadlock, I made the recommendation to evaluate all projections of future eruptive scenarios probabilistically, and to express these within a framework of quantified uncertainties typically associated with such forecasts. In this way, and by processing elicited judgements with the Classical Model, those scenarios that had the least degree of scientific dispute were accompanied by smaller credible intervals, while wider credible intervals characterised differences in judgement that were more extreme.

The big challenge, then, was to convey to the decision-makers what the elicitation findings were signalling for the purposes of decision support. While the uncertainty estimates might be thought likely to dilute the strength of scientific advice, it was only fair that the problem owners were apprised of the ranges of views being expressed on any particular issue. In addition, there was—at least from my perspective—a selfish motivation: scientists providing advice in critical, life-or-death situations can lay themselves open to litigation or criminal charges if a disaster ensues. The Classical

Model seemed to offer a neutral, depersonalised basis for articulating such judgements although, as far as I know, the procedure, and results, have never been tested in a court of law.

17.1.2 An Application to Airline Operations Safety

A year or so after the Classical Model was introduced into observatory practice in Montserrat, an opportunity came up for me to apply it in an important flight safety improvement initiative within the operations branch of a major airline. This exercise benefitted greatly from the enlightened involvement of the problem owner, the late Capt. John Savage (who just happened to be an old school friend). With John's insightful guidance on aviation issues, we learned a lot about the art of devising meaningful 'calibration' (or 'seed') questions for a Classical Model elicitation. Once the elicitation had been undertaken, what surprised me most was the finding that a large group of highly trained, senior and experienced airline pilots can be just as diverse as others, such as volcanologists, when it comes to calibrated individual abilities to judge uncertainties within their own domain of expertise. Moreover, we found that the Classical Model calibration score rankings of individual pilots accorded closely with the chief pilot's judgement of the professional capabilities of those same captains. This made it easier to convince the airline management of the efficacy of the Classical Model as a worthwhile decision support tool and, in effect, validate it for their application.

17.1.3 Spreading the Word

Sad to say, this progressive endeavour—to break into the realms of airline corporate management—foundered subsequently with the 11 September 2001 attack on the Twin Towers in New York. Because of the financial impact of that episode on the airline, a guillotine was dropped on external consultants like me, and the Classical Model became neglected in that particular organisation. However, the experience gained in that case history was not lost.

In 1996, Roger and his colleagues organised a second CPI course and, this time, I was kindly invited along to give a presentation, sharing the Montserrat and the (unpublished) airline applications of the Classical Model with a new group of interested industry and academic participants. These annual courses continued for several more years and gradually I insinuated more and more applications into my contribution.

It is testimony to the generosity and openness of Roger and the others that I was allowed to use some of their valuable course time to grandstand my tentative attempts at facilitating expert elicitations, using the Classical Model.

17.2 Some Observations from Classical Model Application Experience

17.2.1 Calibration Anxieties

Whereas the management and problem owner in the airline case were quick to recognise the validity, and value, of a structured expert elicitation for their purpose, it has been a recurring theme, in many of the elicitations which I have facilitated, that the problem owner has some anxiety about the way experts are calibrated. This extends to the ‘political’ implications of zero weighting many, or nearly all of the participating experts. For participants and problem owner alike, there is a perhaps understandable apprehension about how harshly the full optimisation Decision Maker (DM) of the Classical Model can reduce a group of experts to just one or two people with positive weights.

In some applications, the problem owner sought to have the calibration power of the model reduced so that a reasonable quorum or even a majority of a group received some weight. In practice, this did not have a big impact on the quantiles calculated through the DM; the judgements of those experts with greater weights naturally dominated a pooled combination of those admitted with some weight. As a consequence, the quantile values obtained with reduced power tended to be very similar to those obtained with full optimisation.

These days, with the findings of an in-depth analysis of many classical model elicitations, researched in the TU Delft elicitation database and reported by Colson and Cooke (2017), the challenge of providing conviction to a problem owner about his or her elicitation findings has become easier. If the problem owner does not wish to avail themself of the optimised DM results, the facilitator no longer needs to make an arbitrary choice, on an ad hoc basis, about how to reduce calibration power or statistical accuracy threshold. Abby Colson and Roger found that if the statistical accuracy threshold in the Classical Model is set to about $P = 0.01$, then about one-third of any group of experts receive positive weights.

This ‘one-third’ rule of thumb is found to work fairly consistently over a wide range of elicitations in all sorts of subject matter areas. As such, it provides an arguable basis for widening the catchment of expert judgements in any particular study in order to assuage a problem owner’s concerns about inflexible or over stringent expert scoring. Moreover, in case the problem owner is tempted by the apparent democracy of equal weights, Colson and Cooke demonstrate—with an out-of-sample validity index—that Classical Model performance weighting outperforms equal weighting in twenty-six of thirty-three post-2006 TU Delft studies; if there were no difference between performance weighting and equal weighting, this number of successes in 33 trials has a probability of only 0.001.

17.2.2 Probabilistic Forecasts

Apprehension about capturing a reasonably representative sample of judgements for a distinct problem can be all the more pronounced in circumstances where the experts are being asked to provide probabilistic forecasts of future events. There can be an intrinsic difficulty in devising seed questions which meaningfully and accurately reflect experts' quantitative judgement capabilities in making probability estimates and related credible intervals. Sometimes it is feasible to formulate some seed questions of a predictive nature that can be used for calibration, if observations or measurements become available within a reasonable future timescale. However, this is rarely the case when the elicitation is required to provide urgent decision support, and calibration scores can't be delayed. What can be said is that many repeated probability forecasts were elicited for prospective hazardous events at the Montserrat volcano, and that, (with performance weights based on physical volcanology seed questions), the volcanologists displayed good overall forecast skills (Wadge and Aspinall 2014).

In the latter case, one or two apparent forecast failures, out of more than eighty made over two decades, could be explained by inadequacies in the way the target questions had been posed. For instance, on one occasion the volcanologists were asked to ascribe probabilities to defined ranges for the volume of an impending dome collapse. One range covered volumes from 50 to 100 million m^3 magma, a very large collapse of the type. The team afforded this scenario a high probability of occurrence within a given limited timescale. When the event happened, the estimated volume of collapse was 105 million m^3, which put it into the next higher range, for which the elicited probability was lower. The reality was that a big, and potentially dangerous collapse had been anticipated by the volcanologists and, given the uncertainties associated with estimating such collapses in the field (much of the material disappears into the sea), it is far from certain that the volume lost actually exceeded 100 million m^3. On the face of it, while the forecast appeared poor in terms of a numerical skill score it was apposite for the circumstances.

A similar 'bad' forecast at the volcano involved an eruption event which the volcanologists judged was most likely to take place within a certain time window. The incident in question eventuated just a day or two after that defined time span, again suggesting low forecast skill. The skill measurement does not concede that the scientists were, indeed, providing a valuable outlook for decision-makers, the arbitrary thresholds notwithstanding. It is, therefore, important that target questions that concern future event likelihoods of occurrence are framed very carefully, perhaps by defining scenarios in terms of escalating cumulative exceedance probabilities.

17.2.3 Experts and Their Calibration Scores

All DM solutions are, of course, derived by pooling individuals' quantile judgements using their performance scores as weights. In practice, with the plenary approach (i.e. eliciting a group of experts in a workshop setting, rather than via one-on-one interviews with the facilitator), it is usual to report back to the group their responses to the calibration questions as range graphs showing individuals' credible intervals and median values, together with the known realisation values. Then the resulting personal performance scores are shown (i.e. the product of their statistical accuracy score with their information score), and the make-up of the DM and its scores. For neutrality, this is done without identifying the experts by name.

It is noteworthy that, for the many elicitations I have conducted, involving participants numbering in the hundreds, there has been only one person who demanded to be told which expert they were in the ranking and what their own performance score was. With this single exception, it seems that once experts are sufficiently familiar with the concept and principles of the Classical Model and are satisfied that this is a procedure for determining a rational collective consensus for a constructive purpose, then they were content to contribute their judgements under conditions of anonymity and non-attribution. An important feature of the process under these conditions is that it is depersonalised, and the attendant objectivity—and opportunity to express true beliefs without fear of criticism or ridicule—appeals to (nearly) all scientists.

Now, this is not to say individual experts are totally disinterested in their performance scores, as the following true anecdote illustrates. The incident presented me with a challenge that demonstrated that I was not fully informed myself about the way the Classical Model measured an expert's statistical accuracy. After one elicitation at the Montserrat Volcano Observatory, a very senior colleague, scrutinising the set of calibration range graphs and the performance scores within the group, as recorded in one of the observatory internal reports, had managed to identify himself in the plots (he had kept his own notes of his responses to seed and target questions). He realised that his statistical accuracy score was identical to that of another expert and he spotted also that his median values for the calibration items were generally closer to the realisation values than those of his fellow expert. The affronted scientist challenged me, quite politely, as to how this could be right and fair.

I must confess that my explanation was struggled and likely mathematically defective. At that stage in my foray into the more arcane reaches of structured expert judgement, I hadn't fully appreciated that there is no implicit distance metric in the Classical Model formula for computing statistical accuracy and that, as long as the two experts had identical counts of seed item realisations in their respective quantile-defined probability bins, then they would achieve the same P-value.

17.2.4 No Statistical Distance Metric in the Classical Model

The statistical distance of a seed item realisation—relative to the expert's median—does not enter the equation in the Classical Model and, as a matter of numerical principle, is not called for. This insight was not immediately obvious to me, even after using Roger's method for a couple of years, and, I suspect, it is not evident to many fellow scientists if it is not pointed out to them explicitly when briefing them for an elicitation. The way statistical accuracy is measured guarantees one of the key attributes of the Classical Model: scale invariance is preserved across a set of calibration questions that, almost inevitably, will involve different scale units. Otherwise, how does one combine accuracy and information metrics from variables enumerated against different units, e.g. mass, length, time, velocity, etc.?

This didn't stop my colleagues (and I) from endeavouring to uncover a 'better' formulation for expert aggregation that incorporates some element of statistical distance measurement in the expert scoring. We developed an alternative scheme, the Expected Relative Frequency method (Flandoli et al. 2011), which can do better than the Classical Model under certain circumstances but if, and only if, the sole purpose is to obtain accurate estimates of target item mean (expected) values from quantile elicitations; as we discussed in our paper, the Relative Frequency Method is demonstrably out-performed by the Classical Model when the need is for reliable and informative credible intervals for uncertain variables.

17.2.5 A Potential Numerical Infelicity with Intrinsic Range

One of the first steps that the Classical Model algorithm undertakes in the EXCALIBUR software, before computing experts' calibration scores and decision-maker solutions for target questions, is to set up an 'intrinsic range' for each item in the elicitation (i.e. every seed item and every target question has its own unique intrinsic range). This is done by finding the lowest value ascribed by any member of the expert group to a particular item, at the lower elicited quantile (usually 5th percentile) and, likewise, by identifying the highest value offered by any expert at the upper quantile level (usually 95th percentile). The spread from lowest value to highest is then expanded by adding increments at each end; these increments are usually 10 % of the span from lowest-to-highest values, although other options can be implemented in the EXCALIBUR programme. The purpose of this intrinsic range is to provide a reference variable space over which to establish a uniform (or log-uniform) distribution against which individual expert's informativeness can be measured (for details, refer to p. 128 et seq. of Cooke 1991).

This intrinsic range construction is normally straightforward and uncontentious. But there are potential numerical consequences if some expert gives a 5th percentile value for a variable that is enumerated at, or very close, to zero. In such cases, the intrinsic range extension can take the variable baseline span into negative territory (or, possibly, beyond some limiting upper value at the other end). For some variables, negative values may be non-physical, and yet the decision-maker(s), computed via EXCALIBUR distribution file output, can report percentiles that are negative. And, if this distribution is then sampled in post-processing for further analysis, then negative values can emerge in those samples, too.

Such an outcome might be regarded by some as compromised by the fact that some experts, who have influenced item intrinsic ranges, are subsequently zero weighted for the decision-maker solution. Under normal circumstances, however, I am tempted to describe this as a virtue of the model, in that all experts contribute to intrinsic range definition and the delineation of bounds to the problem items; this notwithstanding the information provided for target item distribution support, via their elicited quantile values, is ultimately zero weighted.

The point is that an uncritically adopted intrinsic range can lead sometimes to spurious or ill-constrained extremes in uncertainty spreads or, worse yet, to non-physical values for variables of interest. This is an incipient condition which the facilitator/analyst needs to be alert to, although it is not clear-cut how best it can be mitigated, at the elicitation stage or in subsequent data processing.

Roger Cooke argues, forcefully and almost certainly correctly, that the analyst cannot, on a post hoc basis, censor out experts who have already given their judgements by elicitation—no matter their performance scores are negligibly low to zero. They remain selected experts and their contributions to intrinsic range definition, at least, must be regarded as valid. When such a situation arises, the analyst will, inevitably, struggle to decide what to do once the condition emerges into the light, and how to message this numerical syndrome back to the experts and the problem owner.

17.2.6 Not just Another Opinion Survey Technique

Elsewhere, in the literature, at conferences and in discussions about alternative elicitation procedures with problem owners, it is apparent that an elicitation with the Classical Model at its heart is presumed by many to be just another polling or opinion surveying technique—a prejudice likely boosted in recent times by the increasingly easy accessibility of online tools which don't call for any great mathematical assurance, or astuteness. Furthermore, the distinctive, and formally proper, numerical principles of the Classical Model are sometimes incorrectly implemented in open source codes, or erroneously misrepresented in the literature; it is difficult to refute all such misconstruals and attempts to do so are not always welcomed by journal editors (for one published rebuttal, see Bamber et al. 2016, or see Chap. 11).

Sadly, as with many other areas in contemporary science, engineering and medicine, it is easy to belittle or simply dismiss on specious grounds that which is not properly understood. A search for a new model, which improves on the Classical Model, must go on—and Roger has always encouraged this. But I, for one, will be very surprised if anything substantially better emerges any time soon. For those interested, there is a recent sister book (Dias et al. 2017), with chapters which demonstrate the authority and formal standing of the Classical Model, as adduced by praxis and mathematical validation evidence, and by its appraisal in mainstream decision and risk analysis.

17.3 An Apologia and an Encomium

It must be stressed the history of events and the views expressed here are my own and should not be attributed to Roger or any other colleague. It is fair to say that, like Winnie the Pooh, I am a ‘bear of very little brain’. Thus, all throughout my professional and academic career, I have been generously and unselfishly tutored by many clever people—who have been willing also to overlook or ignore my persistent, sometimes fatuous, scientific querying of theories, hypotheses and models. With fellow Earth Scientists, it has been possible to ensure they remain mainly supportive by buying a few cold beers at the end of long field days; in the case of Roger, alcohol has not been necessary, but an occasional supply of jars of my wife’s Trinidadian pepper sauce—‘voodoo sauce’, as Roger terms it—has oiled our collaboration nicely.

In short, Roger has been excellent and patient collaborator who, over more than a quarter of a century, has never failed to answer my inane queries about the workings of the Classical Model. Despite, or because of, the fact I have been using his model in earnest for more than 25 years, still I am not convinced I fully understand all the subtleties and strengths of the underlying mathematics. If I have a single message for young colleagues, who might be contemplating alternative ways of eliciting expert judgements, it is to opt for the Classical Model, and try to gain as much understanding of its workings as possible.

Thus, my debt of gratitude to the ineffably erudite Roger is great, and I hope he will, himself, remain very active in expert judgement praxis for many years to come—he is pretty much irreplaceable.

References

Aspinall, W. P. & Woo, G. (1994). An impartial decision-making procedure using expert judgement to assess volcanic hazards. Accademia Nazionale dei Lincei-British council symposium large explosive eruptions, Rome, 24–25 May 1993. *Atti dei Convegni Lincei* 112, 211–220.

Bamber, J. L., Aspinall, W. P., & Cooke, R. M. (2016). A commentary on “how to interpret expert judgment assessments of twenty-first century sea-level rise” by Hylke de Vries and Roderik SW van de Wal. *Climatic Change, 137*(3), 321–328. https://doi.org/10.1007/s10584-016-1672-7.

Colson, A. R., & Cooke, R. M. (2017). Cross validation for the classical model of structured expert judgment. *Reliability Engineering and System Safety, 163,* 109–120. https://doi.org/10.1016/j.ress.2017.02.003.

Cooke, R. M. (1991). *Experts in Uncertainty* (p. 321). Oxford University Press.

Dias L. C., Morton A. & Quigley J. (Eds.). (2017). Elicitation: The science and art of structuring judgement. *International Series in Operations Research & Management Science* (p. 542). Springer, New York. (ISBN 978-3319650517 First edn. 20 December 2017).

Flandoli, F., Giorgi, E., Aspinall, W. P., & Neri, A. (2011). Comparison of a new expert elicitation model with the Classical Model, equal weights and single experts, using a cross-validation technique. *Reliability Engineering and System Safety, 96,* 1292–1310. https://doi.org/10.1016/j.ress.2011.05.012.

Hincks, T. K., Komorowski, J.-C., Sparks, R. S. J., & Aspinall, W. P. (2014). Retrospective analysis of uncertain eruption precursors at La Soufrière volcano, Guadeloupe, 1975–77: Volcanic hazard assessment using a Bayesian belief network approach. *Journal of Applied Volcanology, 3,* 3. https://doi.org/10.1186/2191-5040-3-3.

Wadge, G. & Aspinall, W. P. (2014) A review of volcanic hazard and risk assessments at the Soufrière hills volcano, montserrat from 1997 to 2011. Ch. 24. In: G. Wadge, R. E. A. Robertson & B. Voight (Eds.), *The Eruption of Soufriere Hills Volcano, Montserrat, from 2000 to 2010* (Vol. 39, pp. 439–456). Geological Society Memoirs. London: Geological Society.

Chapter 18
Dealing with Imperfect Elicitation Results

Rens van de Schoot, Elian Griffioen, and Sonja D. Winter

Abstract The trial-and-roulette method is a popular method to extract experts' beliefs about a statistical parameter. However, most studies examining the validity of this method only use 'perfect' elicitation results. In practice, it is sometimes hard to obtain such neat elicitation results. In our project about predicting fraud and questionable research practices among Ph.D. candidates, we ran into issues with imperfect elicitation results. The goal of the current chapter is to provide an overview of the solutions we used for dealing with these imperfect results, so that others can benefit from our experience. We present information about the nature of our project, the reasons for the imperfect results and how we resolved these supported by annotated R-syntax.

18.1 Introduction

The trial-and-roulette method, also called the chips and bins method or the histogram method, is a popular method to extract experts' beliefs about a statistical parameter (Clemen et al. 2000; Goldstein et al. 2008; Goldstein and Rothschild 2014; Gore 1987; Haran et al. 2010; Haran and Moore 2014). During the elicitation procedure experts are provided with a number of 'chips' to allocate probability to specific values of the parameter space. The number of chips placed over a certain value reflects the expert's view on the probability of that value. The method has been used in different fields (e.g. Johnson et al. 2010), but hardly in the social and behavioural

R. van de Schoot (✉) · E. Griffioen
Department of Methodology, & Statistics, Faculty of Social and Behavioural Sciences, Utrecht University, Padualaan 14, 3584 CH Utrecht, The Netherlands
e-mail: a.g.j.vandeschoot@uu.nl

E. Griffioen
e-mail: griffioenelian@gmail.com

S. D. Winter
Department of Psychological Science, University of California, 5200 Lake Rd, Merced, CA 95340, USA
e-mail: swinter@ucmerced.edu

A. M. Hanea et al. (eds.), *Expert Judgement in Risk and Decision Analysis*,
International Series in Operations Research & Management Science 293,
https://doi.org/10.1007/978-3-030-46474-5_18

sciences. One exception is the study by Zondervan-Zwijnenburg et al. (2017), who made a case for using the method as developed by Johnson et al. (2010) in the social and behavioural sciences. They tested the method with behavioural practitioners who provided their judgements with respect to the correlation between cognitive potential and academic performance for two separate populations enrolled at a special education school for youth with severe behavioural problems: youth with Autism Spectrum Disorder (ASD), and youth with diagnoses other than ASD. They also investigated face validity, feasibility, convergent validity, coherence and intra-rater reliability and concluded that the method can also be used in the social and behavioural sciences. Veen et al. (2017b) adjusted the method by adding a step to the procedure, providing the experts with visual feedback during the elicitation process.

However, these studies by Johnson et al. (2010), Veen et al. (2017b) and Zondervan-Zwijnenburg et al. (2017) use only 'perfect' elicitation results. That is, in an ideal situation the expert places the stickers neatly on top of and equally spaced next to each other within the allowed parameter space, see Fig. 18.1a. Subsequently, the best fitting probability distribution is computed with software like SHELF (Sheffield Elicitation Framework; Oakley and O'Hagan 2010). The accompanying hyperparameter values, see Fig. 18.1b, can then be used for other purposes, like Bayesian updating with data (see, e.g. Zondervan-Zwijnenburg et al. 2017) or computing prior-data conflicts (Veen et al. 2017a). In practice, however, such neat elicitation results are hard to obtain and the results will look more like the ones in Fig. 18.2.

In a project about predicting fraud among Ph.D. candidates, we ran into issues with such imperfect elicitation results. The goal of the current chapter is to provide an overview of the solutions we used to deal with these imperfect results, so that

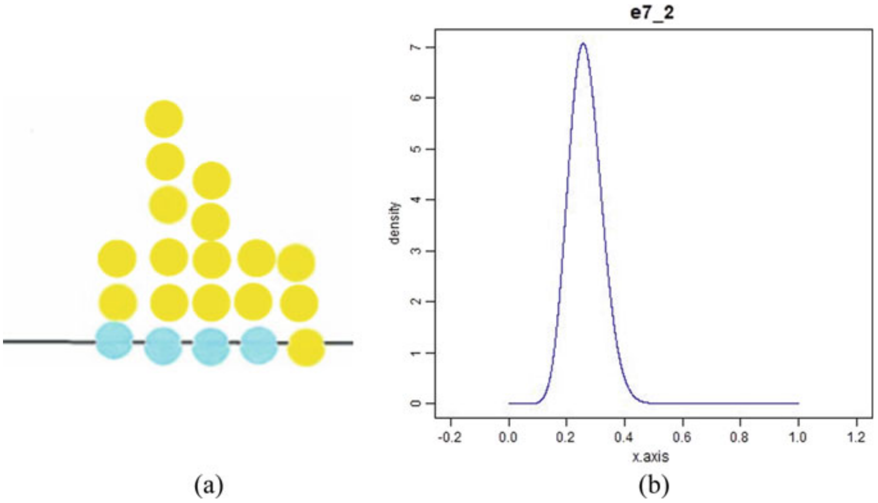


Fig. 18.1 The result of **a** 'perfect' elicited distribution using the trial roulette method and **b** the probability distribution obtained with the SHELF software resulting in a Beta distribution with hyperparameters 21.40 and 80.67

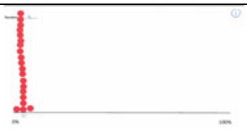
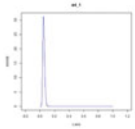
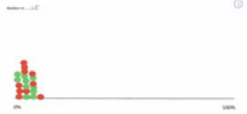
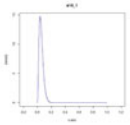
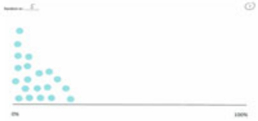
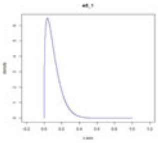
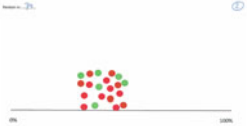
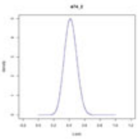
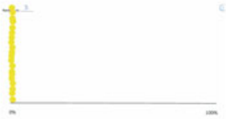
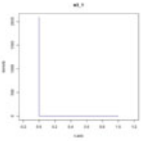
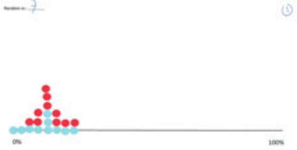
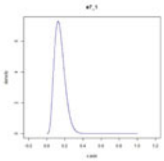
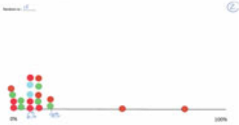
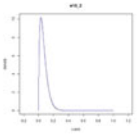
Issue	Stickered Distribution	Parametric Distribution	Hyper parameters (shape - scale)
Issue 1: The expert includes numbers/ percentages with their distribution			14.21 – 267.02
Issue 2: Stickers are pasted in neat vertical stacks with different vertical distancing			3.07 – 52.21
Issue 3: Stickers are not stacked exactly on top of each other			1.40 – 11.74
Issue 4: The distribution lacks stickers in a specific part			16.06 – 22.41
Issue 5: All stickers at one point			4.59 – 10963.51
Issue 6: Stickers fall (partially) outside of the x-axis			5.42 – 32.22
Issue 7: One sticker is an outlier			1.70 – 22.89

Fig. 18.2 Examples of imperfect elicitation results for seven different situations, the statistical distribution obtained via our solutions, and the results from SHELF. Results for all experts can be found on the OSF

others can benefit from our work. In what follows, we first present more information about our project and the reasons for the imperfect results, followed by a discussion of how we dealt with this. We will present our solutions with annotated R-syntax. All elicitation results, the R-code to apply our solutions, the SHELF input, all resulting parametric distributions, etc., can be found on the Open Science Framework (https://osf.io/bq28j/).

18.2 Case Study: Predicting Fraud Among Ph.D. Candidates

Academic integrity has attracted increasingly more attention over the past years (Steneck 2006; John et al. 2012). Studies quantifying the prevalence of questionable research practices and fraud reveal that a substantial number of scientists does not behave according to academic integrity standards (Fanelli 2009; Martinson et al. 2005; Tijdink et al. 2014). Yet, obtaining grants or getting a job is still highly dependent on (the number of) publications (Sonneveld et al. 2010; Van de Schoot et al. 2012). According to Hofmann et al. (2013) there is a tendency amongst young scholars to respect and learn from the scientific norms and practices of other scholars. With the increasing time pressure and publication pressure and the growing number of scholars and the interdisciplinary and international studies being conducted, academic norms have become too diverse and complicated; we cannot and should not simply copy them from one another.

However, young scholars and especially Ph.D. candidates rely heavily on their supervisors and will mimic their behaviour (Van de Schoot et al. 2013). They are in a dependent relation with one or more senior faculty member which makes them prone to senior pressure. In other words, the senior faculty member has a great influence on the Ph.D. candidate and his or her behaviour may thus also influence the Ph.D. candidate's scientific behaviour. Our project is about investigating the ways in which behaviour of senior scholars influences the behaviour of Ph.D. candidates with respect to questionable research practices. Note that the entire study was approved by the Ethics Committee of the faculty of Social and Behavioural Sciences at Utrecht University (FETC15-108), and that the questionnaires were co-developed and pilot-tested by two university-wide organization of Ph.D. candidates as well as the Dutch National organization of Ph.D. candidates.

We developed several scenarios based on different types of questionable research practices/fraud, ranging from objectively fraudulent behaviour (data fabrication), via serious forms of misconduct (deleting outliers to get significant results), to arguably milder forms of questionable research practices (salami-slicing); see the text we used in the Box 18.1.

Box 18.1 Text used for the three Scenarios

Suppose you have been working on this research project with the project leader and senior team member for a few months. The following situation occurs. Together with the project leader and the senior team member you are developing an article. You are in charge of the data analysis.

Scenario 1—data fabrication:

When you are working on the analysis, you discover that something is wrong with the data: you have good reasons to assume the data has been made up, most likely by the senior team member, who was responsible for data collection. You discuss this point of concern with the senior team member, but you have not discussed this with the project leader. The senior team member advises you to use the data anyway because it leads to very interesting conclusions. You figure that publishing the results based on these data might result in a very good article which will be crucial in allowing you to finish your thesis in time.

Scenario 2—deleting outliers to get significant results:

You checked the data and there appears to be no problem with it. However, your most important hypotheses are not supported by the data. You discuss this point of concern with the senior team member, but you have not discussed this with the project leader. The senior team member proposes to reanalyse the data together. Before doing this he removes some outliers/interview quotes that, according to the senior member, disturb the data. He provides no further information. The new analysis shows support for your hypotheses. The senior team member advises you to use this data because it leads to very interesting conclusions. You figure that publishing the results based on this data might result in a very good article which will be crucial in allowing you to finish your thesis in time.

Scenario 3—salami-slicing:

You checked the data and there appears to be no problem with it. Your analysis shows support for your main hypotheses. However, with the current analysis it seems as if you will be able to publish just one article based on this research project. You discuss this point of concern with the senior team member, but you have not discussed this with the project leader. The senior team member asks you to analyze the data in such a way that the group can publish three similar articles instead of one, based on the same dataset. The three proposed articles will differ from each other marginally. You figure that publishing three articles instead of one will be crucial in allowing you to finish your thesis in time.

Question: would you try to publish the results of this study?

The project involved asking 36 senior scholars working at 10 different faculties of Social and Behavioural Sciences or Psychology in The Netherlands—such as deans, vice deans, heads of department, research directors and confidentiality persons—what they know about the behaviour of Ph.D. candidates regarding questionable research practises. We asked them to indicate the percentage of Ph.D. candidates who would answer ‘yes’ to the question at the end of the vignette in Box 18.1: would you try to publish the results of this study?

In a face-to-face interview with the first author (RvdS), the participants were asked to place twenty stickers, each representing five percent of a distribution, on an axis where the left indicates that 0% of the Ph.D. candidates in their faculty would say ‘yes’ to the scenario, and the right indicates 100%. The placement of the first sticker on a certain position on the axis indicated perceived likeliness by the expert for that value. The other stickers represented other plausible values. If all stickers are placed exactly on top of each other, this indicates the expert was 100% certain the observed percentage would be that particular value. Stickers placed next to each other resemble uncertainty about the estimate.

Ideally, the elicitation procedure resulted in a neat stickered distribution like the one in Fig. 18.1a which could easily be transformed into a probability distribution using the software `SHELF`. However, most of the results were not this ‘perfect’: see Fig. 18.2. This was due to lack of time or stickers being too small for large hands (really!). The challenge was to translate the intended distribution of the expert into meaningful input to be fed to the software `SHELF` in order to obtain logical statistical distributions. The goal of the current chapter is to describe our procedure.

18.3 The Perfect Situation

To transform the stickered distribution into a parametric distribution, we applied the following steps:

1. The x-axis was divided into 1.000 sections. The density of each section depended on the height of the stickers, each sticker representing 5% of the density mass. The 5% was divided across the sections where the sticker was placed, using the following rule: 5/[number of sections on x-axis] = [density per section]. With a triangle ruler, for each stack, lines were drawn from both sides of the lowest sticker, perpendicular to the x-axis. The distance from the left edge of the x-axis (0%) to the left and the right line was then measured and rounded to a tenth of a cm. This distance was then used to compute the proportion of the x-axis (rounded to 1 decimal) that the sticker-stack occupies, using the following formula: [position on x-axis in cm]/[length x-axis (here 25.8 cm)] * 100%. This delivered a left and right edge of the interval and the proportion corresponds to a number of sections on the x-axis (each section = 0.1%). We used this information to create a vector of numbers in R using the following formula: [number of stickers] * [percent per sticker (here 5%)]/[number of sections this stack of stickers covers]. If there were no stickers for a certain interval, we repeated the values 0 for those sections.

For the distribution in Fig. 18.1a, the following procedure was applied. In the sections from 0 to 17.5% on the x-axis, no stickers are pasted, represented in a vector with `rep(0,175)`. The first stack containing three stickers (so 3*5% of the total density mass) covers 31 sections on the x-axis and is therefore represented as `rep(3*5/31,31)`. After an empty space of 8 sections `[rep(0,8)]`, the procedure repeats itself ending at the empty interval at the right of the distribution `[rep(0,634)]`. Finally, a vector that shows the density at each section of the x-axis is acquired using the following R-syntax:

```
c(rep(0,175),rep(3*5/31,31),rep(0,8),rep(6*5/31,31),rep(0,12),rep(5*5/31,31)
,rep(0,8),rep(3*5/31,31),rep(0,8),rep(3*5/31,31),rep(0,634))
```

2. The resulting vector with section densities was used as input for the elicitation programme `SHELF` using the following syntax:

```
source("shelf2.R")
elicit.group.values(N.experts=1,method="rp",Lo=0,Up=1)
```

After the final line of code, a window opens in which the density of each section is filled in by hand. The software then computes the best fitting beta distribution, and the hyperparameters of this distribution can be requested using the following command: `elicited.group.data`

The hyperparameters can be used to create a plot of the distribution using the following syntax (Fig. 18.1b): `curve(dbeta(x, 21.40791, 80.6748))`

3. The parametric distribution plot (Fig. 18.1b) was then compared to the sticker distribution (Fig. 18.1a) by the authors (RvdS, EG) to assess the face validity of the elicitation. In the case of Fig. 18.1, the parametric distribution nicely represents the stickered distribution.

18.4 Seven Elicitation Issues with Seven Solutions

The following section describes how the ideal method was adjusted to fit the 'problematic' sticker distributions. We refer to Fig. 18.2 for example distributions. We developed seven solutions for seven issues.

18.4.1 Issue 1: The Expert Includes Numbers/Percentages with Their Distribution

The numbers experts include on the x-axis often do not correspond to the actual interval at that point of the x-axis. In these cases, the numbers included by the expert are leading. This means that the part of the x-axis that the expert used for their distribution needs to be rescaled. Some examples are

1. A start and end percentage is included. The distance between these two points is then equal to the difference in percentages that are noted by the expert. This distance is used to compute the intervals.
2. There is only a percentage/number at the centre of the stickered distribution. The distance between the 0% point and this number is then equal to the percentage at the centre (this could be any percentage, not necessarily 50%). The rest of the x-axis is rescaled according to this distance and percentage.

3. There is a start, end and centre percentage of the stickered distribution (Fig. 18.2—scenario 1). There are now two intervals (left to centre, and centre to right). Both have to be rescaled separately based on the distance between the two points. The R-script to reproduce these results is shown below. The first nonzero density is present after ten percent of the x-axis, which is clearly supported by the zero density for the first 100 sections `[rep(0,100)]`. The right edge of the distribution is set at 40%, which is shown by the zero density of the final 600 sections of the x-axis `[rep(0,600)]`, resulting in:

```
c(rep(0,100),rep(10*0.6/43.07,28),rep(0,18),rep(10*2.4/43.07,29)
,rep(0,11),rep(10*4.3/43.07,25),rep(0,15),rep(10*3.4/43.07,30),r
ep(0,11),rep(10*2.3/43.07,26),rep(0,3),rep(10*1.3/43.07,30),rep(
0,7),rep(10*0.5/43.07,30),rep(0,7),rep(10*0.7/43.07,30),rep(0,60
0))
```

18.4.2 Issue 2: Stickers Are Pasted in Neat Vertical Stacks with Different Vertical Distancing

For the situation that stacks are not neatly placed next to each other, we relied on the perpendicular distance between the x-axis and the top of the stack. To compute the proportion of x-axis taken up by the stack, we first used the percentages for the left and right edges. If the stickers also overlap horizontally, we use the highest sticker of the stack to measure the proportion of the stack on the x-axis. After computing all these proportions on the x-axis, and their corresponding heights of the stack, we can compute the total area of the distribution using $\sum(x_2 - x_1)y_1$, where x_1 and x_2 represent the percentages (one decimal) for the left and right edges and y_1 represents the height of the stack in cm (one decimal). For each interval, we computed the percentages of total area, using

$$\frac{100(x_2 - x_1)y_1}{\sum(x_2 - x_1)y_1},$$

where $(x_2 - x_1)y_1$ is the area of the specific interval which is divided by the sum in the demominator (the total surface area). To decide the percentage of total area per 1/1.000th part of the x-axis (the info we need for `SHELF`), we used

$$\frac{100(x_2 - x_1)y_1}{\sum(x_2 - x_1)y_1} \times \frac{1}{10(x_2 - x_1)} = \frac{10y_1}{\sum(x_2 - x_1)y_1}$$

where the 10 in the denominator was added to the second fraction to convert from percentages to 1/1.000th parts. After computing these numbers for every interval (stack of stickers), we created another series of numbers for `SHELF`. For example, the array `rep(10*0.7/50.54,32)` represents a stack with height **0.7**, a total area of **50.54,** and **32** sections.

```
c(rep(0,49),rep(10*0.7/50.54,32),rep(0,134),rep(10*1.6/50.54,32)
,rep(0,16),rep(10*5.2/50.54,33),rep(0,68),rep(10*6.0/50.54,33),r
ep(0,141),rep(10*1.4/50.54,33),rep(0,247),rep(10*0.5/50.54,32),r
ep(0,150))
```

18.4.3 Issue 3: Stickers Are not Stacked Exactly on Top of Each Other

If the stickers were pasted in a disorderly fashion, more like a cloud than neat stacks (see Fig. 18.2—Issue 3), we used the highest sticker at each point of the x-axis to compute the input for `SHELF`. So, instead of basing the stacks edged on the x-axis on the lowest sticker, you use the highest sticker to find x_1 and x_2 for each interval and apply the same approach as used in Issue 2.

18.4.4 Issue 4: The Distribution Lacks Stickers in a Specific Part of the Parameter Space

When the expert sketches a distribution he or she occasionally fails to fill it all up with stickers, see Fig. 18.2—Issue 4. If we had applied our default strategy, the distribution would have been bumpy. To solve this, we relied on linear interpolation and we added the minimum number of points needed to fill out the distribution. Any newly added point was added in the centre of its two surrounding points (sticker stacks), both with regard to the x- and y-axis. Each point simulates a sticker and is thus of the same width and height. Each point is thus equal to 3.1% of the x-axis. To find the location of this new point or stack, we added 1.55% to the left and right of the point and noted the location on the x-axis. For example in Fig. 18.2 two stickers were added between the stacks with height y of **9.4** and **8.3**. First, 8.85 is the midpoint of **9.4** and **8.3**. The inserted y of **9.1** is the midpoint of **9.4** and 8.85; the inserted y of **8.6** is the midpoint of **8.3** and 8.85.

```
c(rep(0,47),rep(10*9.4/149.67,31),rep(0,6),rep(10*9.1/149.67,31)
,rep(0,11),rep(10*8.6/149.67,31),rep(0,6),rep(10*8.3/149.67,31),
rep(0,4),rep(10*6.7/149.67,31),rep(0,19),rep(10*4.1/149.67,31),r
ep(10*1.5/149.67,12),rep(0,11),rep(10*1.5/149.67,31),rep(0,667))
```

18.4.5 Issue 5: All Stickers at One Point

This is a special case since the expert indicated to be 100% certain that the number of Ph.D. candidates answering ‘yes’ to the question was a specific value, typically zero. During the elicitation process, it was explained to these experts what the consequences were if they still decided to go for this particular answer. It appeared some experts truly believed that 0% of the Ph.D. candidates would never ever agree with publishing the results if they did not trust the data. If this was truly their answer we had to apply a trick because SHELF does not allow to add only one value. So, we used two intervals instead and put 99% of the density mass from 0.0 to 0.1%, and only 1% to the 0.1–0.2 interval:

```
c(rep(99,1),rep(1,1),rep(0,998))
```

18.4.6 Issue 6: Stickers Fall (Partially) Outside of the X-Axis

Pasting stickers outside the limits of the x-axis can take on two forms:

1. There is a stack at the limit, and some of these stickers go beyond the limit of the x-axis. If the expert was clear in saying that this stack should represent 0 or 100%, we put all stickers in the stack in the first (0.0–0.1) or last (99.9–100.0) interval of the x-axis. If they were not clear, then the interval was decided by the edge of the sticker that is just inside the parameter space (0 or 100%), and is the entire density added to the first or last interval.
2. There is only one sticker, or a small part of the stickered distribution indicated by the stickers outside the limit of the x-axis. In this situation, we decided to lengthen the x-axis to fit this sticker. This meant that any computations using the length of the x-axis need to be adjusted to a new length (instead of 25.8 cm).

18.4.7 Issue 7: One Sticker Is an Outlier

In some cases, experts added some ‘outlier’ stickers to their distribution that SHELF cannot handle. In this case, we made the interval on the x-axis wider, while at the same time making these ‘outlier’ stickers ‘flatter’, so that the total percentage of the distribution accounted for by these stickers did not change. That is, the density of 2 stickers is reduced to 2*5/800 and redistributed over 800 steps. This leads to the following R-code:

```
c(rep(2*5/800,800),rep(10*5/31,31),rep(0,45),rep(8*
5/31,31),rep(0,93))
```

18.5 Results

18.5.1 *Perfect and Imperfect Elicitation Results*

First, we applied the seven strategies to the results of the 36 (experts) * 3 (scenarios) = 108 stickered distributions. One expert misunderstood the instructions about how to interpret the x-axis (as only appeared after the interview was done) and none of the three stickered distributions could be used, see for example Fig. 18.3a. For one other stickered distribution it was completely unclear what the intended distribution should look like, see Fig. 18.3b. And for another distribution, see Fig. 18.3c, a beta distribution could not be applied because of the bi-modal shape. So, for 103 stickered distributions the seven solutions could be applied. For one expert, see Fig. 18.4, we decided that the obtained distribution did not fit the stickered distribution and this result was therefore also omitted from future analyses, resulting in a total of 102 distributions.

Out of the 102 stickered distributions, only 23 (22.5%) could be entered in the software SHELF without any adjustments. This implies that most of the elicitation results suffered from one or more of the issues described above and would have been useless without adjustments. Table 18.1 provides an overview of the prevalence of each of the seven issues.

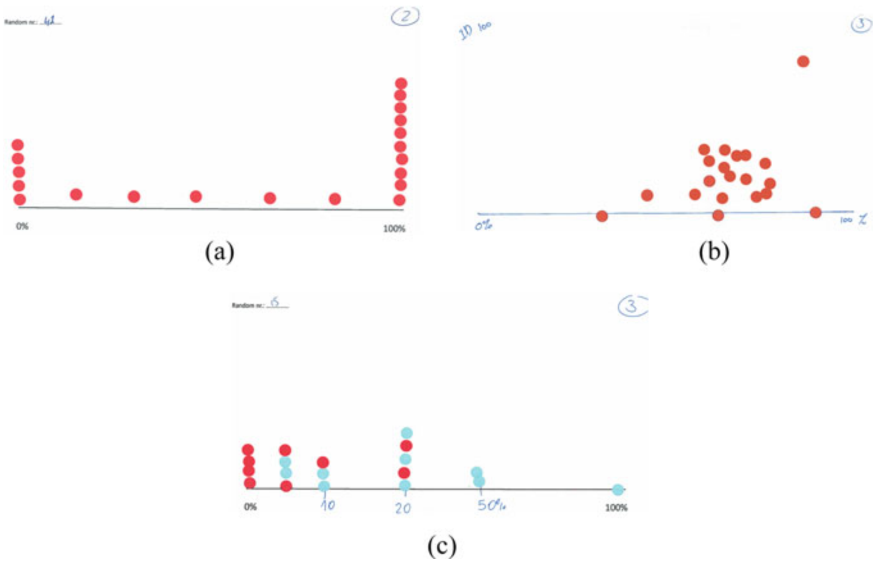
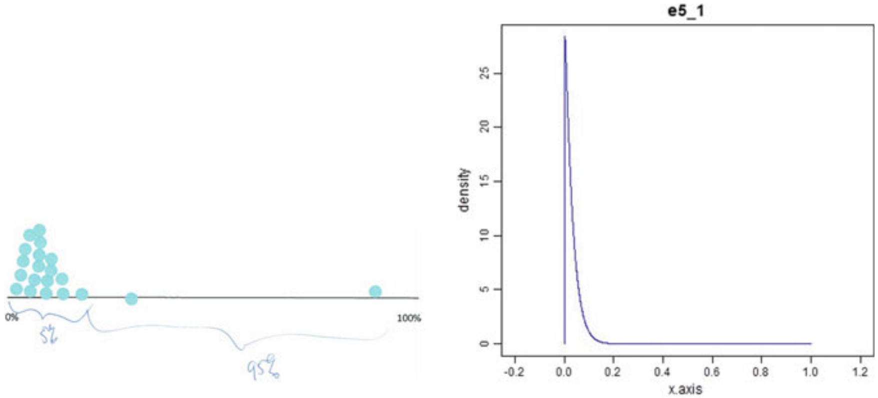


Fig. 18.3 Examples of stickered distributions we omitted: **a** the expert misunderstood the x-axis (n = 3) and (**b** and **c**) the expert placed the stickers in a non-identifiable distribution



(a) (b)

Fig. 18.4 For one expert we decided the parametric distribution (**b**) did not resemble the stickered distribution (**a**)

Table 18.1 Overview of the prevalence of the perfect results, the seven issues and possible combinations for each of the three scenarios with a total of 79 problematic distributions with 109 issues and 23 perfect distributions

	Scenario 1 (n = 42)	Scenario 2 (n = 47)	Scenario 3 (n = 43)
Perfect results	6	9	8
Issue 1	4	6	7
Issue 2	4	5	4
Issue 3	9	15	15
Issue 4	0	1	1
Issue 5	8	3	2
Issue 6	10	7	5
Issue 7	1	1	1
Of which are combinations	7	9	9

18.5.2 Mixture Distributions

As a second step, to summarize the parametric distributions, we merged the individual distributions of all experts into three clusters:

1. Experts who indicated to be certain that zero percent of the Ph.D. candidates would publish the results (almost all density put on zero percentage; blue line);
2. Experts who believed the percentage to be low, but not exactly zero (zero had to be in the 95% density mass; orange line);
3. Experts who believed the percentage to be clearly higher than zero (zero fell outside the 95% density mass; yellow line).

To do so, we applied the following procedure for scenario 1 and the group of experts who expected exactly zero percent. We first constructed a data frame:

```
sc1_null <- data.frame(exp3[[1]], exp28[[1]], exp71[[1]],
exp62[[1]],exp350[[1]], exp532[[1]], exp2807[[1]],
exp35792[[1]])
```

Then, we assigned an equal weight-value to all expert priors [1/total number of experts]:

```
sc1_g1_W <- rep(1/length(sc1_null),length(sc1_null))
```

Next, we created a new density distribution (Y1g1) and looped over all the experts:

```
Y1g1=rep(0,length(x))
for (e in 1:length(sc1_null))
{
 y = sc1_null[,e]
 Y1g1=Y1g1+y*sc1_g1_W[e]
}
```

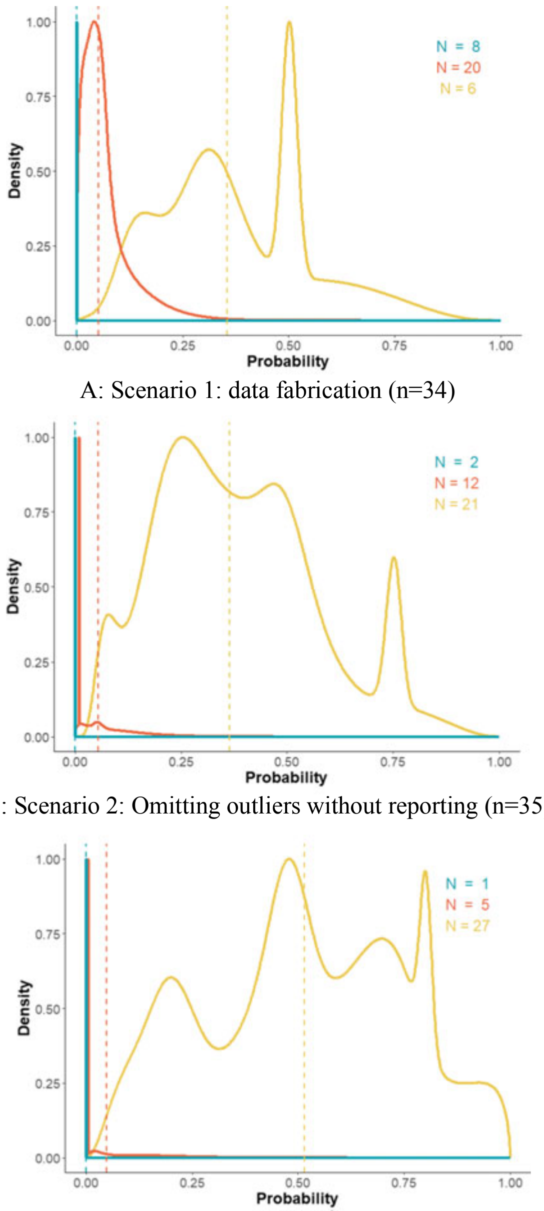
This procedure was repeated for all other expert-groups and for each scenario separately. The complete syntax is available on the OSF (https://osf.io/bq28j/).

18.5.2.1 Mixture Results—Scenario 1

Eight seniors indicated to be hundred percent sure not one single Ph.D. candidate would agree to publishing the results in the situation as described in Scenario 1 (i.e. data fabrication), see Fig. 18.5a. In addition, 20 seniors indicated the percentage of Ph.D. candidates to be close to zero, but not exactly zero. The majority of these second groups' combined probability mass is well below 20%. Combining these two groups, this shows that 82% of the seniors believed the percentage of Ph.D. candidates willing to publish a paper, even if they did not trust the data because of potential data fabrication, to be zero or close to zero. A third group of six seniors (18%) believed the percentage to be higher than zero, but they vary widely in their beliefs between, roughly, between 5 and 75%.

18.5.2.2 Mixture Results—Scenario 2

There are only two seniors who were very sure zero percent of the Ph.D. candidates would publish the results in the second scenario (i.e. deleting outliers to obtain a significant effect) and another 12 seniors believing the percentage to be very close to zero. These two groups represent a total of 40% of the seniors, a much lower



A: Scenario 1: data fabrication (n=34)

B: Scenario 2: Omitting outliers without reporting (n=35)

C: Scenario 3: Salami-slicing (n=33)

Fig. 18.5 Combined results of all parametric distributions split into three categories: (1) Experts who indicated to be certain zero percent of the Ph.D. candidates would publish the results (blue); (2) Experts who believed the percentage to be low, but not exactly zero (orange); (3) Experts who believed the percentage to be clearly higher than zero (yellow)

percentage compared to the first scenario; see also Fig. 18.5b. The remaining 22 seniors (60%), who believed the percentage to be larger than zero, disagreed even more than on the first scenario and provided distributions covering the entire range up until, roughly, 95%.

18.5.2.3 Mixture Results—Scenario 3

Only one senior indicated the percentage of Ph.D. candidates willing to publish a paper in scenario 3 (i.e. splitting results from one study across multiple publications) to be exactly zero, see also Fig. 18.5c. Another five indicated their belief to be close to zero, but with much more variability (i.e. larger variance of the combined distribution) than in the previous two scenarios. Most of the seniors ($n = 27$; 82%) believed the percentage to be much higher, and some even close to 100%.

18.6 Conclusion—Empirical Data

In general, the seniors believed the Ph.D. candidates are very likely to 'salami-slice' their papers (82%) or to delete outliers (60%) and some even believe they are likely to go ahead with fabricated data (18%). Based on these distributions the senior administrators seem to believe that the acceptance of serious misconduct is relatively low amongst Ph.D. candidates when compared to questionable research practices such as deleting outliers without a proper reason and, especially, salami tactics, which are believed to be quite common. Even so, some seniors believe that Ph.D. candidates, if feeling sufficiently pressured, would go ahead and publish even with fake data.

18.7 Conclusion—Elicitation Procedure

Ideally, an elicitation procedure should be properly prepared by allowing for enough time to train the experts, provide them with feedback, etc. (see, e.g. Johnson et al. 2010; Zondervan-Zwijnenburg et al. 2017). However, usually time constraints make it difficult or sometimes even impossible to obtain 'perfect' elicitation results which can directly be entered in elicitation software like SHELF (Oakley and O'Hagan 2010). It would be a pity if results from such an elicitation procedure had to be discarded. Moreover, the experts, at least in our empirical example, had a clear idea of how the distribution should have looked like, but simply lacked the time or skills for the correct placement of the stickers. In our chapter, we provided seven different issues with 'imperfect' elicitation results, and we provided solutions for translating these results into empirical distributions reflecting the original stickered distributions.

Another way of obtaining 'perfect' results when time for the elicitation process is extremely limited, is to use digital procedures for the trial-and-roulette methods. Veen et al. (2017b) developed a five-step method for the trial-and-roulette method which can be used on a mobile device. Lek and Van De Schoot (2018) developed another app for mobile devices in the context of educational testing for the elicitation of a beta distribution for primary school teachers. In both procedures, a direct feedback step is included in which the expert can approve the translation of their stickered distribution into an empirical distribution. Such new developments are to be preferred when compared with the solutions we presented in our chapter. On the other hand, many experts indicated the placing of stickers was fun and inspiring. Filling out the online apps would 'just' be another task at a computer, of which they already have too many.

Acknowledgments This work was supported by Grant NWO-VIDI-452-14-006 from the Netherlands organization for scientific research.

References

Clemen, R. T., Fischer, G. W., & Winkler, R. L. (2000). Assessing dependence: Some experimental results. *Management Science, 46*(8), 1100–1115.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS One, 4*(5), e5738.

Goldstein, D. G., Johnson, E. J., & Sharpe, W. F. (2008). Choosing outcomes versus choosing products: Consumer-focused retirement investment advice. *Journal of Consumer Research, 35*(3), 440–456.

Goldstein, D. G., & Rothschild, D. (2014). Lay understanding of probability distributions. *Judgment and Decision Making, 9*(1), 1.

Gore, S. (1987). Biostatistics and the medical research council. *Medical Research Council News, 35,* 19–20.

Haran, U., & Moore, D. A. (2014). A better way to forecast. *California Management Review, 57*(1), 5–15.

Haran, U., Moore, D. A., & Morewedge, C. K. (2010). A simple remedy for overprecision in judgment. *Judgment and Decision Making, 5*(7), 467.

Hofmann, B., Myhr, A. I., & Holm, S. (2013). Scientific dishonesty—a nationwide survey of doctoral students in Norway. *BMC medical ethics, 14*(1), 3.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524–532.

Johnson, S. R., Tomlinson, G. A., Hawker, G. A., Granton, J. T., Grosbein, H. A., & Feldman, B. M. (2010). A valid and reliable belief elicitation method for Bayesian priors. *Journal of Clinical Epidemiology, 63*(4), 370–383.

Lek, K., & Van De Schoot, R. (2018). Development and evaluation of a digital expert elicitation method aimed at fostering elementary school teachers' diagnostic competence. *Frontiers in Education, 3,* 82.

Martinson, B. C., Anderson, M. S., & De Vries, R. (2005). Scientists behaving badly. *Nature, 435*(7043), 737.

Oakley J, O'Hagan A (2010) SHELF: The sheffield elicitation framework (version 2.0). Sheffield, UK: School of Mathematics and Statistics, University of Sheffield.

Sonneveld, H., Yerkes, M. A., & Van de Schoot, R. (2010). *Ph.D. Trajectories and labour market mobility: A survey of recent doctoral recipients at four universities in The Netherlands*. Utrecht: Nederlands Centrum voor de Promotieopleiding/IVLOS.

Steneck, N. H. (2006). Fostering integrity in research: Definitions, current knowledge, and future directions. *Science and Engineering Ethics, 12*(1), 53–74.

Tijdink, J. K., Verbeke, R., & Smulders, Y. M. (2014). Publication pressure and scientific misconduct in medical scientists. *Journal of Empirical Research on Human Research Ethics, 9*(5), 64–71.

Van de Schoot, R., Yerkes, M. A., Mouw, J. M., & Sonneveld, H. (2013). What took them so long? Explaining Ph.D. delays among doctoral candidates. *PLoS One, 8*(7), e68839.

Van de Schoot, R., Yerkes, M. A., & Sonneveld, H. (2012). The employment status of doctoral recipients: an exploratory study in the Netherlands. *International Journal of Doctoral Studies, 7,* 331.

Veen, D., Stoel, D., Schalken, N., & van de Schoot, R. (2017a). Using the data agreement criterion to rank experts' beliefs. arXiv:170903736.

Veen, D., Stoel, D., Zondervan-Zwijnenburg, M., & van de Schoot R (2017b) Proposal for a five-step method to elicit expert judgement. *Frontiers in Psychology 8*, 2110.

Zondervan-Zwijnenburg, M., van de Schoot-Hubeek, W., Lek, K., Hoijtink, H., & van de Schoot, R. (2017). Application and evaluation of an expert judgment elicitation procedure for correlations. *Frontiers in psychology, 8,* 90.

Chapter 19
Structured Expert Judgement for Decisions on Medicines Policy and Management

Patricia Vella Bonanno, Alec Morton, and Brian Godman

Abstract Many decisions related to the marketing authorisation of medicinal products as well as decisions for processes such as Health Technology Assessment (HTA), reimbursement and pricing of medicines, and the setting of clinical guidelines, are taken in the face of significant uncertainties. Moreover, decision-making can be impacted by biases resulting from psychological heuristics. In other domains where decisions have to be taken with imperfect or incomplete evidence, Structured Expert Judgement (SEJ) has been found to be useful in making the best use of available evidence, and synthesising it with professional expertise, stakeholders' values and concerns. To date, formal SEJ has only been used to a limited extent in healthcare. Aspects affecting decisions for marketing authorisation and health technology assessment, reimbursement and pricing of medicines are described and the main risks and uncertainties are identified. Some considerations and recommendations for the use of SEJ to strengthen these decisions are made.

19.1 Background

> We look for medicine to be an orderly field of knowledge and procedure. But it is not. It is an imperfect science, an enterprise of constantly changing knowledge, uncertain information,

P. Vella Bonanno (✉) · B. Godman
Department of Pharmacoepidemiology, Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, Glasgow G4 0GE, UK
e-mail: patricia.vella-bonanno@strath.ac.uk

A. Morton
Department of Management Studies, Strathclyde Business School, University of Strathclyde, Glasgow G4 0GE, UK

B. Godman
Division of Public Health Pharmacy and Management, School of Pharmacy, Faculty of Health Sciences, Sefako Makgatho Health Sciences University, Pretoria, South Africa

Division of Clinical Pharmacology, Karolinska Institute, Karolinska University Hospital Huddinge, Stockholm, Sweden

A. M. Hanea et al. (eds.), *Expert Judgement in Risk and Decision Analysis*,
International Series in Operations Research & Management Science 293,
https://doi.org/10.1007/978-3-030-46474-5_19

> fallible individuals, and at the same time lives on the line. There is science in what we do, yes, but also habit, intuition, and sometimes plain old guessing.
>
> AtulGawande, *Complications: A Surgeon's Notes on an Imperfect Science*

Although it is considered important that decisions related to medicinal products and their use are optimal, in real life the situation differs. In spite of the use of evidence-based practice, uncertainties remain. Moreover, as in all human decision-making, the decision processes are subject to psychological biases. There are also other factors affecting the decisions taken such as advertising and other influences.

Our objective is to present the risks and the challenges arising from gaps in knowledge and uncertainty that exist in the decisions for medicines marketing authorisation, health technology assessment, reimbursement and pricing. We argue for the use of Structured Expert Judgement (SEJ), also known as expert knowledge elicitation, as a tool to fill these gaps in knowledge and strengthen the decision-making processes.

Our experiences show that there can be synergy from collaboration between experts in SEJ and domain experts, specifically in the healthcare arena. The expected audience comprises both experts in the field of SEJ and practitioners in healthcare, particularly from the pharmaceutical area. We lead with Sect. 19.2, which presents the application of SEJ in healthcare, mainly based on the literature and the experience of the authors. From Sect. 19.3 onwards, we focus on the area of medicines and start by describing a case study which illustrates the problems of pharmaceutical policy and management, and makes clear the importance and pervasiveness of scientific uncertainty in this domain. In Sect. 19.4, we describe the pharmaceutical policy framework and its regulatory risk governance structure, and present this complex system emphasising the key areas for decision-making. In Sect. 19.5, we describe these key decisions, focusing on marketing authorisation and Health Technology Assessment (HTA), reimbursement and pricing. In Sect. 19.6, we present some challenges for decision-making in this area. Section 19.7 presents the case for the application of SEJ to processes of the pharmaceutical policy framework, including considerations to support the introduction and implementation of SEJ in this area. Section 19.8 concludes and summarises our recommendations.

19.2 The Application of Structured Expert Judgement in Healthcare

As seen in different chapters in this book, one method to fill the gaps in evidence and to counter biases in judgement is through practitioners' professional expertise in the form of SEJ. A number of areas in which uncertainty looms large such as vulcanology, natural disasters and risk management have experience with the use of SEJ (Barends et al. 2014; Bedford and Cooke 2001; Cooke 1991; Garthwaite et al. 2005). Formal methods to elicit SEJ can improve the accountability and the transparency of the decision-making process and can reduce bias and the application of heuristics (Soares and Bojke 2018).

19.2.1 SEJ for Healthcare Decision-Making

Soares and Bojke (2018) show that to-date formal SEJ has only been used to a limited extent in healthcare decision-making and they mainly link this to the lack of clear guidance on what methodologies may be appropriate for the purpose. They explain that there are a number of features of healthcare decision-making that distinguish it from other disciplines. Currently available guidelines and protocols for SEJ need to be subject to further considerations, particularly where such protocols describe multiple options for particular elements of the design process. Decision-making in healthcare is increasingly becoming explicit, accountable, evidence-based and focused on an explicit normative framework based on the maximisation of aggregate health. Although there are elicitation protocols proposed in the literature, there has not yet been consideration of protocols or elements of protocols that are appropriate for healthcare decision-making.

The experience of the authors of this chapter supports the observation by Soares and Bojke (2018). It is difficult for practitioners in the field of health to understand the methodologies and master them. Consequently, it is important that practitioners in healthcare decision-making collaborate with academic experts in the field of SEJ to develop and implement the methodology which suits the purpose. The authors consider that there could be other reasons for this lack of use of SEJ in health beyond those stressed by Soares and Bojke. Healthcare prides itself as the flagship of evidence-based practice and that all decisions are based on high quality evidence. Expert opinion is considered to be at the bottom of the 'hierarchy of evidence', and therefore there can be a lack of motivation for use of SEJ.

Nevertheless, the demand for transparency of decisions and of the decision process, as well as the growth of public interest and advocacy organisations, have resulted in demands for higher levels of accountability from decision makers (Baba and HekemZadeh 2012). Some areas of healthcare have been slow in developing this culture. Where there is lack of high quality evidence, rather than admitting that there is a gap in knowledge practitioners tend to hide the situation and there is lack of transparency of such cases. One reason could be that practitioners in the field can be too proud to admit that there are gaps in knowledge. Healthcare is an area with a traditionally high paternalistic culture. Another reason for this lack of transparency could be the concerns from litigation.

Notwithstanding these barriers, there is an increasing call for including patients, representatives of patient organisations, healthcare professionals and industry representatives as experts in committees involved in decision-making (European Commission European Commission 2018a, b) and SEJ can support this. A number of decisions are taken by committees and SEJ can be used to bring about a structured decision-making framework for committees and thus improve these decisions.

19.2.2 Supporting the Implementation of SEJ in Healthcare

Soares and Bojke (2018) have made a number of suggestions for the implementation of SEJ in health and make recommendations for the most appropriate use of available resources for implementation of SEJ. SEJ requires resources and time. Moreover, audit is important to consider which elements of formal elicitation are necessary requirements for decision-making in healthcare, and which elements make a more marginal contribution. In practice, there are usually time constraints for carrying out of processes and taking decisions, and it is important to consider how formal SEJ would work alongside the different processes. In some circumstances a 'gold-plated' SEJ may not be achievable in order for decisions to be made in a timely manner. It is recommended to consider which available software facilitates the process, or alternative software which needs to be developed. Other areas have managed to combine the expertise of experts in SEJ and expertise from the technical area to address the specific needs and healthcare should learn from these experiences.

It is evident that the introduction of SEJ requires a change in culture and more discipline for the decision-making process in healthcare. It is important that experts involved in decision-making are coached about the possibility of biases. For feedback and de-biasing to work, decision makers should be convinced that their own judgements are just as vulnerable to biases as others and ignore their strong intuition that they are not biased. People may have a 'bias blind spot' whereby they consider that they are less biased than others in the same circumstances. People afflicted with this bias blind spot are more likely to ignore the advice of peers or experts and are least likely to learn from de-biasing training and de-biasing strategies. De-biasing training involves teaching the decision maker different thinking strategies to help improve critical thinking through education as well as by providing formal decision aids which support better thinking. There could also be de-biasing through modification of the environment by altering the setting and the choice options where decisions are made.

Decision analytic and SEJ techniques are useful to avoid bias and to structure decision-making. Decision analytic and SEJ techniques can also distinguish between errors of ignorance (mistakes made because there is not enough knowledge) and errors of inaptitude (mistakes made because there is not proper use of what is known). These can include checklists, protocols and rubrics derived from scientific evidence and provide guidance for information gathering and for action under specific circumstances to reduce bias and improve quality of decisions (Rousseau 2012; Rousseau and Gunia 2015). Another method recommended for preventing cognitive biases from affecting decisions is to draw up an algorithm for a decision in advance and to apply it consistently. This simplifies the decision-making process without compromising its quality and enables the decision maker to avoid potential cognitive bias problems (Otuteye and Siddiquee 2015).

Inclusion of diversity within the decision-making team is considered to reduce biases by bringing forward real alternatives. Working within a group where there is diversity helps to anticipate alternative viewpoints and expect that reaching consensus will take effort (Pfeffer and Sutton 2011).

19.3 The Vioxx Case Study

The case study of Vioxx helps demonstrate the substantial risks involved with medicines and the critical role of uncertainty in decision-making in pharmaceutical policy and management.

Vioxx (rofecoxib), a COX-2 selective Non-Steroidal Anti-Inflammatory Drug (NSAID), was launched on the market by Merck in 1999. This was a blockbuster drug which at the height of its use in 2003 represented 10.3% of all NSAID prescriptions in the UK. NSAIDs are linked with adverse gastro-intestinal effects and when the selective COX-2 inhibitors emerged these were claimed to have fewer (or no) gastro-intestinal effects, which was considered as a major advance for patients with arthritis. However, the medical community was not aware of the cardiovascular risks associated with COX-2 inhibitors. The VIGOR trial had a four times higher risk of myocardial infarction than comparators. Regulators strengthened precautions to reflect this safety information. Merck was under pressure and in 2002 they placed a warning on the product. A second clinical trial, APPROVe, showed a two-fold increase in risk of adverse cardiovascular events compared to placebo. As a result of this study, in 2004 the evidence became very strong, the trial was stopped early and Vioxx was voluntarily withdrawn from the market by Merck (O'Connor 2005; Sukkar 2014).

During the Vioxx case, the regulators and the regulatory systems were deeply scrutinised and criticised for being 'too cosy' with the drug industry, neglecting their obligations as a regulator. Merck was blamed for deceiving the medical community, the regulators, consumers and its own researchers into believing that Vioxx was safe and allegedly trained sales representatives to avoid questions about the cardiovascular risks of Vioxx. The safety issues were not highlighted to the regulator; nor were consumers warned of the cardiovascular risks in its advertising. Merck insisted that it was only obliged to warn doctors (Pritts 2006). It was noted that Merck's promotional materials and activities minimised the potentially serious cardiovascular findings that were observed in the VIGOR study and omitted crucial risk information associated with Vioxx, contained unsubstantiated comparative claims and promoted unapproved uses. Eventually the FDA objected to Vioxx's promotional materials, considering them to be unsubstantiated (O'Connor 2005).

Merck eventually added a warning about cardiovascular risks in June 2001, 14 months after receiving the results of the VIGOR study (Pritts 2006). By the time of withdrawal from the market, 80,000 people had taken Vioxx. Following the withdrawal of Vioxx, the prescriptions of the other NSAIDs which remained on the market were also highly reduced (Sukkar 2014). In 2014 Merck was still fighting a number of lawsuits world-wide. Vioxx earned Merck approximately US$ 11 billion in revenue during its marketing and up to 2014 reportedly cost US$ 6 billion in litigation, one of the largest settlements being with the US government in 2011 for US$ 950 million (Sukkar 2014). In 2012, the United States Attorney in Massachusetts imposed a fine of nearly US$ 322 million for illegal promotional and marketing activity related to off-label marketing of Vioxx and false statements about the drug's cardiovascular safety (FDA Office of Criminal Investigations 2012). The Vioxx scandal caused considerable harm to patients and their families, misled prescribers, fomented mistrust with the system of medicines regulation and with the regulatory institutions, brought medicines marketing into disrepute and damaged the company. The financial and managerial aspects related to the case are also important. Like many block-buster drugs, the research and development of Vioxx was funded in large part by the investment of private capital and the financial fallout of this drug resulted in lessons for designing future investment strategies.

19.4 The Policy and Regulatory Landscape for Pharmaceuticals

19.4.1 The Pharmaceutical Policy Framework

To understand the context in which the Vioxx story unfolded, we present a logic model of the Pharmaceutical Policy Framework of the European Union (see Fig. 19.1). This is an adapted and updated version of the earlier logic model by Vella Bonanno (2003, 2010).

The logic model supports systematic representation and evaluation of this complex system. The elements of a logic model (resources, activities, stakeholders [customers reached], outputs and outcomes) and the logical linkages among them support the description and evaluation of the pharmaceutical framework. The main processes within this framework concern research and development, marketing authorisation and pharmacovigilance, pricing and reimbursement, manufacture and the supply chain, prescribing and dispensing as well as administration and monitoring of medicines in clinical care. The processes involve the relevant resources (structural, legislative and policy as well as institutions), the activities, the outputs (with the main challenge being the balance between public health and competitiveness) and the different stakeholders (policymakers, health care professionals, patients and the industry). There are different outcomes—those related to the quality, efficacy and safety of medicinal products and those related to access, availability, affordability and use of medicinal products. The final outcomes of the framework depend on the outcomes of the individual processes and also on the logical flow from one process to another. The framework also highlights the contextual factors external to the programme and not under its control ('external influences') that could influence its success either positively or negatively.

The processes of the pharmaceutical framework involve different activities and a number of decision processes. Different stakeholders (who are also experts for the

Resources		Activities	Stakeholders (customers reached)	Outputs	Outcomes
Legislation & policy	Structural & human				
European legislation European policy National Legislation National policy	Academia Investigators Pharma industry	Research &Development, clinical trials Application for Marketing Authorisation	Regulators Policy makers Pricing and reimbursement authorities Pharma industry Healthcare professionals Patients Health Service providers Payers	Risk governance Free movement of goods Innovation Intellectual Property Protection Public health Allocation of resources Equity Sustainability of healthcare systems	Unmet medical needs Quality, safety, efficacy Affordability Access & availability Rational use
	Regulatory agencies: EMA National agencies	Scientific advice to industry pre-authorisation Evaluation for marketing authorisation Post-authorisation monitoring			
	Health Technology Assessment (HTA) bodies Reimbursement agencies	HTA advice to industry pre-authorisation Horizon scanning Health technology assessment Reimbursement decisions Monitoring of medicines use			
	Pricing authorities	Setting of price			
	Operators of the supply chain	Manufacture, distribution, storage			
	Payers Drug & Therapeutics Committees Procurement agencies	Procurement, negotiation, managed entry agreements Monitoring of treatment			
	Health care professionals	Prescribing, dispensing, administration, monitoring of treatment			
	Patients and carers	Access to medicines, Administration, Monitoring of treatment			
External Influences					
The Treaties; The European Commission, the Council and the European Parliament; Regulatory governance framework set by the pharmaceutical legislation; Associations/groupings for stakeholders; Pricing and Reimbursement bodies; Pressures by / on different stakeholders; Research and funding initiatives					

Fig. 19.1 General logic model representing the EU pharmaceutical policy framework (March 2018). Updated from Vella Bonanno (2003, 2010)

specific systems) play an important role in decision-making. Most stakeholders are represented through groupings and associations (e.g. industry and trade organisations, healthcare and medical professional bodies and patient groups and networks). These are generally non-governmental organisations. The stakeholder associations serve to protect the interests of group and to coordinate positions. In recent years, groups of countries have formed networks for joint HTA, joint negotiation and possibly joint procurement. Such collaborations include the BENELUXA, the Valletta Declaration, the Baltic Collaboration and Visigrad (European Observatory on Health Systems and Policies 2017).

A key objective for the pharmaceutical policy framework is a high level of protection of human health and the improvement of public health through the use of medicines. The Treaties of the European Union set the mandate for the legislation covering these areas. They stress a high level of human health protection in the definition and implementation of all Union policies and activities (Council of the EU 2016, 2017). Clinical trials are governed through Regulation (EU) No. 536/2014, aimed at reducing risks and protecting clinical trial subjects. Regulation 726/2006 covers marketing authorisation. In 2006, this Regulation was updated to allow for early access of new centrally authorised medicinal products including conditional approval (Article 14 (7)), on the basis of less complete data than is required for a normal submission, granted 'subject to certain specific obligations' of fulfilment of the data required (Regulation (EC) No 507/2006). Once marketing authorisation is granted, medicinal products are followed through post-authorisation pharmacovigilance activities (Regulation (EU) No 1235/2010). Pricing and reimbursement are regulated by the 'Transparency Directive' (Directive 89/105/EEC 1988) and are of national competence and jurisdiction.

19.5 Key Decisions in Pharmaceutical Policy and Management

The above description of the pharmaceutical policy framework sets the stage for the detailed discussion of the key decisions in pharmaceutical policy and management. As shown, there are three main areas for decision-making: marketing authorisation decisions; HTA, reimbursement and pricing decisions and clinical decisions. The three areas impact each other. For the purpose of the remainder of this chapter the focus will be on the first two areas.

19.5.1 Marketing Authorisation and Post-authorisation Activities

European legislation demands that a marketing authorisation gives the reassurance that a medicinal product has proven efficacy and safety, shows a positive benefit/risk, based on high quality evidence and address the risks and governance.

The European Medicines Agency is responsible for the marketing authorisation evaluation and the post-authorisation monitoring, and the European Commission is legally responsible for the marketing authorisations of medicinal products authorised through Regulation (EC) No 726/2004. The marketing authorisation process involves the evaluation of data generated from research and development. This evaluation is based on the assessments by two experts (the rapporteur and the co-rapporteur) and feedback from the rest of the committee while the final decision for recommendation is undertaken by voting by a committee of experts (the Committee for Human Medicinal Products—CHMP).

The investors in new medicinal products face high risks with a high attrition rate and many products do not make it through to marketing authorisation. The marketing authorisation holder of the medicinal product is responsible for monitoring the safety of the product. Health care professionals and patients should report adverse drug reactions to support the pharmacovigilance system. Decisions based on pharmacovigilance are taken by a committee of experts within the EMA (the Pharmacovigilance Risk Assessment Committee—PRAC).

19.5.2 Health Technology Assessment, Reimbursement and Pricing Decisions

A marketing authorisation gives the right for the marketing authorisation holder to have a medicinal product considered for pricing and reimbursement. It is the prerogative of the marketing authorisation holder whether and when to place a product on the market of each EU country. The processes and decisions of pricing and reimbursement are undertaken by the agencies and authorities of each Member State by experts within national committees established for this purpose.

Often, pricing and reimbursement decisions are supported by a form of analysis called 'Health Technology Assessment' or 'HTA'. According to the World Health Organization: 'Health technology assessment (HTA) refers to the systematic evaluation of properties, effects, and/or impacts of health technology. It is a multidisciplinary process to evaluate the social, economic, organisational and ethical issues of a health intervention or health technology' (World Health Organization 2017). The European Network for Health Technology Assessment (EUnetHTA) which was set up and developed to achieve joint health technology assessment, did not achieve the expected cooperation between Member States and there was lack of utilisation of the joint HTA evaluations at national level. On 31 January 2018, the European Commission issued a proposal for new legislation 'Proposal for a Regulation of the European Parliament and of the Council on HTA and amending Directive 2011/24/EU', which suggested a framework for joint HTA for new centrally authorised products (European Commission 2018a).

HTA is based on evidence, and the interpretation of clinical data. It is a multidisciplinary process and involves a systematic evaluation of the properties, effects and impacts of health interventions and technologies. It is designed to enhance decision-making, including reimbursement and pricing decisions. There are different tools to support pricing and reimbursement decisions and their monitoring (Paris and Belloni 2013) and multi-criteria decision analysis have been utilised for analysing the value of medicines where more than one criterion is relevant (Irwin and Peacock 2015; Godman et al. 2016; Soares and Bojke 2018).

There are various challenges during the HTA evaluation. In practice only a limited number of new medicines are seen as innovative (Prescrire 2016). From 2006 to June 2014, twenty-six (26) products were granted a conditional marketing authorisation, and a number of these have not fulfilled post-authorisation obligations (Banzi et al. 2015; Joppi et al. 2016). Conditional approval also resulted in a significant impact on other processes of the pharmaceutical policy framework including problems with the evidence required for HTAs; increased burden, demands and costs for payers; increased monitoring requirements during the post-authorisation phase and possibly reduced safety and effectiveness for new medicines for patients due to uncertainties (Joint Briefing Paper 2015; Garattini and Curto 2016; Davis et al. 2016). This has evolved into Adaptive Pathways to further accelerate the introduction of new medicines; however, there are concerns (Refer to Sect. 19.6.1) (Ermisch et al. 2016; Vella Bonanno et al. 2017).

The increasing prices of new medicines are a major concern for Member States and threaten the sustainability of national healthcare systems. There are different mechanisms for pricing of medicines in Member States (Paris and Belloni 2013; Godman et al. 2016). Each member state undertakes its own pricing negotiations with pharmaceutical companies, with at times a lack of trust between different players (Pharma Diplomacy 2016). Concerns with trust are exacerbated by apparently limited correlation between Research and Development (R&D) costs, the costs of producing medicines, their value and requested prices (Gagnon 2015; Godman et al. 2016; Hill et al. 2016). Pressures are applied on governments to reimburse new medicines in high priority areas such as cancer and orphan diseases with high prices and in often with limited health gain despite high prices (Cohen 2017; Godman et al. 2018; Simoens et al. 2013; Haycox 2016).

The initiatives among payers to support reasonable pricing of medicines are based on their perceived value, and on the principle of rewarding and incentivising innovation. A clear and concerted definition of innovation is required (Aronson et al. 2012; Ward et al. 2014). Different pricing mechanisms are employed and despite extensive application of external reference pricing, countries do pay different prices for medicinal products (Leopold et al. 2012). Concerns with high prices have resulted in the growth of risk-sharing arrangements including confidential discounts, often referred to as managed entry agreements (Godman et al. 2016; Ferrario et al. 2017). Although there are more than ten years of experience with such risk-sharing schemes, there is still limited evidence in support of their effectiveness (Garattini and Curto 2016; Godman et al. 2016). Tools such as Value-Based Pricing (VBP) and the Transparent Value Framework (TVF) were developed in response to concerns about the high prices being requested for new expensive orphan and anti-cancer medicines (European Commission 2012; Godman et al. 2018). The introduction of very expensive new medicines has led to the development of the concept of ‘new payment models’ (European Commission 2018b).

In practice, it can be difficult to delist new medicines on grounds of value rather than potential safety concerns (Godman et al. 2015; Simoens et al. 2013). A number of countries have introduced formal disinvestment procedures to try to address this (Guerra-Junior et al. 2017; Lemos et al. 2018; Parkinson et al. 2015).

19.6 Challenges for Decision-Making

19.6.1 Limitations of the Evidence-Base for Decision-Making

Good quality decision-making should be based on a combination of the best available evidence and critical thinking. In the medical field, adoption of evidence from systematic review and meta-analyses is standard practice, particularly in the evaluation of medical interventions (Centre for Reviews and Dissemination 2008; Tranfield et al. 2003). Systematic reviews allow for large amounts of information to be assimilated quickly by practitioners and academics (Hartling et al. 2014). The 'hierarchy of evidence' which lists a range of study designs ranked in the order of increasing internal validity is used to critically evaluate studies (Petticrew and Roberts 2003; Tranfield et al. 2003).

Although the marketing authorisation process is assumed to adopt evidence-based decision-making and is expected to provide a basis for HTA decisions and for medicines use guidelines, in practice there are major gaps in knowledge which cascade to subsequent processes. Information for marketing authorisation decisions is based mainly on randomised clinical trials. These are typically undertaken in a selected population with limited co-morbidities; however, in clinical practice patients may be of different ages and have additional treatments and have other disease conditions. The introduction of conditional marketing authorisation resulted in a shift in the evidence requirement from the pre-authorisation phase, where the randomised clinical trial is the main methodology for evidence, to the post-authorisation phase, where systematic reviews involving different types of studies may be more appropriate (Ermisch et al. 2016; Vella Bonanno et al. 2017).

Although systematic reviews may constitute high-quality evidence, for many questions about many technologies, systematic reviews are simply not available, or give only limited guidance in the matter under decision. It may be, for example, that: only one or a very small number of clinical trials have been conducted; the trial populations may be unrepresentative of the population for which decisions have to be made (younger, fewer co-morbidities, different ethnic groups); the comparator of the technology in the trials may be different from the comparator which represents the standard of care in the healthcare system about which decisions are to be made; the trial(s) may be underpowered to pick up important side-effects; the treatment modality may differ in the trial population from the population about whom decisions are being made (for example, because compliance cannot be monitored and managed in the general population as it has been in the study population); in the time horizon of the trial does not permit the detection of long term consequences (positive and negative) of interest or the trial population is intrinsically unrepresentative (because they have selected themselves into the study).

Other problems arise not at the level of the individual study, but at the level of the population of all studies on a question of interest. Publication bias refers to a greater likelihood that studies with positive results get published as compared to studies with negative results (Olson et al. 2002). Companies may coordinate or support publications in journals, increasing the risk of publication bias. Scrutiny is recommended for scientific journals, particularly when publishing large trials which involved considerable funding. All of these limitations mean that human interpretation has to be brought to bear on the evidence base.

In spite of the high regulation of medicinal products, there are still causes for concern as reduction of risk depends highly on trust of the different stakeholders and transparency. Quite recently, in the case of dabigatran, the company withheld vital information leading to a number of unnecessary deaths due to excessive bleeding (Cohen 2014).

19.6.2 Limitations of Human Judgement

As described, decisions in the marketing authorisation process as well as in pricing and reimbursement are mainly taken by experts individually or in committees. Such decisions are often taken using crude rules of thumb or ‘heuristics’. Heuristics are defined as ‘a strategy that ignores part of the information, with the goal of making decisions more quickly, frugally and/or accurately than more complex methods’ (Gigerenzer and Gaissmaier 2011). While heuristics are adequate for decision-making in day-to-day life, where they produce ‘good enough decisions’, they can lead to predictable biases, which, in the context of decisions about population health, may lead to squandering considerable sums of money, and substantial unnecessary morbidity and mortality.

As seen above, marketing authorisation evaluation involves comparison of the new drug to alternative treatments and a balance of benefits and risks. During comparison of a medicinal product to alternative treatments, the decoy effect creates a simple relative comparison which makes the object look better not just relative to alternatives but also overall (Ariely 2008, p. 9). Risk perception is generally a cognitive assessment and is therefore susceptible to many biases (Kahneman 2011, p. 252; Simons et al. 1999). Risk perception entails an assessment of the degree of the situational uncertainty, controllability of that uncertainty and the confidence of these estimates. Risk propensity is the general behavioural tendency to take or avoid risk in a specific domain and is affected by factors such as perceived risk, risk attitude and price consciousness (Garling et al. 2009).

Once a decision for the marketing authorisation of a medicinal product is taken, it is difficult to change that decision. Behavioural decision-making literature shows that people show preferences for the status quo over committing to an action that could result in regrettable outcomes (Kahneman and Tversky 1979; Tversky and Kahneman 1991). On the other hand, the status quo is often not satisfying to regulators who are eager to show constituents the impact of legislative initiatives. Moreover, policymakers and regulators may be under pressure from the public to ‘do something’—take visible (even if ineffective) action to signal that a problem is being taken seriously—and uncertainty and distrust increase demand for regulation (Collins and Urban 2014).

Aggressive advertising and publications can create anchoring. Anchoring and adjustment involve sticking to the first piece of information one is exposed to and then adjusting (priming) the estimate upward or downward from the anchor (Epley and Gilovich 2006, Kahneman 2011, p. 122). Even if an initial price is ‘arbitrary’, once this high price is established in people’s minds it will anchor the actual price and also subsequent prices making prices ‘coherent’, leading to biases. A reference price is typically set high because discounts have the important advantage that their subsequent cancellation will elicit less resistance than an increase in posted prices. A temporary surcharge may be unattractive to the seller because it does not have the prospect of becoming a reference price and can only be considered as a loss. The setting of high prices can also be a form of conservatism. Conservatism is demonstrated when once people have formed a probability estimate, they are slow to change the estimate when presented with new information (Kahneman 2011, p. 80).

The involvement of patients in decision-making for marketing authorisation and HTA may have an impact on decisions taken. Some patients will be willing to take medicines while failing to properly take into account high risks of adverse effects. Kahneman and Thaler (1991, 2006) assume that when making a choice at a point in time, the decision-maker makes a forecast of the utility of an outcome that will be experienced at a later time. The evaluations of extended outcomes systematically overweight some parts of the experience and underweight others. These biases result in violations of maximisation of utility (Kahneman and Thaler 2006).

Overconfidence can apply to different stakeholders of the pharmaceutical framework including the regulators, healthcare professionals, patients as well as the industry. Optimism bias and overconfidence can act together in synchronisation. One reason why biases persist is that professionals often do not seek information about the accuracy and the validity of their decisions and they fail to seek feedback, possibly due to overconfidence. This is aggravated by the tendency of people to barrow their focus and not to seek information and feedback to support their decisions (Rousseau and Gunia 2015).

19.7 Application of SEJ for Decisions Related to Medicines

The pharmaceutical policy framework (Fig. 19.1) gives a picture of the complex network of processes involved and descriptive information which empower identification of the areas where formal SEJ can support decision-making related to marketing authorisation and HTA, reimbursement and pricing. Some considerations about the feasibility of the use of SEJ will be made.

19.7.1 Marketing Authorisation Decisions

Although there is a decision process framework for the decision-making committees of the EMA (the CHMP, PRAC and other committees), there is room for improvement and more standardisation and transparency in the operation and work of these committees. SEJ can be considered as a methodology for this improvement.

During marketing authorisation decisions there are guidelines with technical criteria related to the different types of medicinal products. One major shortcoming in the decision-making for the marketing authorisation process for medicines is how to transparently address areas of gaps in knowledge from clinical trials for the assessment of efficacy and safety of medicinal products. Conditional marketing authorisation is granted on the basis of an overall positive benefit/risk balance. The European Food Safety Authority (EFSA) has developed a framework and applied expert knowledge elicitation to address uncertainties in the regulatory field for food products (European Food Safety Authority 2014). Food products may have different considerations compared to medicinal products but regulators of medicinal products can learn from the experience of EFSA.

There needs to be a clear and transparent evaluation of the level and quality of evidence available for the different criteria considered in the evaluation for marketing authorisation, particularly regarding efficacy and safety. At the time of marketing authorisation, these gaps in evidence should be clearly and specifically identified and documented in the marketing authorisation and the resultant information including the Summary of Product Characteristics (SPC). These gaps in knowledge can be filled with the best available evidence available at that time, and SEJ can be used to obtain the best evidence available at that point. The procedures for conditional marketing authorisation should specify the evidence which needs to be presented, and the criteria to be met by the marketing authorisation holder in the post-authorisation phase in order to fill the gaps with evidence-based knowledge and attain an unconditional marketing authorisation. Specific timelines and conditions should be set for the unconditional marketing authorisation.

In the post-authorisation phase, there should be planned and systematic collation of real effectiveness data and other data to fill the gaps in evidence. SEJ can also be used where data is missing, and with time the data from SEJ can be replaced with evidence-based data. The collation of this evidence should follow a clear and robust methodology ideally through the conduct of systematic reviews. There can be a comparison between the data based on SEJ and the real-life data which eventually becomes available, and this can be audited and compared to introduce improvement of the process and of the data.

The marketing authorisation holder is responsible for collating the necessary evidence to support the process of granting of non-conditional marketing authorisation and should give a regular update of the evidence in the post-authorisation phase until all the data required to give an unconditional marketing authorisation is collected. The European Commission, as the granter of the marketing authorisation, is responsible to see that this issue is addressed adequately. The EMA and the scientific expertise provided by the Member States should set guidelines for covering of lack of knowledge in the pre- and post- authorisation phases.

The information from the evaluation and decision for marketing authorisation, including any knowledge filled through SEJ, can be used for the evaluations, decisions and updates which follow the marketing authorisation including the HTA, reimbursement, pricing, procurement and clinical use guidelines.

Although marketing authorisation decisions are taken within a set time-frame (usually 210 days excluding clock stops for regular marketing authorisation procedures, and a shorter period for marketing authorisation procedures for certain medicines such as orphan drugs), the time-table should be set in such a way that there is enough time to allow for SEJ integrated in the decision-making of these procedures.

19.7.2 Reimbursement and Pricing Decisions

Soares and Bojke (2018) described their experience with the use of expert knowledge elicitation to inform HTA. They show that SEJ can provide valuable information in informing decisions utilising HTA, particularly where evidence is missing, where evidence may not be well developed and where evidence is limited. They consider that in HTA, clinical evidence is considered to be of the highest quality if drawn from clinical trials. This may be supplemented by longer term observational studies and real-world data. This evidence needs to be synthesised to allow estimation of total costs and of health benefits associated with competing interventions. Cost-effectiveness analysis often employs decision modelling methods and often involves uncertainty and incomplete evidence. Expert judgement is used to reach decisions in such cases of uncertainty, whereby these judgements are made explicit and incorporated transparently into the decision-making process (Soares and Bojke 2018).

HTA presents particular challenges for SEJ, as the reimbursement and pricing decisions are intrinsically broader in scope than marketing authorisation decisions. Many of the features which drive the cost of a new technology will depend on decisions about the modality of treatment and service use. For example, the cost of introducing a new medicine is only partially driven by the price of the pill: it is also driven for instance by changes in patient length of stay in hospital and the intensity of the follow-up required in outpatient and primary care. At the point of HTA, there will typically be clinical trial data but the clinical trials may give only an imperfect sense how service provision will have to change—particularly as this will depend on the local institutional context, which can be expected to vary widely across Europe. Such concerns are not relevant to market authorisation but may be highly relevant in the context of HTA.

Moreover, reimbursement and pricing decisions are particularly politically charged as millions of Euros may be at stake. In this sort of setting, it is important to be transparent about the process of elicitation of uncertain judgements, and here the existing SEJ literature is clearly a resource to draw upon. But it is also important to define good practice for the selection of experts in the first place, which has been much less of a focus in the SEJ literature to date. Also, in view of the importance of being able to audit the HTA process in order to ensure fair dealing, HTA agencies and the relevant societies should develop good practice guidelines for the practice of SEJ, as they have done for other aspects of the HTA process. This is indispensable both for providing *ex ante* guidance to analysts using SEJ within an HTA framework, and also for enabling *ex post* quality assessment of published studies.

As described above, HTA decisions as well as pricing and reimbursement decisions are taken by committees and therefore as described in Sect. 19.7.1 SEJ can also be used to structure the decision-making process by the committees involved in HTA. In particular as at present there is an initiative to perform joint HTA for all Member States, SEJ can be used to make the methodology of the joint decision-making robust. Having said this, there are serious concerns about a pan-European Joint HTA (Vella Bonanno et al. 2019). A concept which is particular to reimbursement decisions is prioritisation in the allocation of resources, and SEJ can be used to support this challenging decision.

19.8 Conclusion and Recommendations

The legislation for medicines regulation provides for a strong governance framework based on evidence, and one might think that the medicines regulatory process is a robust process which minimises risks particularly related to efficacy and safety of medicinal products. However, there are currently considerable gaps in the evidence available for marketing authorisation decisions, and the decision-making process can be subject to heuristics, biases and other influences.

It is recommended that SEJ is robustly and systematically included as part of the evaluation for marketing authorisation for medicinal products particularly those with conditional approval. First of all, to strengthen the marketing authorisation decision and secondly because this would give the opportunity for the cascade of this information to the subsequent processes including HTA, reimbursement, pricing, procurement and clinical use guidelines.

There is already some experience with the use of SEJ for HTA. It is important that this field is developed further for the clinical evaluation (particularly for products with conditional marketing authorisation) as well as for the economic evaluations. SEJ can have an important role in the development of evidence to support price negotiations, although much work remains to be done.

We see a bright future for the application of SEJ in pharmaceutical policy and management. Of course, this is not a trivial undertaking: the application of SEJ will require the building of robust methodologies and collaboration between experts in SEJ and professionals from the field of medicines and related activities. Nevertheless, we think that not only can SEJ help manage the uncertainty inherent in pharmaceutical policy and management processes, but it can also serve as a framework for the inclusion of patients, healthcare professionals and other stakeholders as part of the decision-making in different parts of the system.

The current regulatory system for pharmaceuticals is a complex system, but it is not static. In the recent past, stakeholders have taken the lead in a number of quality areas in health including the setting of standards for quality and the implementation of quality management systems. It is plausible to believe that stakeholders of the pharmaceutical framework can take the lead to use SEJ to improve their decision-making and implement these changes in a planned manner.

In conclusion, we argue that Structured Expert Judgement has a role to improve decision-making within the pharmaceutical policy framework to increase the benefit, reduce the risks and manage overall system costs for patients who need medicine.

Acknowledgments The content in this chapter draws heavily on experiences and the knowledge acquired through the work of COST Action IS1304: *Expert Judgement Network: Bridging the Gap Between Scientific Uncertainty and Evidence-Based Decision Making*.

This Action supported various activities including a *Workshop on Expert Judgement to Enhance Health Decision Making* held in Malta between 7 October 2015 and 9 October 2015. The first and the second authors of this chapter were part of the organising committee of this workshop. They also collaborated in a short-term scientific mission to Strathclyde Business School, Scotland, and all three authors participated in a workshop for practitioners from various areas related to medicines reimbursement and use in Scotland.

References

Ariely, D. (2008). *Predictably irrational, the hidden forces that shape our decisions* (p. 9). New York: HarperCollins Publishers.

Aronson, J. K., Ferner, R. E., & Hughes, D. A. (2012). Defining rewardable innovation in drug therapy. *Nature Reviews, 11,* 253–254.

Baba, V. V., & HakemZadeh, F. (2012). Towards a theory of evidence based decision making. *Management Decision, 50*(5), 832–867. https://doi.org/10.1108/00251741211227546.

Banzi, R., Gerardi, C., Bertele, V., & Garattini, S. (2015). Approvals of drugs with uncertain benefit-risk profiles in Europe. *European Journal of Internal Medicine, 26,* 572–584.

Barends, E., Rousseau, D. R., & Briner, R. B. (2014). *Evidence-based management—The basic principles*. Amsterdam: CEBMa centre for Evidence-Based Management.

Bedford, T., & Cooke, R. (2001). *Probabilistic risk analysis*. Cambridge: Cambridge University Press.

Centre for Reviews and Dissemination. (2009). *Systematic reviews, CRD's guidance for undertaking reviews in health care*. York: York Publishing Services Ltd.

Cohen, D. (2014). Dabigatran: How the drug company withheld important analyses. *British Medical Journal (Clinical Research ed.), 349,* g4670.

Cohen, D. (2017). Cancer drugs: high price, uncertain value. BMJ. 359, j4543.

Collins, J. M., & Urban, C. (2014). The dark side of sunshine: Regulatory oversight and status quo bias. *Journal of Economic Behaviour and Organisation, 107,* 470–486. https://doi.org/10.1016/j.jebo.2014.04.003.

Cooke, R. M. (1991). *Experts in uncertainty: Opinion and subjective probability in science.* Oxford: Oxford University Press.

Council of the European Union. (2016). Council conclusions on strengthening the balance in the pharmaceutical systems in the EU and its Member States. http://www.consilium.europa.eu/en/press/press-releases/2016/06/17-epsco-conclusions-balance-pharmaceutical-system/.

Council of the European Union. (2017). Council conclusions on encouraging member states-driven voluntary cooperation between health systems. http://data.consilium.europa.eu/doc/document/ST-9978-2017-REV-1/en/pdf.

Davis, C., Lexchin, J., Jefferson, T., Gotzsche, P., & McKee, M. (2016). "Adaptive pathways" to drug authorisation: Adapting to industry? *British Medical Journal, 354,* i4437. https://doi.org/10.1136/bmj.i4437.

Directive 89/105/EEC of 21 December 1988 relating to the transparency of measures regulating the prices of medicinal products for human use and their inclusion in the scope of national health insurance systems https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A31989L0105.

Ermisch, M., Bucsics, A., Vella Bonanno, P., Arickx, F., Bybau, A., Bochenek, T., et al. (2016, September 28). Payers' views of the changes arising through the possible adoption of adaptive pathways. *Frontiers in Pharmacology* (7), 305. eCollection 2016. https://doi.org/10.3389/fphar.2016.00305.

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic. Why the adjustments are insufficient. *Psychological Science, 17*(4), 311–318.

European Commission. (2012). Process on corporate social responsibility in the field of pharmaceuticals platform on access to medicines in Europe working group on mechanism of coordinated access to orphan medicinal products (MoCA-OMP)-transparent value framework. http://ec.europa.eu/enterprise/sectors/healthcare/files/docs/orphans_conclusions_en.pdf2012.

European Commission. (2018a). Proposal for a Regulation of the European Parliament and of the Council on health technology assessment and amending Directive 2011/24/EU (2018/0018(COD). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A51%3AFIN.

European Commission. (2018b). *Innovative payment models for high-cost innovative medicines. Report of the expert panel on effective ways of investing in health (EXPH).* Luxembourg: Publications Office of the European Union.

European Food Safety Authority. (2014). Guidance of EFSA: Guidance on expert knowledge elicitation in food and feed safety risk assessment. *EFSA Journal, 12*(6), 3734.

European Observatory on Health Systems and Policies. (2017). How can cross-border collaboration in public procurement improve access to health technologies in Europe? *Policy Brief 21.* http://www.euro.who.int/__data/assets/pdf_file/0009/331992/PB21.pdf?ua=1.

FDA Office of Criminal Investigations. (2012, April 19). U.S. Pharmaceutical Company Merck Sharp & Dohme sentences in connection with unlawful promotion of Vioxx. https://www.fda.gov/ICECI/Criminalinvestigations/ucm301329.htm.

Ferrario, A., Arāja, D., Bochenek, T., Čatić, T., Dankó, D., Dimitrova, M., et al. (2017). The implementation of managed entry agreements in Central and Eastern Europe: Findings and implications. *PharmacoEconomics, 35*(12), 1271–85.

Gagnon, M. A. (2015). New drug pricing: Does it make any sense? *Rev Prescrire, 35*(380), 457–61.

Garattini, L., & Curto, A. (2016). Performance-based agreements in Italy: 'Trendy outcomes' or mere illusions? *PharmacoEconomics, 34,* 967–969. https://doi.org/10.1007/S40273-016-0420-1.

Garling, T., Kirchler, E., Lewis, A., & van Raaij, F. (2009). Psychology, financial decision making, and financial crises. *Psychological Science in the Public Interest, 10*(1), 1–47.

Garthwaite, P. H., Kadane, J. B., & O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of American Statistical Association, 100*(470), 680–700.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *The Annual Review of Psychology, 62,* 451–482.

Godman, B., Malmstrom, R. E., Diogene, E., Gray, A., Jayathissa, S., Timoney, A., et al. (2015). Are new models needed to optimize the utilisation of new medicines to sustain healthcare systems? *Expert Review of Clinical Pharmacology, 8*(1), 77–94. https://doi.org/10.1586/17512433.2015.990380.

Godman, B., Bucsics, A., Vella Bonanno, P., Oortwijn, W., Rothe, C. C., Ferrario, A., et al. (2018). Barriers for Access to New Medicines: Searching for the Balance Between Rising Costs and Limited Budgets. *Front Public Health*, *6*, 328.

Godman, B., Oortwijn, W., de Waure, C., Mosca, I., Puggina, A., Specchia, M. L., et al. (2016). Links between pharmaceutical R&D models and access to affordable medicines. A study for the ENVI Committee. http://www.europarl.europa.eu/RegData/etudes/STUD/2016/587321/IPOL_STU(2016)587321_EN.pdf.

Guerra-Junior, A. A., Pires de Lemos, L. L., Godman, B., Bennie, M., Osorio-de-Castro, C. G. S., Alvares, J., et al. (2017). Health Technology Performance Assessment: Real-World evidence for public healthcare sustainability. *International Journal of Technology Assessment in Health Care, 33*(2), 279–87.

Hartling, L., Hamm, M. P., Fernandes, R. M., Dryden, D. M., & Vandermeer, B. (2014). Quantifying bias in randomized controlled trials in child health: A meta-epidemiological study. *PLoS One, 9*(2), e88008.

Haycox, A. (2016). Why cancer? *PharmacoEconomics, 34*(7), 625–7.

Hill, A., Gotham, D., Fortunak, J., Meldrum, J., Erbacher, I., Martin, M., et al. (2016). Target prices for mass production of tyrosine kinase inhibitors for global cancer treatment. *British Medical Journal Open, 6*(1), e009586.

Irwin, J., & Peacock, S. (2015). Multi-criteria decision analysis: An emerging alternative for assessing the value of orphan medicinal products. *Regulatory Rapporteur, 12*(1), 12–15.

Joint Briefing Paper. (2015). “Adaptive licensing” or “adaptive pathways”. Deregulation under the guise of earlier access. Brussels: International Society of Drug Bulletins, Medicines in Europe Forum, Nordic Cochrane Centre. http://www.isdbweb.org/en/publications/view/adaptive-licensing-or-adaptive-pathways-deregulation-under-the-guise-of-earlier-access.

Joppi, R., Gerardi, C., Bertele, V., & Garattini, S. (2016). Letting post-marketing bridge the evidence gap: The case of orphan drugs. *British Medical Journal, 353,* i2978. https://doi.org/10.1136/bmj.i2978.

Kahneman, D. (2011). *Thinking fast and slow* (pp. 80, 122, 252). New York: Farrar, Straus and Giroux.

Kahneman, D., & Thaler, R. (1991). Economic analysis and the psychology of utility: Applications to compensation policy. *The American Economic Review, 81*(2), 341–346.

Kahneman, D., & Thaler, R. H. (2006). Utility maximisation and experienced utility. *The Journal of Economic Perspectives, 20*(1), 221–234.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263–291.

Lemos, L. L. P., Guerra Junior, A. A., Santos, M., Magliano, C., Diniz, I., Souza, K., et al. (2018). The assessment for disinvestment of intramuscular interferon beta for relapsing-remitting multiple sclerosis in Brazil. *PharmacoEconomics, 36*(2), 161–73.

Leopold, C., Vogler, S., Mantel-Teeuwisse, A. K., de Joncheere, K., Leufkens, H. G., & Laing, R. (2012). Differences in external price referencing in Europe: A descriptive overview. *Health Policy, 104*(1), 50–60.

O’Connor, M. A. (2005). Vioxx withdrawn from the market: Controversies continue to involve Merck & FDA. *Journal of Legal Nurse Consulting, 16*(1), 19–21.

Olson, C. M., Rennie, D., Cook, D., Dickersin, K., Flanagin, A., Hogan, J. W., et al. (2002). Publication bias in editorial decision making. *Journal American Medical Association, 287*(21), 2825–2828.

Otuteye, E., & Siddiquee, M. (2015). Overcoming cognitive biases: A heuristic for making value investing decisions. *The Journal of Behavioural Finance, 16,* 140–149. https://doi.org/10.1080/15427560.2015.1034859.

Paris, V., & Belloni, A. (2013). *Value in pharmaceutical pricing* (OECD Health Working Papers, No. 63). OECD. http://dx.doi.org/10.1787/5k43jc9v6knx-en.

Parkinson, B., Sermet, C., Clement, F., Crausaz, S., Godman, B., Garner, S., et al. (2015). Disinvestment and value-based purchasing strategies for pharmaceuticals: An international review. *PharmacoEconomics, 33*(9), 905–24.

Petticrew, M., & Roberts, H. (2003). Evidence, hierarchies, and typologies: Horses for courses. *Journal of Epidemiology and Community Health, 57,* 527–529.

Pfeffer, J., & Sutton, R. (2011, September 3). Trust the evidence, not your instincts. *The New York Times*. http://www.nytimes.com/2011/09/04/jobs/04pre.html?_r=0.

PharmaDiplomacy Working Group. (2016). *Principles for collaborative, mutually acceptable drug pricing*. Meteos. http://www.meteos.co.uk/wp-content/uploads/PHARMADIPLOMACY-REPORT-low-res.pdf.

Prescrire. (2016). New drugs, new indications in 2015: Little progress, and threats to access to quality healthcare for all. Editorial. *Rev Prescrire, 36*(388), 133–7.

Pritts, D. (2006). Vioxx …More to the story. *Journal of Legal Nurse Consulting, 17*(2), 11–15.

Regulation (EC) No 726/2004 of the European Parliament and of the Council of 31 March 2004 laying down Community procedures for the authorisation and supervision of medicinal products for human and veterinary use and establishing a European Medicines Agency https://ec.europa.eu/health/sites/health/files/files/eudralex/vol-1/reg_2004_726/reg_2004_726_en.pdf.

Regulation (EC) No 507/2006 on the conditional marketing authorisation for medicinal products for human use falling within the scope of Regulation (EC) No 726/2004 https://ec.europa.eu/health/sites/health/files/files/eudralex/vol-1/reg_2006_507/reg_2006_507_en.pdf.

Regulation (EU) No 1235/2010 of the European Parliament and of the Council of 15 December 2010 amending, as regards pharmacovigilance of medicinal products for human use, Regulation (EC) No 726/2004 laying down Community procedures for the authorisation and supervision of medicinal products for human and veterinary use and establishing a European Medicines Agency, and Regulation (EC) No 1394/2007 on advanced therapy medicinal products. https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2010:348:0001:0016:EN:PDF.

Regulation (EU) No 536/2014 of the European Parliament and of the Council of 16 April 2014 on clinical trials on medicinal products for human use, and repealing Directive 2001/20/EC https://ec.europa.eu/health/human-use/clinical-trials/regulation_en.

Rousseau, D. M. (2012). Envisioning evidence-based management. In D. M. Rousseau (Ed.), *The Oxford handbook of evidence-based management*. New York: Oxford University Press.

Rousseau, D. M., & Gunia, B. C. (2015). Evidence-based practice: The psychology of EBP implementation. *Annual Review of Psychology* https://scholar.google.com/scholar?q=Rousseau,+D.M.+and+Gunia,+B.C.+(2015).+Evidence-Based+Practice:+the+psychology+of+EBP+implementation.+Annual+Review+of+Psychology&hl=en&as_sdt=0&as_vis=1&oi=scholart.

Simoens, S., Picavet, E., Dooms, M., Cassiman, D., & Morel, T. (2013). Cost-effectiveness assessment of orphan drugs: A scientific and political conundrum. *Applied Health Economics and Health Policy, 11*(1), 1–3.

Simons, M., Houghton, S. M., & Aquino, K. (1999). Cognitive biases, risk perception, and venture formation: How individuals decide to start companies. *Journal of Business Venturing, 15,* 113–134.

Soares, M., & Bojke, L. (2018). Expert elicitation to inform health technology assessment. In L. C. Dias, A. Morton, & J. Quigley (Eds.), *Elicitation: The science and art of structuring judgement*. New York: Springer.

Sukkar, E. (2014, September 19). Still feeling the Vioxx pain. *The Pharmaceutical Journal*. http://www.pharmaceutical-journal.com/news-and-analysis/features/still-feeling-the-vioxx-pain/20066485.article.

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management, 14,* 207–222.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics, 106*(4), 1039–1061.

Vella Bonanno, P. (2003). *The managed entry of new drugs into a national health service—A case study for Malta.* Ph.D. Dissertation. Aberdeen: The Robert Gordon University. The British Library.

Vella Bonanno, P. (2010). *The managed entry of new drugs into a national health service—A case study for Malta, prior to its becoming a Member State of the European Union* (pp. 103–105). Saarbrucken: LAP Lambert Academic Publishing.

Vella Bonanno, P., Bucsics, A., Simoens, S., Martin, A. P., Oortwijn, W., Gulbinovic, J., et al. (2019). Proposal for a regulation on health technology assessment in Europe—opinions of policy makers, payers and academics from the field of HTA. *Expert review of pharmacoeconomics & outcomes research*, *19*(3), 251–256.

Vella Bonanno, P., Ermisch, M., Godman, B., Martin, A.P., Van Den Bergh, J., Bezmelnitsyna, L., Bucsics, A., Arickx, F., et al. (2017). Adaptive pathways: Possible next steps for payers in preparation for their potential implementation. *Frontiers in Pharmacology, 8,* Article 497. https://doi.org/10.3389/fphar.2017.00497. http://journal.frontiersin.org/article/10.3389/fphar.2017.00497/full; https://www.ncbi.nlm.nih.gov/pubmed/28878667.

Ward, D. J., Slade, A., Genus, T. M., & Stevens, A. J. (2014). How innovative are new drugs launched in the UK? A retrospective study of new drugs listed in the British National Formulary (BNF) 2001–2012. *British Medical Journal, 4,* e006235. https://doi.org/10.1136/bmjopen-2014-006235.

World Health Organisation. (2017). Health technology assessment http://www.who.int/medical_devices/assessment/en/.

Chapter 20
Structured Expert Judgement Issues in a Supply Chain Cyber Risk Management System

Alberto Torres-Barrán, Alberto Redondo, David Rios Insua, Jordi Domingo, and Fabrizio Ruggeri

Abstract The escalation of cyberthreats is a major problem for supply chain managers with potentially enormous impacts affecting service availability and reputation, among other performance indicators. We sketch a framework and system to support supply chain cyber risk management. As data regarding impacts of cyberattacks are scarce and difficult to obtain, we describe how we acquire the required operational parameters through structured expert judgement techniques. We then describe how the whole framework is set up and implemented.

Keywords Supply chain risk management · Cybersecurity · Structured expert judgement

20.1 Introduction

Organisations worldwide are suffering cyberattacks with important consequences. This is increasingly perceived as a major global problem, as reflected, e.g. in the World Economic Forum (2018) Global Risks Report, and becoming even more important as companies, administrations and individuals get more and more interconnected, facilitating the spread of cyberthreats. As an example, the recent WannaCry attack affected around 45,000 systems globally, including the UK NHS, Renault and Telefónica, causing major service interruptions; its ransomware caused estimated financial losses of nearly $4 billion. Another relevant example is the Target data breach, McGrath (2014), in which a cyberattack to that company through one of its suppliers led to the loss of 70 million credit card details, entailing major reputational damage.

A. Torres-Barrán (✉) · A. Redondo · D. Rios Insua
Institute of Mathematical Sciences, ICMAT-CSIC, Madrid, Spain
e-mail: alberto.torres@icmat.es

J. Domingo
Blueliv, Barcelona, Spain

F. Ruggeri
CNR-IMATI, Milano, Italy

A. M. Hanea et al. (eds.), *Expert Judgement in Risk and Decision Analysis*,
International Series in Operations Research & Management Science 293,
https://doi.org/10.1007/978-3-030-46474-5_20

Thus, organisations face significant risks due to the need of using interconnected suppliers for their services. To alleviate such problem, the discipline of Supply Chain Cyber Risk Management (SCCRM) aims at implementing strategies to oversee cyber risks with the objective of mitigating service interruptions and decreasing their eventual impact, Redondo et al. (2018). To further complicate matters, for reputational reasons, there is a reluctance to release information concerning attacks, as this could affect relations with stakeholders and entail a loss of customers (Pelteret and Ophoff 2016). In order to supplement such lack of data, we may appeal to structured expert judgement elicitation techniques, Cooke (1991), O'Hagan et al. (2006) and Clemen and Reilly (2013), exploiting the knowledge available from cybersecurity experts to support cyber risk management.

This paper briefly sketches a framework for SCCRM in Sect. 20.2. As data regarding occurrences and impacts of cyberattacks are difficult to obtain, we need to rely on various expert judgement techniques to assess the parameters[1] in the required impact and preference models in Sects. 20.3, 20.4 and 20.5. Section 20.6 illustrates operational aspects of our framework and system. We end up with a discussion in Sect. 20.7.

20.2 A Framework for SCCRM

We aim at supporting a company c interconnected with k suppliers in its supply chain cyber risk management activities. We briefly sketch the framework that we use for such purpose, with full technical details in Redondo et al. (2018). Our focus will be on the expert judgement techniques and processes used to extract beliefs and preferences from experts to make the framework operational.

The company faces three cyberattack scenarios: *direct attacks*; *attacks to its suppliers not transferred to the company*, but affecting it through the unavailability of the corresponding product or service and, finally, *attacks targeting the suppliers that are eventually transferred to the company*. Some of these attacks could be successful in the sense of producing noticeable harm to the company. We assume we have access to a Threat Intelligence Service (TIS) (Tittel 2017) which compiles data, both for the company and its suppliers, about: *potential or actual attacks through various attack vectors*, such as the number of malware-infected devices or the number of phishing attempts suffered; their *security environment*, as reflected in e.g, the number of negative tweet mentions about the company and its suppliers and, finally, the *security posture*, covering, for instance, the number of open ports or installed firewalls.

Based on the TIS data, and other available information, we aim at assessing the following basic ingredients in our SCCRM framework: the probabilities that the company and its suppliers are attacked; the probability that an attack to a supplier gets transferred to the company; the impacts over the company associated with eventual attacks, direct or indirect, during the relevant security planning period and how

[1] For confidentiality reasons, data have been masked when presented.

does the company evaluate various impacts. We then integrate such assessments to evaluate the supply chain cyber risks that the company could face and support risk management decisions both at strategic and operational levels.

We start by estimating the probability that the company and its suppliers are attacked through various vectors. First, we aggregate the information about the security environment and posture of the company and the suppliers through two Indicators which are a linear combination of the corresponding collected variables. Next, various attack vectors are considered conditionally independent given the posture and the environment, as we model all attack probabilities through logistic regressions with explanatory variables referring to the indicators of the corresponding attack vector, and the environment and posture indicators. As companies are reluctant to provide their attack data, we indirectly estimate the corresponding logistic regression weights with the aid of expert judgement. Besides, we need to be able to assess the probabilities that attacks to suppliers get eventually transferred to the company, which we obtain directly from experts. With all this information, we may assess the relevant attack probabilities directly to the company or indirectly through its suppliers, duly apportioning their sources.

We next estimate the impacts that an attack might have over the company, taking into account the three types of attack scenarios mentioned above. The relevant impacts may vary across organisations. Some examples are the costs associated with the rupture of a service provided by a supplier, as in the Wannacry case with Telefonica; the costs associated with the unavailability of the company's service or product, as in the Wannacry case with the UK NHS or the loss of reputation associated with a major attack, which might induce a loss of customers or stock value, as in the Target case. We typically use continuous distributions assessed based on quantiles obtained from experts. We then aggregate various impacts through a multi-attribute utility function, if we need to cater for risk attitudes, González-Ortega et al. (2018).

Based on the above probability and preference models, we assess the expected impacts and risks, duly apportioning them to various sources (suppliers, transferred attacks from suppliers, or direct attacks, as well as the different attack vectors) and use such assessments to rank suppliers, negotiate service level agreements, or allocate cybersecurity risk management resources, including cyber insurance products, among other possibilities.

We present now how the expert judgement elicitation tasks described above are actually implemented and how we integrate all the information for risk management purposes.

20.3 Expert Calibration

We start by calibrating eight available experts based on their cybersecurity knowledge. After a training session, we passed them a questionnaire which served for weighting purposes.

20.3.1 Calibration Process

We used reports, such as Kaspersky (2016) or Imperva (2016), to elaborate a questionnaire about cybersecurity attacks impacting SMEs and large companies. The questionnaire was built using the Google Forms tool and was ran initially with two colleagues to check for comprehensibility. It included ten questions concerning attack likelihoods and impacts. Two example questions are as follows:

> What was the number of new ransomware types over the last year?
>
> What was the average cost in dollars of a ransomware incident over the last year?

Before interviewing the experts, we suggested that they watched YouTube *It's a Risky Life* videos 2, 3 and 4 to refresh the basic issues and concepts required for the session. When beginning, we also provided a review of the concepts, objectives and process to be followed. Some of the experts were interviewed physically, the remaining ones through the communication tool Skype. We introduced the process as follows:

> We present here a few general questions in relation with cyber security attacks, their likelihood and impacts. Answer them with what represents for you the quantities described.
>
> At each question, we shall ask you about an interval which covers with high probability (0.90) the actual value based on the 5% and the 95% quantiles and what is, according to you, the median value. For example, the interval could be [30-40] and the median value 35, so the answer would be 30, 35, 40.

Several motivating and warming up examples were included to further facilitate understanding, together with additional explanations about cognitive and motivational biases. In such a way, we tried to make sure that the experts understood the questions and response format correctly. They were also encouraged to ask for further clarification whenever they felt like. We also provided graphical support (fortune wheels) to facilitate the assessments. In the end, we verified whether the experts had answered all questions according to the instructions and checked that the results had been submitted correctly, allowing them to modify responses upon reflection. Upon observing inconsistent results in one of the experts (Ex2), we checked that he actually misunderstood the concepts underlying some of the questions and we decided to suppress his responses from the study, due to lack of time to repeat the process. We also eliminated questions Q6 and Q7 as the experts' answers were astray, possibly because of inadequate wording on our behalf.

20.3.2 Exploratory Analysis of Experts' Responses

We start with some exploratory data analysis with the responses. We display the experts' point and interval responses in Table 20.1, as well as the actual values. We double checked whether some of the questions had been misunderstood (consider e.g.

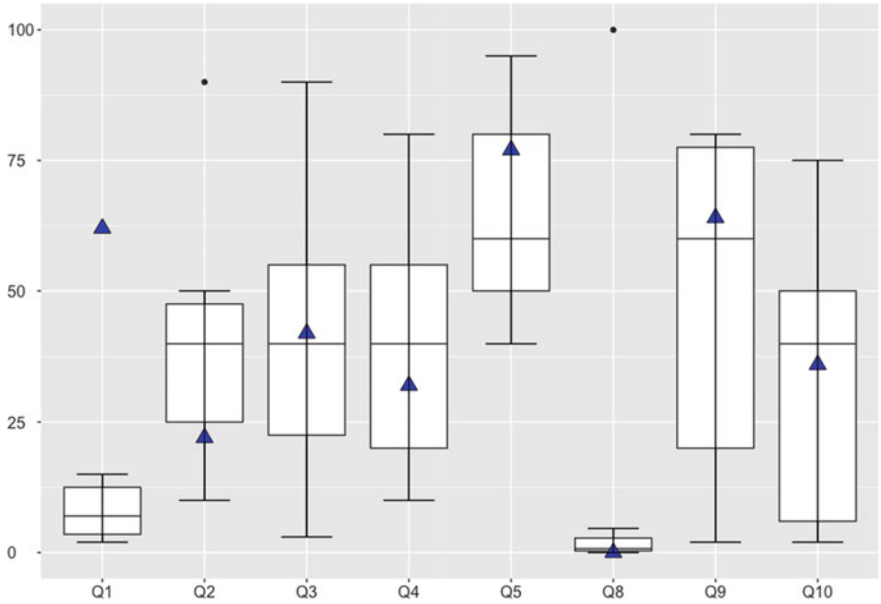


Fig. 20.1 Boxplots of median experts answers for the retained experts and questions. True value is shown as a blue triangle. Q8 responses normalised to [0, 100]

the responses of Ex8 for Q1 and Ex6 for Q9 and Q10) but the participants confirmed their results. Incidentally, this pointed out towards somewhat unknown cybersecurity topics about which even experts seem to be not sufficiently aware of.

Boxplots in Fig. 20.1 show the experts' median responses and the true value is marked with a blue triangle. We have normalised the answers of question Q8 to the [0,100] range and removed the extreme outliers (responses that lie more than 3 times the interquartile range below the first quartile or above the third quartile). We note that the boxplots tend to cover the true values, except for Q1 (which was clearly underestimated) and Q8 (which was overestimated), reflecting also large variability in the responses.

We next display the scatter plots of the experts' responses and their correlation matrix, Fig. 20.2, in which we have also included the true values (as the responses of a ninth expert). We have removed the very extreme observation of Ex8 for Q1 from this figure. We do not observe very high correlations. For example, if we use 0.5 as cutting value for noticeable correlations, only Ex1 with Ex3, Ex4 and Ex7, as well as Ex3 with Ex7 show a relevant correlation between themselves. Also, only Ex1, Ex3 and Ex7 show some correlation with the true values.

Table 20.2 summarises the performance of the experts over the 8 retained seed questions, presenting how many observed responses appeared in each of the intervals, compared with the expected responses in such intervals. Ex1, Ex7 and, to a lesser extent, Ex6 seem to perform better as they have more hits in the central intervals.

Table 20.1 Responses of experts and actual values for the 7 retained experts (Ex) and 8 retained questions (Q). Expert responses in format 5% quantile, median, 95% quantile

Experts	Q1	Q2	Q3	Q4	Q5	Q8	Q9	Q10
Ex1	1,15,50	2,50,100	20,50,100	10,45,90	50,60,100	500,750,1000	50,60,75	30,40,40
Ex3	5,10,10	40,45,50	50,60,60	10,10,20	40,50,50	1000,1000,3000	75,80,85	10,10,15
Ex4	2,2,3	20,20,40	2,3,5	25,30,35	30,40,50	5000,15000,400000	70,75,80	70,75,80
Ex5	5,7,8	6,10,50	18,30,35	40,40,50	90,95,99	20000,1500000,2000000	20,25,30	1,2,6
Ex6	3,4,5	10,30,30	30,40,50	60,80,100	40,50,60	100,10000,15000	1,2,5	1,2,5
Ex7	1,3,5	1,40,100	1,90,100	1,10,100	1,80,100	1,12000,100000	1,80,100	1,60,100
Ex8	500,750,1000	70,90,100	10,15,25	50,65,80	70,80,90	50000,70000,150000	10,15,20	20,40,45
True value	62	22	42	32	77	700	64	36

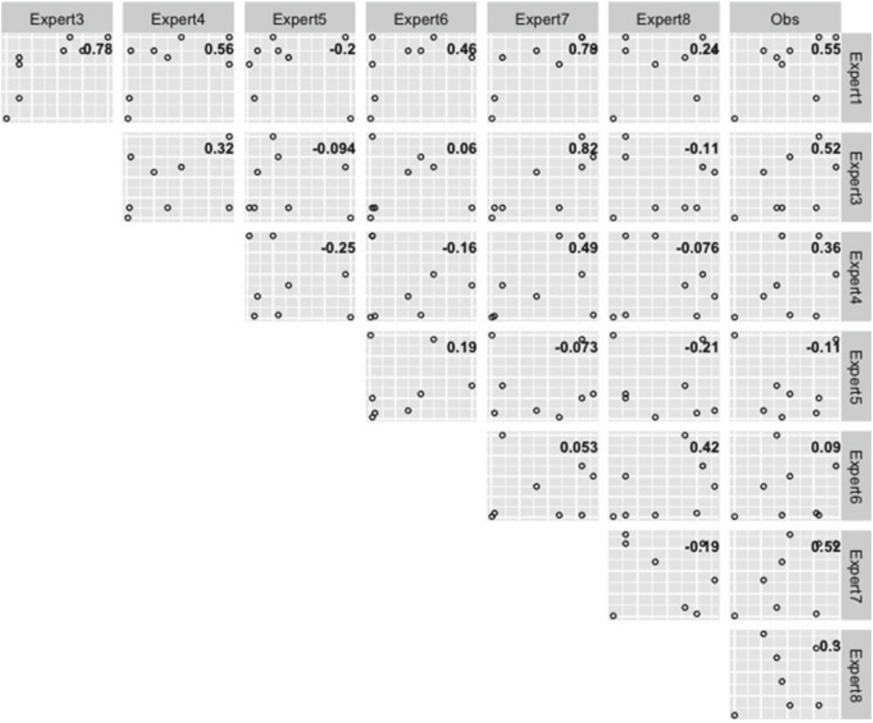


Fig. 20.2 Scatterplot of expert answers and correlation matrix

Table 20.2 Performance of experts over the 8 seed questions, together with expected frequencies

Expert	Below 5th	5th–50th	50th–95th	Above 95th
Ex1	0	5	2	1
Ex3	4	0	0	4
Ex4	3	0	2	3
Ex5	3	0	1	4
Ex6	1	2	1	4
Ex7	0	6	1	1
Ex8	4	2	0	2
Exp. Freq.	0.4	3.6	3.6	0.4

20.3.3 Calibration

We used Excalibur (Lighttwist 2018) to score the experts as summarised in Table 20.3, based on Cooke's classical method (Cooke 1991), which provides also the calibration and information scores of the experts retained. We did not use the Decision Maker (DM) optimisation and adopted a significance level of 0.001. Ex1 and Ex7 retained most of the weight.

We performed a robustness analyses and found questions Q5 and Q9 to be the most influential over the results. Also, experts Ex1 and Ex7 showed the lowest discrepancy.

20.4 Attack Probabilities' Assessment

We describe now how to extract the cybersecurity knowledge from the experts to enable us building our underlying SCCRM model. For this, we created a second questionnaire with Google Forms, which included a short introduction outlining the procedure to answer the questions:

> The following questions will aid us in extracting your expertise on cyber security so as to build a model that allows us to forecast sufficiently important attacks to a company. Please feel confident. There are no right or wrong answers. We shall be posing questions that take advantage from your cyber security expertise.

The questions were divided into two groups: first, attack probability questions and, then, questions related to the environment and posture.

20.4.1 Attack Probabilities

With this first group of questions, we aimed at obtaining for each expert i the probability q_i of various events. We then aggregate the probabilities through $p = \sum \omega_i q_i$

Table 20.3 Calibration scores, weights and information scores of the experts using Cooke's classical method

Expert	Calib.Sc.	Weight	Info.Sc.
Ex1	0.429	0.820	1.834
Ex3	0.000	0.000	3.440
Ex4	0.000	0.000	2.592
Ex5	0.000	0.000	2.143
Ex6	0.002	0.004	2.636
Ex7	0.145	0.176	1.168
Ex8	0.000	0.000	1.441

(effectively, $i \in \{1, 6, 7\}$, as the other experts have weight zero), where the weights ω_i are the result of the calibration process reflected in Table 20.3. Based on such probabilities, at this stage, we extracted the judgements required to obtain the logistic regression parameters mentioned in Sect. 20.2. Each question included a description of a relevant scenario with the answer interpreted as a probability.

We illustrate the procedure for attacks due to malware infections. The TIS we use is able to detect three types of malware. Thus, the model has three coefficients besides β_0, and the logistic regression model we use is

$$Pr(y = 1 \mid \beta_0, \boldsymbol{\beta}, \mathbf{n}) = \frac{\exp(\beta_0 + \boldsymbol{\beta} \cdot \mathbf{n})}{1 + \exp(\beta_0 + \boldsymbol{\beta} \cdot \mathbf{n})} \tag{20.1}$$

where $\mathbf{n} = (n_1, n_2, n_3)$ is the vector containing the counts of the three types of malware, β_0, $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$ are the logistic parameters and $y = 1$ indicates that the attack through malware was successful (sufficiently harmful). First, we ask the experts for the attack probability in a scenario in which no such infections were detected by the TIS, $\mathbf{n} = (0, 0, 0)$. The actual question posed to the experts was as follows:

> Assume that the TIS has detected no evidence of malware infections in your network, what would be the probability of actually suffering an attack based on malware?

We then aggregate the responses of the (three) experts to obtain the estimated probability p_0. Since we are assuming that the attack probability follows Eq. (20.1) and no infections are found, we have

$$p_0 = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)},$$

and thus $\beta_0 = \sigma(p_0)$, where $\sigma(\cdot)$ is the logit function or simply log-odds, i.e.

$$\sigma(p) = \log\left(\frac{p}{1 - p}\right). \tag{20.2}$$

We compute the remaining required coefficients in a similar manner, asking the experts to provide an estimate of the attack probability p_j if the TIS detects a certain number m_j of j-th level infections of the attack vector and none of the rest, aggregating the responses and solving for the corresponding β_j. We need to ask at least one question per coefficient to each expert. A typical question would be as follows:

> Assume that the TIS has detected 5 malware infected devices of level 1 in your network, what would be the probability of actually suffering an attack based on malware?

The previous question proposes a scenario in which $n_k = m_j$ if $k = j$ and $n_k = 0$ if $k \neq j$, with $m_j = 5$ and $j = 1$. We aggregate the expert responses in p_j and, using Eqs. (20.1) and (20.2), we finally get

Table 20.4 Number of reference infections (m_j), responses of three retained experts (Ex1, Ex6, Ex7), aggregated probabilities (p_j) and estimated (β_j) parameters

j	m_j	Ex1	Ex6	Ex7	p_j	β_j
0	0	0.01	0.05	0.01	0.01076	−4.52110141681262
1	5	0.02	0.1	0.02	0.02032	3.74597734
2	5	0.15	0.1	0.02	0.12692	4.13540737
3	1	0.35	0.2	0.05	0.296	3.65468251

$$\beta_j = \frac{\sigma(p_j) - \beta_0}{m_j}. \tag{20.3}$$

Table 20.4 includes the responses of the three experts, their aggregation and the corresponding parameters. Recall that the probabilities are aggregated using the weights from Table 20.3. Also, note that the coefficients regarding different infection levels are independent between them and depend only on β_0. We also posed additional questions for each level j with different values of m_j to check whether the experts are consistent in their answers, as similar β_j values should be obtained.

The previous procedure is fully general and can be applied to any attack vector with a variable number of infection levesl k. Thus, we perform the above for each attack vector detectable by the TIS.

20.4.2 Environment and Posture

We describe now how to incorporate information about the security environment captured by the TIS. Examples of relevant variables would include the number of negative mentions about the company in major social media or the number of mentions in security blogs.

Let e_i be the i-th incumbent variable, $i = 1, \dots, k$, rescaled to [0, 1]; we assume that the bigger the e_i is, the worse the security environment is. For example, the bigger the number of negative mentions in major social media about the company, the more irritated hacktivists would be, therefore being more prone to launching an attack. We define an environment index e which aggregates the k environment variables through a multi-criteria value function (González-Ortega et al. 2018),

$$e = \sum_{i=1}^{k} \lambda_i e_i$$

with $\lambda_i \geq 0$, $i = 1, \dots, k$ and $\sum_{i=1}^{k} \lambda_i = 1$. We determine the weights λ_i by asking experts to compare pairs of security environment contexts T_i, $i = 1, \dots, k - 1$, identifying the corresponding system of equations and solving it.

For example, for the variables mentioned above, we could pose the question:

> How would you weight the relative importance of the number of mentions in security blogs and negative mentions in social media regarding the likelihood of receiving a successful harmful attack? Both numbers should add up to 100; the higher the weight, the more impact you give to such variable (in the sense of deeming more likely an attack).

In this case, the expert's answer should be a pair of numbers $T_i = (a_i, b_i)$, which adds up to 100 for ease of interpretation (but are later rescaled to add up to 1). If both are 50, the expert considers both variables equally relevant when assessing the security environment of the company. This leads to a system of equations

$$\begin{aligned}\lambda_1 \times a_1 &= \lambda_2 \times b_1\\ \lambda_2 \times a_2 &= \lambda_3 \times b_2\\ &\vdots\\ \lambda_{k-1} \times a_k &= \lambda_k \times b_k\\ \sum \lambda_{i=1}^k &= 1\end{aligned}$$

The solution of the previous system is the final weights λ_i. The minimum number of questions to be posed to each of them is $k-1$, as the k-th equation relies on the restriction that all weights should add up 1. We select overlapping pairs of environment variables for the questions, comparing variables 1 and 2; 2 and 3 and so on, until the $(k-1)$-th and k-th variables are compared. To mitigate biases, the order in which the questions are posed is randomised. Moreover, additional questions using other combinations of variables are added to check for consistency. We then find the value function corresponding to each expert and aggregate them with equal weights.[2] Table 20.5 includes the expert responses for the above type of questions with four environment variables ($k = 4$) and the seven retained experts. The first three columns reflect the response of the pairwise comparisons, whereas the last four contain the computed weights for each expert.

Table 20.5 Environment responses of experts and weights

Ex	T_1	T_2	T_3	λ_1	λ_2	λ_3	λ_4
1	(70, 30)	(30, 70)	(50, 50)	0.292	0.125	0.292	0.292
3	(60, 40)	(30, 70)	(60, 40)	0.235	0.157	0.365	0.243
4	(10, 90)	(10, 90)	(50, 50)	0.006	0.052	0.471	0.471
5	(15, 85)	(40, 60)	(85, 15)	0.060	0.340	0.510	0.090
6	(60, 40)	(30, 70)	(70, 30)	0.257	0.171	0.400	0.171
7	(60, 40)	(40, 60)	(50, 50)	0.273	0.182	0.273	0.273
8	(70, 30)	(30, 70)	(50, 50)	0.292	0.125	0.292	0.292
				0.202	0.165	0.372	0.262

[2] Note that this refers to value judgements, not belief judgements as in Sect. 20.3.

A similar procedure is applied to combine the posture variables l_i into a posture indicator, $l = \sum_i v_i l_i$, with $\sum_i v_i = 1$ and $v_i \geq 0$, $\forall i$. Finally, we incorporate the environment and posture indices into the logistic regression model using the procedure in Sect. 20.4.1. First, we construct a scenario where no infections are found by the TIS and ask the experts how much the attack probability would increase assuming a certain value for one of the security environment variables, as in the following example:

> Assuming that the TIS has detected no evidence in the network concerning malware infections, how much would the attack probability increase if we detect a value for the first environment variable equal to 10 and none for the others?

Since the rest of the environment variables are zero, we can compute the index value as $e = \lambda_1 \times e_1$, where $e_1 = 10$, and expand the attack vector. Continuing with the example of malware infections, we have that now $\mathbf{n} = (0, 0, 0, e, 0)$, where the last two elements correspond now to the environment and posture indices. Then, since we are asking for the increase in probability, assuming the expert's answer is Δ with $p_0 + \Delta \leq 1$, we have $p_e = p_0 + \Delta$. We then substitute in Eq. (20.3) to obtain

$$\beta_e = \frac{\sigma(p_e) - \beta_0}{e}.$$

As before, consistency questions are posed (using different environment variables and values for e). We apply the same procedure to incorporate the security posture.

20.4.3 Probability of Attack Transfer

Finally, we also ask to the experts questions regarding the probability of an attack being transferred from a supplier to the company. There are as many questions as attack types, as we use the same probabilities for all suppliers. For each of the types, we aggregate the experts' probabilities. As an example, the probabilities for the transfer of malware attacks were 0.3, 0.1 and 0.2, respectively, for experts 1, 6 and 7. The final aggregated probability, using the weights in Table 20.3, is 0.282.

20.5 Attack Impact Assessment

The final group of questions refers to information concerning the impacts of a successful attack. They are different from the previous questions as they are company specific: for example, even if the consequence of a successful attack may be the same for two companies, like losing 1% of their customers, their economic impact will typically differ; the unavailability period will depend on the company's recovery capacity; moreover, distinct companies assess impacts differently. Thus, these questions are answered by in-company experts rather than by general cybersecurity experts. In the same manner, such type of information from the suppliers may not be readily available as its experts could have no incentives to answer the required questions or even be unavailable for the necessary elicitation exercise. In this case, the in-company experts may try to estimate what would be the answer to the questions for the corresponding suppliers.

As before, we introduce a training session with the local experts as well as an eventual aggregation procedure, if there are several of them available.

20.5.1 Relevant Impacts

Relevant cyberattack impacts might change across organisations. In our supply chain area, we have focused on downtimes, for both the company and its suppliers, and the induced reputational damage.

We model the downtimes in hours with Gamma distributions (for the company and suppliers). To obtain estimates of the parameters for these distributions, we ask the experts for at least two of its quantiles, for instance, the first and the third quartiles. An example question would be as follows:

> What is the duration of the downtime in hours due to malware at your organisation such that you would expect 25% of the downtimes to be below this value?

Once we obtain the quantiles, we use a least squares approach to estimate the parameters of the distributions, as in Morris et al. (2014). We may ask additional quantiles to perform consistency checks. As an example, in one case, an expert provided as first and third quartiles, respectively, 2 and 6. The best fitting gamma distribution was a Gamma(1.79, 0.40). After obtaining the distributions, we may compute centrality measures such as the mean or the median, if required, or use them for simulation purposes.

We performed similarly in relation with reputational damage, estimating the proportion of lost customers due to a certain type of attack with a Beta distribution.

20.5.2 Aggregating Impacts

We aggregate the three types of impacts taking into account the costs associated with unavailable services and the percentage of lost customers due to reputational damage. For such purpose, we require τ, the market share for the company; η, the (monetary) market size and κ_s and κ_c, the cost per hour of supplier and company service unavailability, respectively. For the required additional information, the corresponding questions are straightforward and directly posed to in-company experts.

The downtime costs of the supplier s and company c after a sufficiently harmful attack are

$$c_{i_s} = \kappa_s \times i_s,$$

$$c_{i_c} = \kappa_c \times i_c.$$

The reputational cost after a successful attack is approximated through the cost associated with clients abandoning the company, which would be

$$c_d = d \times \tau \times \eta.$$

Recall now that there are three types of attacks: direct attacks to the company, with entailed cost $c = c_d + c_{i_c}$; attacks to the supplier that disrupt its service but are not transferred to the company, with cost $c = c_{i_s}$; attacks to the supplier that disrupt its service and are transferred to the company. The cost in this case would be $c = c_d + c_{i_c} + c_{i_s}$.

When necessary, we shall use the expected costs

$$\bar{c}_{i_s} = \kappa_s \times \bar{i}_s, \quad \bar{c}_{i_c} = \kappa_c \times \bar{i}_c, \quad \bar{c}_d = \bar{d} \times \tau \times \eta.$$

20.5.3 Utility Elicitation

If we wish to cater for the company's risk attitude, we would introduce an utility function. A simple but very useful form of utility function arises when the relative risk aversion is set to a constant, in which case we have $u(x) = 1 - \exp(-\rho x)$ with $\rho > 0$, Keeney and Raiffa (1993), where x is the relevant attribute. To assess the risk tolerance parameter ρ, we ask the DM to determine the largest stake $x_{\max}$ for which she would accept the 50–50 gamble

$$\begin{cases} 2x_{\max} & \text{with probability } \frac{1}{2}, \\ -x_{\max} & \text{with probability } \frac{1}{2}. \end{cases}$$

This leads to the approximate expression $\rho \approx \frac{1}{2x_{\max}}$ (González-Ortega et al. 2018). Consistency checks would lead us to elicit additional values and iterative attempts to assess such value.

20.6 Operational Uses

We now have the necessary components to implement our SCCRM framework. We begin first by sketching some of its potential uses. We then describe how to implement it, and finally provide a numerical example.

20.6.1 Some Uses

The above information may be summarised in several measures and indices that may be used for risk monitoring and management purposes. These include attack probabilities through different attack vectors, both through the various suppliers or the company, resulting in a successful attack; the direct attack to the company probability; the induced attack probabilities and the total attack probability. Recall that if an attack is successfully transferred from a supplier, there are unavailability and reputational costs. Thus, we include also the expected impact due to direct attacks, the expected impact induced from attacks through suppliers and the total expected impact generated. Finally, we would also employ the corresponding expected utilities.

As an example, we provide the expressions for two of such indices whose use we illustrate in Fig. 20.3. First, the attack probability to the company c through a specific attack vector a is

$$p_c^a = \frac{\exp(\beta_0^a + \boldsymbol{\beta}^a \cdot \mathbf{n}_c^a)}{1 + \exp(\beta_0^a + \boldsymbol{\beta}^a \cdot \mathbf{n}_c^a)},$$

where $\mathbf{n}_c^a$ represents the a-th attack vector count for the company c, including the environment and posture indicators. Based on them, the direct Attack probability to the company is

$$\mathrm{AP}_c = \sum_{k=1}^{|\mathcal{A}|} \sum_{\mathcal{I} \in \mathcal{C}_{\mathcal{A},k}} \left(\prod_{a \in \mathcal{I}} p_c^a \prod_{a \in \mathcal{A} \setminus \mathcal{I}} (1 - p_c^a) \right),$$

where $\mathcal{C}_{\mathcal{A},k}$ is the set of all possible combinations of k elements taken from $\mathcal{A}$, the set of all incumbent attacks.

We describe now how we use in our framework the risk indicators:

- *Risk management*. We set up warning and critical level alarms for the indices to advise when specially dangerous situations have been detected. When such levels are reached, as we have apportioned them to various sources (vector attacks and suppliers), we may point out to the most critical ones to try to act over them.
- *Risk forecasting*. As the framework is running over time and the indices are periodically re-evaluated, each of the indices mentioned above may be viewed as an observation of a time series. We may, therefore, introduce forecasting models (specifically, we use dynamic linear models, West and Harrison 2006) for such series to forecast whether we shall reach the critical levels (through long-term forecasts) or which levels should we expect in the near future (through short-term forecasts). These forecasts can also detect sudden changes in the behaviour of the series and, consequently, suggest potential security issues.
- *Supplier negotiations*. We can use the indices produced to rank suppliers according to the risk they induce. We can also employ them to negotiate minimum induced security requirements to demand actions to suppliers or negotiate service level

agreements, say requiring to maintain a risk induced level below a certain value to preserve business continuity.

- *Insurance*. We may use the risk series generated to demonstrate low risk levels at the incumbent company and, consequently, negotiate lower insurance premiums. Alternatively, from the point of view of an insurance company, we could introduce insurance products with variable premium depending on the risk indices integrated over time. For example, an incentive could be introduced if, say, the average and maximum risk indices fall below a certain level over the contracted period of time.

20.6.2 Operations

The above framework has been implemented to support a dynamic approach to SCCRM in conjunction with an available TIS. For a given company, the TIS periodically gathers data, and the system computes the risk indices, provides various forecasts, issues warnings and performs update operations as follows:

1. Obtain new attack vectors evaluating the security posture and environment of the company and its suppliers.
2. Compute attack probabilities for suppliers and company, for various attacks and globally.
3. Estimate the expected impacts and utility for the company.
4. Launch alarms depending on limits defined.
5. Display risks associated with attack vectors and suppliers.
6. Predict risks for the next k-periods ahead.
7. (Update the probability models).
8. Proceed to the next period.

All of the tasks have been described above, except for the seventh one which refers to updating the parameter distributions in the logistic regression and impact models through MCMC methods as standard in Bayesian inference, e.g. French and Insua (2000). This would be possible as long as the company releases relevant data about attacks.

Figure 20.3 provides a trace of the model which runs periodically acquiring new probabilities and costs. Specifically, we show the evolution for $T = 100$ time steps of the direct Attack Probability (AP), the induced probabilities from two suppliers (IAP1, IAP2) and the Global Attack Probability (GAP). Here, we can observe that from $T = 0$ to $T = 40$ suppliers 1 and 2 induced similar risks. However, from $T = 40$, we may prefer supplier 1 since it seems to induce a lower risk to the company. We may fit DLMs (West and Harrison 2006; Petris et al. 2009) to forecast the attack probabilities k-steps ahead. Figure 20.3 presents the predictive distribution for $k = 1, \ldots, 20$, from period 100, with the corresponding predictive intervals.

The framework (its components, its output and its implementation) has been validated by cybersecurity and interface experts and is currently operational.

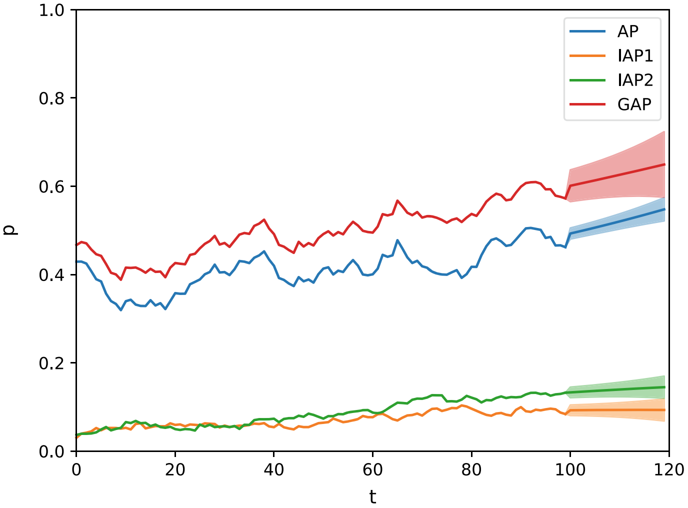


Fig. 20.3 Trace of risk indices over time

20.7 Discussion

The proliferation of cyberattacks and the increasing interconnectedness of organisations is framing the new field of SCCRM with several commercial solutions available. For reputational reasons, organisations are reluctant to release data concerning attacks. Therefore, we have sketched an approach to SCCRM which uses structured expert judgement techniques to assess the parameters required to make the approach implementable. We have focused on how suppliers may affect organisations, but the ideas extend to the impact of suppliers, and so on.

In line with the contents of this volume, we have presented the SEJ aspects of the framework as well as its operational implementation, covering issues concerning calibration of experts; eliciting attack probabilities indirectly through logistic regression models; aggregating environment and posture variables through multi-attribute value functions; directly eliciting transfer attack probabilities; eliciting impact distributions through quantiles; aggregating the impacts and, finally, eliciting utilities to cater for risk attitudes. We have also described how such information is integrated for various risk management purposes. Mathematical details may be seen in Redondo et al. (2018).

The whole framework has been implemented through Python routines based on a specific TIS and is running successfully supporting several companies in their SCCRM duties. The experience gained will allow us to further refine the framework; improve and/or expand the attack vectors as well as the assessment of the environment and posture.

Acknowledgments The authors acknowledge the support by the Spanish Ministry of Economy and Innovation program MTM2017-86875-C3-1-R and RTC-2017-6593-7, the ESF-COST Action IS1304 on Expert Judgement and the AXA-ICMAT Chair on Adversarial Risk Analysis. This work has also been partially supported by the Spanish Ministry of Economy and Competitiveness, through the "Severo Ochoa" Program for Centers of Excellence in R&D (SEV-2015-0554), and has received funding from the European Union's H2020 Program for Research, Technological Development and Demonstration, under grant agreement no. 740920 (CYBECO). AR work acknowledges the Spanish Ministry of Economy for his PTA position. FR acknowledges the support of the Chair of Excellence of the Comunidad de Madrid. FR and DRI acknowledge the support of the US National Science Foundation through grant DMS-163851.

References

Clemen, R. T., & Reilly, T. (2013). *Making hard decisions*. Duxbury Press.

Cooke, R. (1991). *Experts in uncertainty: Opinion and subjective probability in science*. Oxford University Press.

French, S., & Insua, D. R. (2000). *Statistical decision theory*. Wiley.

González-Ortega, J., Radovic, V., & Insua, D. R. (2018). Utility elicitation. In L. Dias, A. Morton, & J. Quigley (Eds.) *Elicitation: The Science and Art of Structuring Judgement*, pp. 241–264. Springer.

Imperva. (2016). DDoS Threat Landscape Report 2015–2016. https://lp.incapsula.com/rs/804-TEY-921/images/2015-16%20DDoS%20Threat%20Landscape%20Report.pdf.

Kaspersky. (2016). Story of the year: The ransomware revolution. https://securelist.com/kaspersky-security-bulletin-2016-story-of-the-year/76757/.

Keeney, R. L., & Raiffa, H. (1993). *Decisions with multiple objectives: Preferences and value trade-offs*. Cambridge University Press.

Lighttwist. (2018). Excalibur. http://www.lighttwist.net/wp/excalibur.

McGrath, M. (2014). Target Data Breach Spilled Info On As Many As 70 Million Customers. https://www.forbes.com/sites/maggiemcgrath/2014/01/10/target-data-breach-spilled-info-on-as-many-as-70-million-customers/#2d90e3b2e795.

Morris, D. E., Oakley, J. E., & Crowe, J. A. (2014). A web-based tool for eliciting probability distributions from experts. *Environmental Modelling & Software*, *52*, 1–4.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., & Rakow, T. (2006). *Uncertain judgements: Eliciting experts' probabilities*. Wiley.

Pelteret, M., & Ophoff, J. (2016). A review of information privacy and its importance to consumers and organizations. *Informing Science*, *19*, 277–301.

Petris, G., Petrone, S., & Campagnoli, P. (2009). *Dynamic linear models with R*. Springer.

Redondo, A., Torres-Barran, A., Rios Insua, D., & Domingo, J. (2018). Assessing supply chain cyber risk management. Tech. Report, 1–20.

Tittel, E. (2017). Comparing the top threat intelligence services. https://searchsecurity.techtarget.com/feature/Comparing-the-top-threat-intelligence-services. Accessed: 2019-09-24.

West, M., & Harrison, J. (2006). *Bayesian forecasting and dynamic models*. Springer.

World Economic Forum. (2018). The Global Risks Report 2018. http://www3.weforum.org/docs/WEF_GRR18_Report.pdf.

Chapter 21
Structured Expert Judgement in Adversarial Risk Assessment: An Application of the Classical Model for Assessing Geo-Political Risk in the Insurance Underwriting Industry

Christoph Werner and Raveem Ismail

Abstract For many decision and risk analysis problems, probabilistic modelling of uncertainties provides key information for decision-makers. A common challenge is lacking relevant historical data to quantify the models used in decision and risk analyses. Therefore, experts are often sought to assess uncertainties in cases of incomplete or non-existing historical data. As experts might be prone to cognitive fallacies, a structured approach to expert judgement elicitation is encouraged with the aim to mitigate such fallacies. Further, it enhances the assessment's transparency. An area, in which the assessment and modelling of uncertainties are particularly challenging due to incomplete or non-existing historical data is adversarial risk analysis (ARA). In contrast to more traditional application areas of decision and risk modelling, in ARA intelligent adversaries add more complexity to assessing uncertainties given that their behaviour and motivations can be versatile so that they adapt and react to decision-makers' actions, including actions based on traditional risk assessments. This often inhibits the availability of historical data. This additional complexity is also shown by the challenges that machine learning methods face when informing adversarial risk assessments. As such, using expert judgements for assessing adversarial risk (at least supplementary) often provides a more robust decision. In this chapter, we discuss the importance of structured expert judgement for ARA and present an application of the *Classical Model* as a structured way for eliciting uncertainty from experts on geo-political adversarial risks. We elicit the frequency of terrorist attacks and strikes, riots and civil commotions (SR & CCs), including insurgencies and civil wars, in various global regions of interest. Assessing such uncertainties is of particular interest for insurance underwriting.

C. Werner (✉)
Department of Management Science, University of Strathclyde, Glasgow, UK
e-mail: WernersChristoph@web.de

R. Ismail
Qomplx:Underwriting, Oxford, UK
e-mail: raveem.ismail@raveem.com

A. M. Hanea et al. (eds.), *Expert Judgement in Risk and Decision Analysis*,
International Series in Operations Research & Management Science 293,
https://doi.org/10.1007/978-3-030-46474-5_21

21.1 Introduction

Probabilistic uncertainty modelling is fundamental for decision and risk analysis. It allows for considering the variability in model inputs and the uncertainty propagation onto its outputs. For decision-makers, this information is often of key importance for understanding the robustness of their decisions and actions. Nevertheless, as modellers and analysts building models that inform decision and risk analyses, we commonly face the challenge of lacking relevant, historical data to quantify our models. In this case, and when (in addition) simplifying assumptions on the uncertainties of interest are not sensible, we should use experts' judgements to assess the uncertainties. This can also be a way of quantifying uncertainties when other forms of data gathering are too costly.

In order to consider assessments elicited from experts as scientific data and at the same time ensure they are defensible in front of and transparent for any stakeholders involved in the decision problem, we require a formal process for obtaining expert judgements. Therefore, an elicitation includes the careful definition of the target variables, formulating and pilot testing the elicitation questions, training the experts, eliciting the uncertainty assessments, and analysing together with documenting the elicited results. Further, we need to choose a defensible way for aggregating various assessments as elicitations typically involve multiple experts to capture diverse backgrounds, knowledge and opinions. Clemen and Winkler (1999, 2007) provide an overview of aggregation methods, commonly classified as either *behavioural*, aiming at obtaining a single consensus distribution through group interaction, or *mathematical*, considering analytic ways for yielding one combined distribution from the experts' assessments, usually without expert interaction. In the elicitation presented in this chapter, we use the *Classical Model* for structured expert judgement (Cooke 1991). It provides a mathematical aggregation method that is based on validating experts' assessment performance against empirical data. For overviews and discussions on elicitation processes and their specific elements, see Dias et al. (2018) and Chap. 1.

The area of decision and risk analysis on which we focus in this chapter is *adversarial risk analysis* (ARA). It considers risks which are due to intentional acts of intelligent adversaries and their impact on uncertain outcomes (Rios and Rios Insua 2012; Chap. 7). Therefore, in ARA a lack of relevant historical data is common, in particular due to the versatile and adaptive nature of the risks to be modelled. A particularity, in contrast to many other research areas in risk analysis, is that the complexity of the risks considered poses specific challenges on the more recent advances in machine learning which is why using or at least including human expert judgement is regarded as more reliable for assessments (Cederman and Weidmann 2017). In other words, while in many fields of decision and risk analysis machine learning-based methods, such as expert systems, are used more and more often to assess risk (see for instance Abdelgawad and Fayek (2010) for construction risk, Hadjimichael (2009) for aviation risk, Fares and Zayed (2010) for water supply risk, or as well Idrus et al. (2011) for project risk), in ARA they face several limitations and challenges that human experts can overcome.

The purpose of adversarial risk models is usually to inform counter-terrorism decisions, such as investments and resource allocations for responding to terrorism risk (Rios and Rios Insua 2012). This often involves geo-political considerations and concerns. Traditionally, counter-terrorism intelligence has been available for and used by governmental decision-makers. However, many industries require and invest in similar information today. As such, the industry of interest for this chapter, that of insurance underwriting, is also more and more often in need of rigorous adversarial risk models on geo-political risks. These inform for instance decisions on global insurance portfolios that are possibly impacted by terrorism threats. Therefore, in the elicitation presented in this chapter, our experts are insurance underwriters with an expertise in terrorism analysis.

The objective of this chapter is to explore how the Classical Model works within ARA by presenting one of its first applications for geo-political adversarial risk, in particular with regard to the availability of suitable seed questions for calibrating experts' performance on terrorism events and the general acceptance of SEJ elicitation by experts and decision-makers in this domain.

The remainder of this chapter is structured as follows. The next Sect. 21.2 provides more background on ARA problems, their foci and the role structured expert judgement can play for improving these. This section also contrasts human experts to machine learning approaches in adversarial contexts. Section 21.3 presents some recent developments in the insurance industry due to large-scale terrorism risk given that this is the industry from which our experts come from and for which the elicitation is done. In Sect. 21.4, we then outline our elicitation protocol together with the seed and target questions, before in Sect. 21.5 we present the elicitation results. Lastly, Sect. 21.6 provides a discussion on alternative seed questions for elicitation in geo-political ARA and their availability before we conclude the chapter in the final Sect. 21.7.

21.2 Adversarial Risk Analysis and Structured Expert Judgement

In recent years, there has been an increased interest in advanced analytical methods and models that consider uncertain events and outcomes triggered or are at least affected by intelligent opponents who intend to cause harm and about whose behaviour, actions, motivations and utilities we have imperfect information. This research area is often referred to as ARA. Structured expert judgement and machine learning methods, both face particular challenges when used for adversarial risk which determine their different opportunities for enhancing models in this area.

21.2.1 Brief Background on Adversarial Risk Analysis

Loosely, ARA combines traditional probabilistic risk analysis (PRA) with game-theoretic methods (Roponen and Salo 2015).

The traditional methods and models evolved from the need to assess risk when uncertain outcomes are due to chance (nature) directly without the inclusion of intelligent adversaries. While they have been proposed to be used directly for assessing adversarial risks, e.g. by Ezell et al. (2010), at the same time their use has also been criticised, for instance, by Brown and Cox (2011) and Cox (2009). One of the main potential issues is that an attacker's decision rule for selecting a target is dynamic and as such might be even informed by the anticipated defender's assessment of targets' likelihoods. In this way, a defender's initial assessment of the most likely to be attacked target(s), which as a result obtains most defence resources, has now zero probability of being attacked given that the defender's PRA informs the attacker's choice. This can also happen if the attacker cannot access the defender's assessment directly but rather anticipates his way of thinking. Therefore, traditional risk analysis tools, such as influence diagrams and probabilistic reasoning, are extended for adversarial problems. Examples are Pinker (2007), who uses influence diagrams for informing the supply of countermeasures to terrorism, Merrick and McLay (2010), applying decision trees for modelling the instalment of sensors for screening cargo containers under threat of terrorists, and Parnell et al. (2010), modelling terrorists' objectives for biological weapon usage with decision trees.

Similar to using traditional PRA methods on their own, considering only game-theoretic approaches can also be problematic. For these, min-max solutions, i.e. ones in which both opponents seek to minimise their expected maximum losses across all actions available to them, might lead to sub-optimal solutions (Roponen and Salo 2015). This is due to the attacker and defender not respecting the min-max rationality principle whereas modelling such rational solutions requires particular strong assumptions on the common knowledge available to both opponents (Kadane and Larkey 1982). For instance, the worst possible outcome can have such a low probability that (in reality) it is not considered at all (Roponen and Salo 2015). ARA does not need such strong assumptions on the knowledge of opponents' aims and resource capabilities (Roponen and Salo 2015) and Banks et al. (2011) provides an overview on how classical game-theoretic approaches compare to ones modified for use in ARA.

21.2.2 Structured Expert Judgement for Adversarial Risks

In order to understand the role of structured expert judgement for assessing adversarial risk and hence for enhancing ARA models, we first note briefly how adversarial aspects have been integrated in some more recent definitions on risk. A main advent of new risk definitions that include adversaries followed the terrorist attacks on the USA in September 11th, 2001 (9/11) (Haimes 2009). For overviews see Aven and

Guikema (2015), Aven and Krohn (2014) or Aven (2012). As such, Garrick (2002), for instance, extends the common, quantitative risk definition by Kaplan and Garrick (1981), based on the triplet $\langle s_i, p_i, x_i \rangle$ of i scenarios, their probabilities and outcomes, by a threat (outcome) likelihood as the conditional probability of a successful attack given that the attack is planned.

When using expert judgement for adversarial risk, this altered definition together with the discussion on ARA models shows that experts face more complex uncertainties. This is why it is often necessary to consider a decomposition of the assessments. For example, Paté-Cornell and Guikema (2002) propose assessing a probability of an attack through modelling an attacker's objective from the viewpoint of a defender first before the attacker's probabilities and utilities are assessed through point estimates. In a similar way, expert judgement is used in the *Probabilistic Terrorism Model* by Risk Management Solutions (RMS[1]) to assess target selection probabilities, capabilities of attack modes and attacks' overall probabilities. Here, experts consider the attackers' motivations, resources and capabilities together with defenders' vulnerabilities for an assessment. This shows how experts need to be able to assess probabilities by taking into account the aims, knowledge and skills of attackers as well as defenders. See Willis et al. (2007) for a more detailed discussion of the model.

Similarly, Chap. 22 suggests that experts should assess the probability of operational success and failure, conditional on terrorists' technical capabilities and their modus operandi. He recommends that thereby enhancing our understanding of terrorists' technical capabilities and the modus operandi is what we should use experts for, while highlighting that some (other) uncertainties of terrorism events cannot be expected to be assessed. He provides an example of a failed terrorist attack on an Algerian gas plant due to an accidental cut of the power supply, which ultimately prevented the plant to explode, a contingency we cannot expect to be reliably assessed.

Such decompositions can comprise a lot of information to elicit and therefore their elicitation needs to be well-structured or otherwise we need to make assumptions on the information that is considered by experts for making an assessment. Further, this underlines the importance of other elements in an elicitation process, such as structuring experts' knowledge and beliefs prior to the quantitative elicitation as well as the training of experts. This is similarly the case for SR & CC events.

21.2.3 Machine Learning Methods for Adversarial Risks

This additional complexity of assessments not only affects human experts but also machine learning approaches which are being developed for assessing uncertainties. This is an important aspect to consider given that in particular the recent focus on the terms "data analytics" and "big data" has resulted in an increased interest

[1] RMS, founded at Stanford University in 1989, provides services in the area of catastrophe modelling for (re-)insurers.

in more applications of machine learning methods to do uncertainty assessments in risk analyses. However, Cederman and Weidmann (2017) provide an overview of the challenges that machine learning methods, such as neural networks and expert systems, face when used for predicting political violence and terrorism events whereas it is noteworthy that several of these challenges are less crucial for or can even be overcome by human experts. This is despite more recent machine learning methods having become more reliable at conflict prediction than earlier prediction models, often based on linear regression. For example, remaining challenges are geo-political variations of borders and territories as well as changing power of actors and their, by definition, rule-breaking behaviour. These significantly impede the ability to obtain suitable training data necessary for machine learning methods. Further, even if techniques, such as data scraping from online sources, generate vast data-bases to be used for training machine learning methods, it has been shown that only the quantity of conflict data alone does not enhance prediction accuracy, often due to additional noise. Rather, we need to consider the quality of our information. In this regard, sources like news reports on political violence seem to be stronger predictors than other, more conventional predictors of conflict, such as level of democracy. However, the potential issue with these is that for secondary sources the level of observed violence depends either on the level of actual violence or the probability of reporting, or both of them. Human experts on the other hand can infer knowledge and beliefs about causal mechanism and broader patterns about future changes of power relationships among geo-political actors and hence decide how much of historical data they take into account. In this way, human experts might even guide machine learning models given that they provide insight into the amount and type of information they use for an assessment. The advantage of explanation for certain assessments also enables decision-makers to make more informed decisions. That is, even if a machine learning method offers highly accurate predictions, a black-box model might not be usable in high-risk situations. Therefore, Subrahmanian and Kumar (2017) suggest that experts should be used to propose relevant independent variables that are included in a data set and explain predictions through corresponding narratives of their domain to enhance the understandability of predictions.

21.3 Recent Developments in Insurance Underwriting Due to Risk of Terrorism and SR & CC Events

We already established that while ARA might be of interest in a variety of industries, an industry in which a rigorous approach to quantifying and modelling adversarial risk is particularly key is insurance underwriting. In non-life insurance, so-called *low frequency-high impact* events are by definition observed only rarely and as such a main concern is the lack of relevant historical data for model quantification. Of main interest with regard to *non-natural perils* are terrorism events (Woo 2002; Chap. 22). In addition to the previous brief outline, Parnell et al. (2010), Enders and

Sandler (2009) and Chap. 22 provide overviews about models and research issues for terrorism risk analysis.

The pricing of terrorism risk in insurance has traditionally not been assessed from actuarial principles. Instead, it has been covered by the supply and demand balance in the insurance market while adjustments have been made based on less formal risk selection from site surveys Woo (2002). In the United States, for example, terrorism coverage was included in standard commercial insurance policies as an unnamed peril as part of all-risk commercial and private coverage for property and contents (Michel-Kerjan and Pedell 2006).

The more recent loss developments however led to the necessity of approaching its risk assessment more rigorously. A major turning point for the insurance industry and the reason for an increased focus on terrorism risk were the 9/11 attacks on the United States. These attacks caused an estimated monetary loss up to 60 billion US dollars whereas this amount is spread across various lines of business, such as property insurance, business interruption insurance and workers' compensation (DeMey 2003). Further, on a global scale, the 15 terrorist attacks with the highest casualty numbers have all happened since the year 1982 whereas many more near-miss events occurred which could have ranked among these (Michel-Kerjan and Pedell 2006). In this context, the relationship between the frequency of attacks and their severity can be modelled by a power law (Clauset et al. 2007; Clauset and Woodward 2013; Spagat et al. 2018), a finding similarly provided for war sizes already by the British polymath Lewis Fry Richardson (Richardson 1948). This means that attack severities several orders of magnitude larger than the mean can be common. This (global) development of terrorism risk through an increase in the number of frequencies and in severities underlines the urgent need for improved assessment methods for insurance underwriters.

21.4 Elicitation Protocol and Presentation of Seed and Target Questions

While the complete elicitation protocol of this study can be found in Werner (2017), in this section we briefly outline the main aspects of our elicitation. The method used for this elicitation is the Classical Model (Cooke 1991). Hence, a particular focus here is on the seed and target questions. For detailed overviews and introductions, see the original reference and more recently, Quigley et al. (2018) and other chapters in this book presenting the Classical Model.

After having introduced the experts to the Classical Model and provided them with training on assessing probabilities through quantiles and on the interpretation of the framing of our questions, we proceeded with eliciting first the seed and then the target questions.

In this study, we used both, predictions and retrodictions, as seed questions. The former are seed questions on variables which are about the future but will become

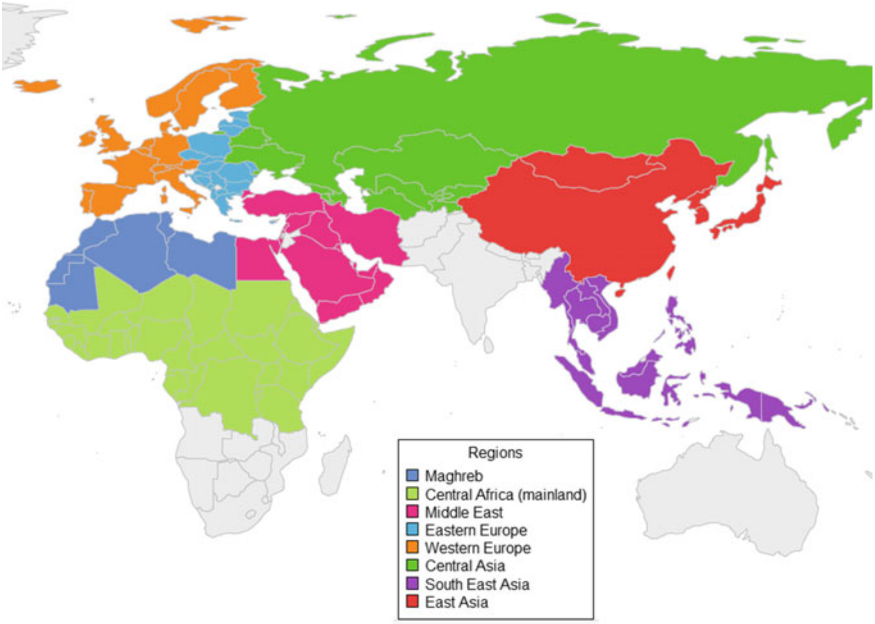


Fig. 21.1 Regions of interest for seed and target questions

known during the time frame of the study. The latter are seed questions on previously observed events (Quigley et al. 2018). Further, all of our seed questions are domain ones which means that they are from the same field of expertise as the target variables. Domain-specific predictions are usually seen as the ideal seed questions (Quigley et al. 2018).

In order to assess the global risk of terrorist attacks, we elicited expert judgements on the frequencies of terrorist attacks in various regions of the world. The regions of interest are shown in Fig. 21.1. For a complete list see Appendix 21.8.

The seed and target questions are formulated in a similar way and exemplary for all 14 of the former, which are about terrorist attacks (another 14 are on SR & CC events), seed questions S01 to S08 are shown below:

> S01 – S03: For a terrorist attack* to be recorded as such, there must be evidence of an intention to coerce, intimidate or convey some other message to a larger audience (or audiences) than the immediate victims.
>
> According to GTD (2016), what was the total number of terrorist attacks (any number of casualties) during the years 2010 to 2015 in the regions of [. . .]

S01 **Maghreb:**

5%ile:__________ 50%ile:__________ 95%ile:__________

- -

S02 **Central Africa (mainland):**

5%ile:__________ 50%ile:__________ 95%ile:__________

- -

S03 **Middle East:**

5%ile:__________ 50%ile:__________ 95%ile:__________

***Terrorist attack** = Any perpetrator group, any weapon type (e.g. biological, chemical, explosive, firearms etc.), any attack type (e.g. armed assault, bombing, facility/infrastructure attack, hostage taking etc.), any target apart from private persons (i.e. business, infrastructure, military, educational/religious institutions, etc.)

S04 – S06: For a terrorist attack* to be recorded as such, there must be evidence of an intention to coerce, intimidate or convey some other message to a larger audience (or audiences) than the immediate victims.

According to GTD (2016), what was the total number of terrorist attacks (any number of casualties) in ***East Asia*** during the time intervals of [. . .]

S04 **1970–1980:**

5%ile:__________ 50%ile:__________ 95%ile:__________

- -

S05 **1990–2000:**

5%ile:__________ 50%ile:__________ 95%ile:__________

- -

S06 **2005–2015:**

5%ile:__________ 50%ile:__________ 95%ile:__________

*__Terrorist attack__ = Any perpetrator group, any weapon type (e.g. biological, chemical, explosive, firearms etc.), any attack type (e.g. armed assault, bombing, facility/infrastructure attack, hostage taking etc.), any target apart from private persons (i.e. business, infrastructure, military, educational/religious institutions etc.)

S07 – S08: Terrorist attacks* are often targeting businesses. According to GTD (2016), of the total number of these attacks during 2010 to 2015, what has been the percentage of attacks targeting businesses in the regions of [. . .]

S07 Western Europe:

5%ile: __________ 50%ile: __________ 95%ile: __________

S08 Eastern Europe:

5%ile: __________ 50%ile: __________ 95%ile: __________

*__Terrorist attack__ = Any perpetrator group, any weapon type (e.g. biological, chemical, explosive, firearms etc.), any attack type (e.g. armed assault, bombing, facility/infrastructure attack, hostage taking etc.), any target apart from private persons (i.e. business, infrastructure, military, educational/religious institutions etc.)

We observe that different formats of seed questions were elicited. Mainly, we asked the experts to assess frequencies of terrorist attacks whereas the region and (range of) years were modified (S01 – S06). In addition, we also elicited seed questions on percentage values for the target types (S07 – S08). The remaining seed questions varied only in that they were either on different years (and ranges), such as the predictive seed questions used, or on the changes in the number of terrorist attacks from one year to another. For seed questions on SR & CC events, the regions, years and targets were similarly varied and formulated in the same framing shown above.

It is important to note that a particularity for eliciting probabilities on adversarial risks, such as terrorist attacks in the above seed questions, is the definition of what constitutes a terrorist attack. This needs to be clarified and pointed out during the elicitation as it defines the probability space of the questions. Therefore, it has been listed for each question on terrorist attacks and is similarly shown for seed questions on SR & CC events.

Following the seed questions, we framed and elicited target questions, $T01 - T08$. These elicit the number of terrorist attacks for the coming year 2017–2018 as the elicitation was done in March 2017. The next eight target questions, T09 – T16, considered SR & CC events.

T01 – T08: How many terrorist attacks (according to the definition in the seed questions) will occur in the coming year (March 2017–March 2018) in the regions of [. . .]

T01 **Maghreb:**

5%ile: __________ 50%ile: __________ 95%ile: __________

⋮

T08 **East Asia:**

5%ile: __________ 50%ile: __________ 95%ile: __________

The elicitation of seed and target questions was held in a plenary format. This means that the experts worked through the questions individually, however, all experts were together for the introduction, motivation and training as well as feedback session by the facilitator. Further, individual expert's questions on clarifications of the questions have been heard by and explained to all experts which ensures they interpret everything in a similar way as best as possible.

21.5 Discussion of Elicitation Results

In total 16 experts participated in the elicitation, all with similar backgrounds and experiences as professionals in terrorism risk modelling and analysis in insurance underwriting. One expert is additionally also an academic in the field.

In this section, we present how the experts performed in the elicitation with regard to the Classical Model metrics for statistical accuracy and informativeness, and discuss the properties of the resulting aggregated judgement, the so-called Decision Maker (DM), and for comparison the equal weighting combination (EW). The seed questions create the basis for identifying the optimal performance-based weighting of experts which can then be used for combining experts' assessments on the target variables as DM. The EW combination is simply the average of all experts' assessments.

Following the seed questions presented (exemplary for all) in the previous section, Figs. 21.2, 21.3, 21.4, 21.5 and 21.6 show the experts' judgements for these. In addition to each expert's uncertainty range over the variable of interest per question, each figure includes the EW combination together with the performance-based DM weighting. The left-hand side of each horizontal line shows an expert's and the combined judgement's 5th quantile assessment, the right-hand side of the line the

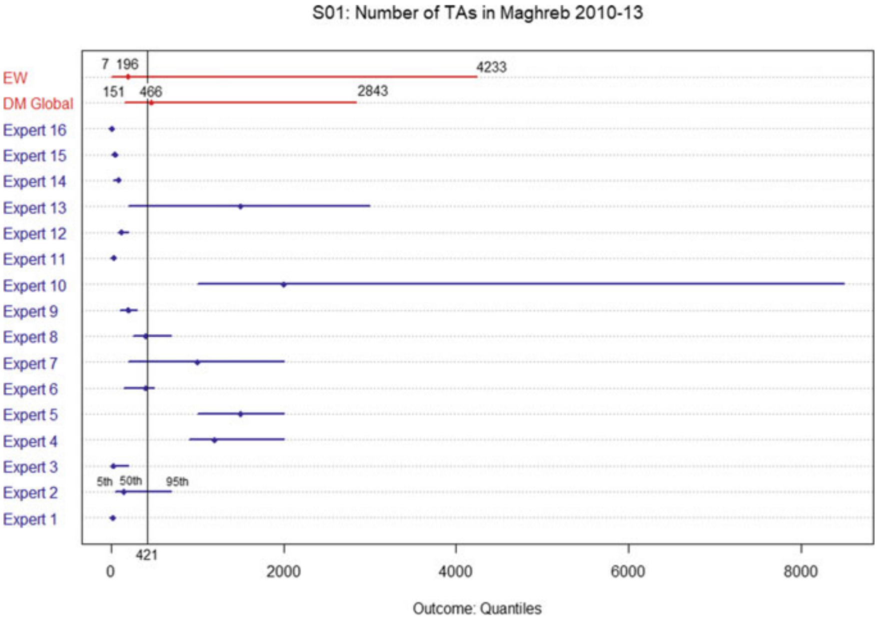


Fig. 21.2 Seed question on number of terrorist attacks in the Maghreb region, 2010–2013 (S01)

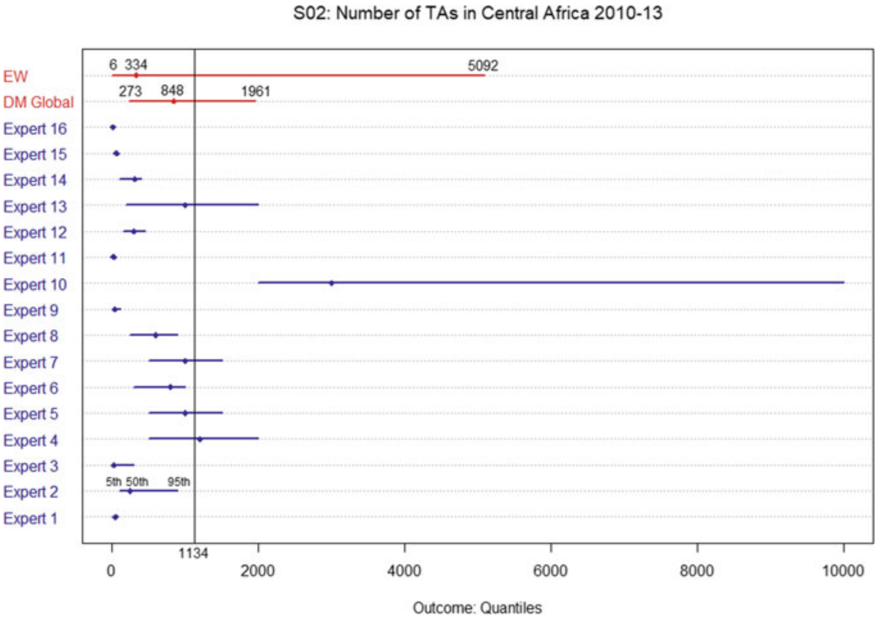


Fig. 21.3 Seed question on number of terrorist attacks in the Central Africa, 2010–2013 (S02)

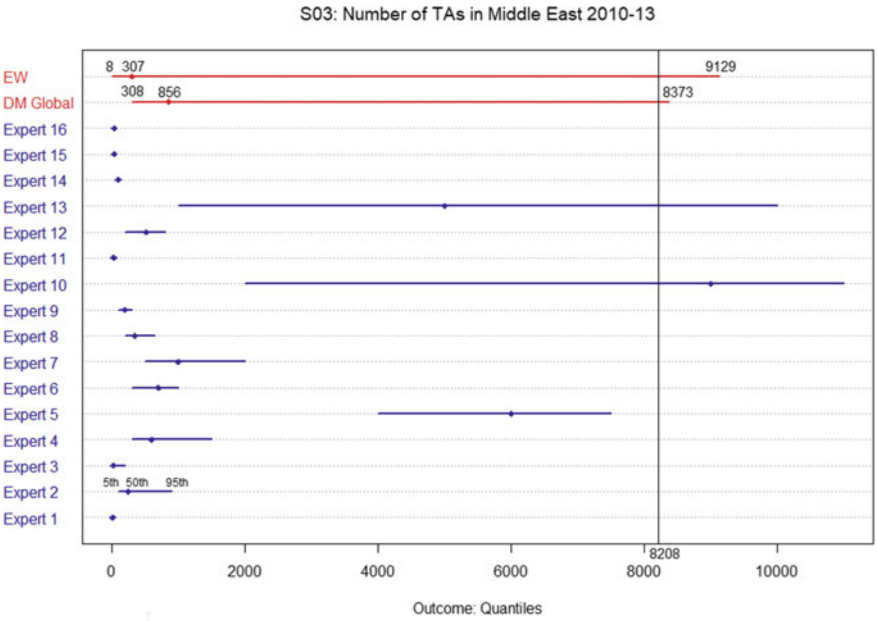


Fig. 21.4 Seed question on number of terrorist attacks in the Middle East, 2010–2013 (S03)

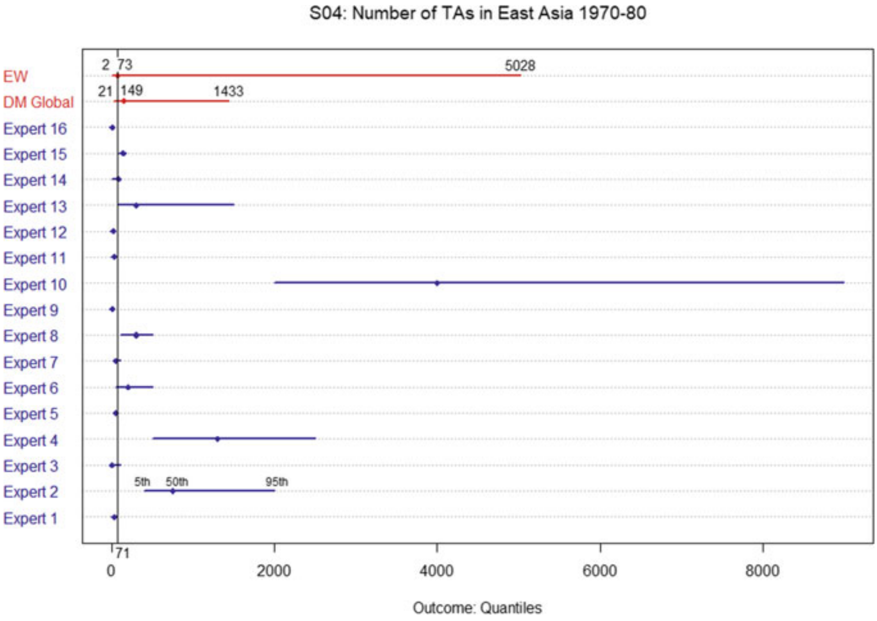


Fig. 21.5 Seed question on number of terrorist attacks in the Middle East, 2010–2013 (S04)

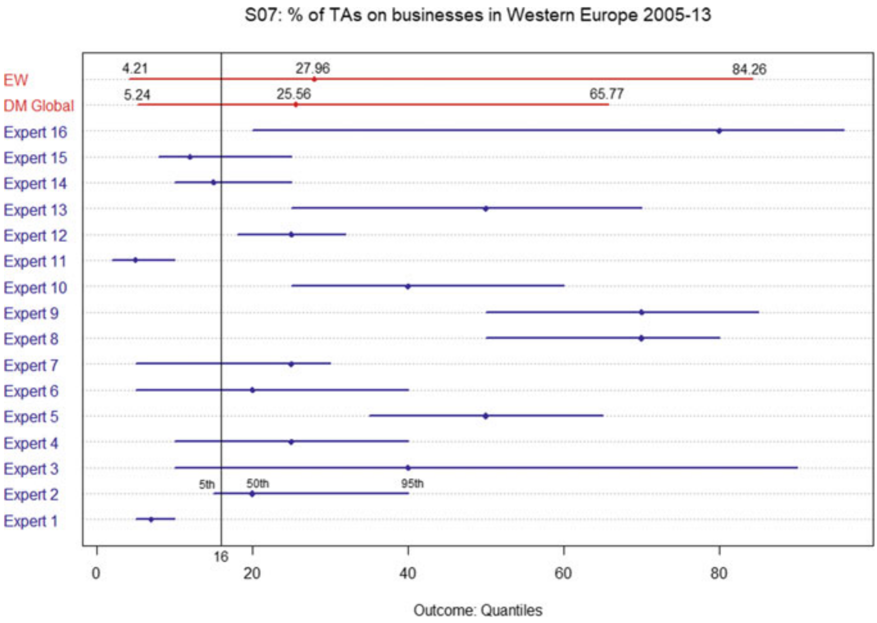


Fig. 21.6 Seed question on percentage of terrorist attacks in Western Europe, 2010–2013, targeting businesses (S07)

95th quantile and the median is given by the dot between the two ends. This is shown exemplary for all assessments for *Expert 1*'s distribution. The realisation is shown through the vertical line.

For the first three seed questions on terrorist attacks' frequencies, S01 – S03 (Fig. 21.2, 21.3 and 21.4), we observe that most experts' assessments are within the same range and that most distributions are narrow. In other words, we see that most experts are confident in their assessment. Nevertheless, several experts (*Expert 1, 3, 9, 11, 12, 14, 15* and *16*) do not include the actual realisation in any of these three questions as a result.

In contrast, *Expert 10* provides for all three questions large uncertainty bounds. Nevertheless, the realisation is missed for the first two, S01 – S02 (Fig. 21.2 and 21.3), and only includes it for the third seed question, S03 (Fig. 21.4).

The remaining experts adjust their assessments more often for each question and include the realisation more often.

The fourth seed question, S04 (Fig. 21.5), is part of a set of questions modifying the year range, in which terrorist attacks happened, for a particular region, in this case *East Asia*. We observe that most experts include the realisation for this 10-year period, even though several experts' assessments have, again, narrow uncertainty ranges.

Seed question S07 (Fig. 21.6) is exemplary for the questions which elicit the percentage of attacks that aim at specific target types, in this case *businesses*. We see

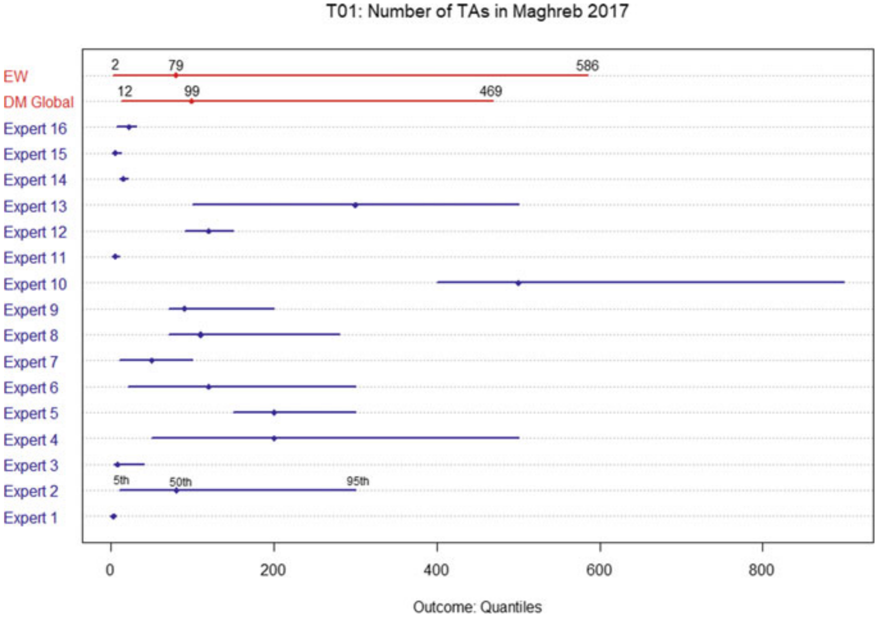


Fig. 21.7 Target question on number of terrorist attacks in the Maghreb region, 2017 (T01)

that the assessment of seven experts includes the realisation whereas one of these experts (*Expert 3*) is not informative in her/his assessment due to the wide uncertainty bounds given.

Next, Figs. 21.7 and 21.8 show the experts' assessments of the target questions together with the aggregated results (EW and DM) exemplary for the first and last target question, T01 and T08. All other target question results (for terrorism risk) are provided in Appendix 21.9.

Considering that the magnitudes on the horizontal axis change for each figure, we can see that the assessments are overall the most informative for *East Asia*. In the complete overview, we observe that they are also informative for *Eastern Europe*, *Central Asia*, *Western Europe* and *South East Asia*. This means that the experts overall are more confident about their prediction with regard to these regions. In contrast, the uncertainty is highest (again, among all experts) for the regions of *Middle East*, *Central Africa* and *Maghreb*.

Across the experts, we observe that similarly to the earlier seed questions (S01 – S03) the same expert (*Expert 10*) provides the widest uncertainty ranges with other ones (e.g. *Expert 4*) providing similarly uncertain judgements only for certain regions. Some experts (*Expert 1, 11, 14, 15* and *16*) consistently give narrow distributions for the target questions whereas their assessments are also the narrowest for the seed questions.

This difference in the experts' uncertainty ranges has implications on the aggregated results. As such, we see that for all target variables the performance-based

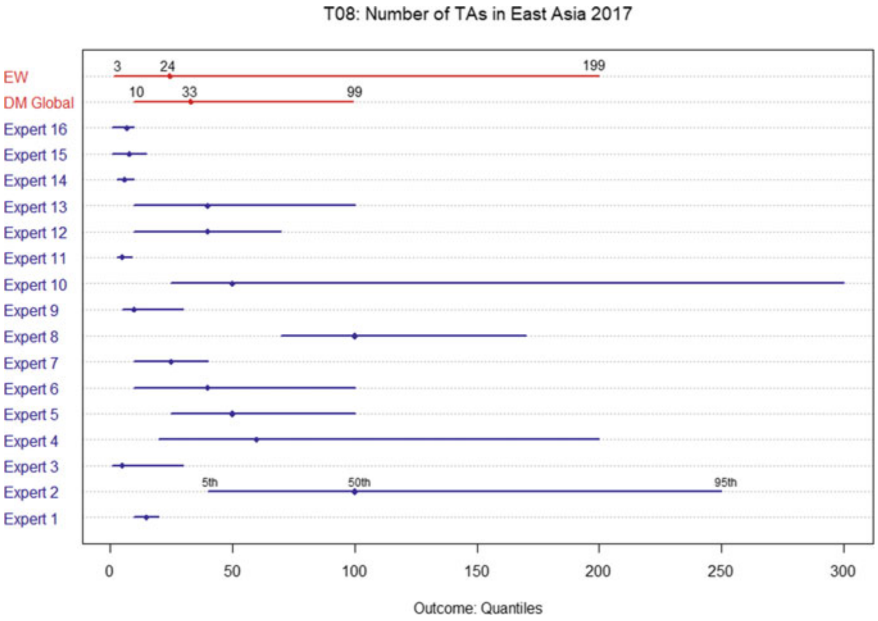


Fig. 21.8 Target question on number of terrorist attacks in the East Asia, 2017 (T08)

combination is more informative than the equal-weighted result. For all but one target question (T03 on the *Middle East*) this difference might be even regarded as considerable. The resulting median assessments of both types of combinations on the other hand are mostly in agreement. This is a frequently observed benefit of the Classical Model (Quigley et al. 2018), i.e. that the performance-based combination typically yields pooled assessments which are more informative than the result obtained by equally weighting judgements while being at least as statistically accurate.

21.6 Alternative Seed Questions for Adversarial Risk Problems

The seed questions used in the elicitation all consider the number of terrorist attacks directly or are based on that, for example, in form of a percentage or change (yearly difference). Nevertheless, in the dry-run of the elicitation, five additional, alternative seed questions were still included. While these were not used further in the later elicitation nor the weighting of experts, they served to test out other seed question types. This is important as we have less experience with using expert judgement methods and the Classical Model in adversarial risks contexts and there is indication that it is cognitively more complex to assess. In future applications, we might take the findings of this section as a basis for developing robust seed question when adversaries are a consideration.

The alternative seed questions were mainly on (1) potential contributing factors of terrorist attacks and SR & CCs which would be commonly included in models, e.g. as exogenous variables and (2) factors and conditions impeding a terrorist attack or SR & CC event.

Regarding the first, several research findings suggest that a relation of climate change to geo-political risks exists (Burke et al. 2014; Barnett 2018) (even though opposing views are also worth mentioning, such as Salehyan 2008 and Theisen et al. 2013). Therefore, a first alternative seed question on this relationship was as follows:

A meta-analysis of studies that examine populations in the post-1950 era suggests that there is a clear statistically significant influence of climate on modern conflict (Burke et al. 2014). Large potential changes in precipitation and temperature regimes are projected for the coming decades with locations throughout the inhabited world expected to warm by +2 to +4 standard deviations (SDs) by 2050.
According to Burke et al. (2014) analysis, what would be the percentage increase in the median frequency of intergroup conflicts due to a +1 SD change in climate toward warmer temperatures?

5%ile: ________ 50%ile: ________ 95%ile: ________

With a similar reasoning, the potential impacts of climate change in the form of resource scarcities are also commonly linked to geo-political risks, mostly with regard to water and food (Hendrix and Brinkman 2013). Hence, another alternative seed question concerned the number of food riots in certain regions of the world over specific time periods.

These alternative seed questions were regarded as cognitively complex, in particular the first one including standard deviations, while the link to geo-political risks was judged as not clear enough. In future, it might be still worth trying out more seed question of this kind, however, new findings in the relevant literature need to be included and possibly new training and framing methods should be considered.

A particular aspect of terrorism in this regard is *stochastic terrorism*. It is commonly defined as acts of violence by random extremists (often “lone wolfs”), motivated and ultimately triggered by political demagoguery in the mass media (Keats 2019; Hamm and Spaaij 2017). Keats (2019) provides the example of US president Trump *tweeting* a video of himself smashing the CNN logo which the Trump fan Sayoc might took as a motivation for supposedly mailing a pipe bomb to the broadcaster’s headquarters. That is, while the attackers are not directly guided, nor provided with resources, to commit terrorist attacks, their attacks are motivated by messages in the media (whether intended as such or not). In other words, they are individually unpredictable, however, their motivating events can be observed and considered similarly as the above as contributing factors.

The seed questions on factors and conditions impeding the success of a terrorism attack or SR & CC event were on the number of military capacities of certain countries and state unions together with the capabilities of the respective national intelligence agencies.

While the connection to geo-political terrorism and SR & CCs risk was clear, in future for these seed question types to be more useful, we should consider two key principles of terrorism risk modelling that stem from the role of security and which Woo (2017) discusses in more detail.

The first is that "target substitution displaces terrorism threat". As terrorist will choose the easier of two similar targets, all terrorist targeting is relative and increasing security efforts for one possible target will often increase the likelihood of other, similar targets. As such, we cannot elicit the likelihood of one particular target in isolation. This is important when eliciting terrorism risk for specific targets on a local scale, for instance, a certain city and its main focal points of infrastructure or places of publicity relevance, but it might be also extended to the global level we have been looking at in this elicitation. That is, for the regions of Fig. 21.1 we need to consider whether additional security efforts have an impact on making other regions more attractive for attacks or whether the terrorist groups active in one particular region only focus on these locally without an interest or the resources for diverting to other countries (targets).

The second principle is that "terrorists follow the path of least resistance in weaponry". Similar to the previous principle, in an elicitation it is important to consider whether an increase in one target's security makes other targets more attractive due to less resistance. Again, this might be extended to the spatial level of this elicitation.

We should include the above principles, for example, by decomposing seed questions, on the resource capabilities and on terrorists' responses to likelihoods of defender actions, in order to account for the relative nature of targeting.

Both types of alternative seed questions will be important in future elicitations on adversarial risk and show the importance of closely following new developments and findings in modelling of terrorism and SR & CCs events. This will ensure that future possible seed questions are suitable and capture experts' knowledge on adversarial risk appropriately.

21.7 Conclusions

In this chapter, we have presented and discussed an expert judgement elicitation for geo-political risks. The adversarial nature of these risks poses a particular challenge for experts and hence their quantification. This study shows one of the first applications of structured expert judgement for adversarial risk and as such we point out several learnings from it to conclude the chapter.

First of all, we have seen that it is sensible to apply the Classical Model in adversarial settings, in particular as appropriate seed questions, a core element of the method can be found even for these types of problems. Overall, the experts' performances on the seed questions show that we can identify experts onto whom we can base the performance-based combined assessment sensibly to yield a more informative (while statistically accurate) distribution for our target variables than achieved with an equal-weight aggregation. When applying the Classical Model within a new application area, it might be problematic if no sensible seed questions can be found for which the experts feel comfortable making assessments or for which we obtain only poor calibration and informativeness scores.

A consideration for future elicitations on terrorist attacks, but also SR & CC events, with regard to the seed questions is the aforementioned importance of defining our events of interest appropriately. Some of our experts provided feedback that they agreed with our definitions of terrorist attacks and SR & CC events, however, also pointed out that other ones are possible and depending on these an assessment can vary considerably. An example is whether we consider only terrorist attacks with casualties or also ones without them. Depending on the region, the former might be considerably lower than the latter.

For some regions, such as the Middle East, we have seen that most experts provide wider uncertainty bounds. In these cases, it might be of interest to include a more rigorous structuring part of experts' knowledge and beliefs about future scenarios in future elicitations. In a related elicitation on the dependence between these regions' frequencies of terrorist attacks (Werner et al. 2018), we have used a structuring method prior to a quantitative elicitation. While the method used is for dependence assessments through exploring conditional scenarios, a similar method could be used also when eliciting marginal distributions (at least for regions with higher uncertainty).

When not only considering the frequency of terrorist attacks and SR & CCs but also the severities in future elicitations it is important that we account for the fat-tailed distributions, often approximated by a power law. This can provide further challenges for experts, however, if dealt with in a structured manner, expert judgements provide an important source of information in particular when, e.g. machine learning methods do not have enough training data (Werner et al., 2017).

Lastly, our experts had all similar experiences by working in the same industry for several years. When eliciting uncertainty from experts on adversarial risk, it might enhance the elicitation results and the discussion thereof if including other types of experts, such as terrorism experts from academic institutions or journalism.

Acknowledgments We would like to thank the editors for the feedback on the previous version of this chapter and Professor Willy Aspinall for the support during the elicitation presented in this chapter together with its preparation and analysis.

21.8 Appendix 1

Detailed list of regions from seed and calibration variables

In detail, the regions of interest for the seed and target questions are as follows:

Maghreb: Algeria, Libya, Mauritania, Morocco, Tunisia

Central Africa (mainland): Benin, Burkina Faso, Burundi, Cameroon, Central African Republic, Chad, Congo, Djibouti, DR Congo, Equatorial Guinea, Eritrea, Ethiopia, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Ivory Coast, Kenya, Liberia, Mali, Niger, Nigeria, Rwanda, Senegal, Sierra Leone, Somalia, South Sudan, Tanzania, Togo, Uganda, Western Sahara

Middle East: Bahrain, Egypt, Iran, Iraq, Israel, Jordan, Kuwait, Lebanon, Oman, Qatar, Saudi Arabia, Syria, Turkey, United Arab Emirates, Yemen

Eastern Europe: Albania, Bosnia-Herzegovina, Bulgaria, Croatia, Cyprus, Czech Republic, Estonia, Greece, Hungary, Latvia, Lithuania, Macedonia, Moldova, Montenegro, Poland, Romania, Serbia (and Montenegro), Slovakia, Slovenia

Western Europe: Austria, Belgium, Denmark, Finland, France, Germany, Iceland, Ireland, Italy, Luxembourg, Malta, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, United Kingdom

Central Asia: Armenia, Azerbaijan, Belarus, Georgia, Kazakhstan, Kyrgyzstan, Russia, Tajikistan, Turkmenistan, Ukraine, Uzbekistan

South East Asia: Brunei, Cambodia, East Timor, Indonesia, Laos, Malaysia, Myanmar, Palau, Papua New Guinea, Philippines, Thailand, Vietnam

East Asia: China, Japan, Mongolia, North Korea, South Korea, Taiwan.

21.9 Appendix 2

Target variables elicitation results of other regions

The other region's target variable elicitation results are (Figs. 21.9, 21.10, 21.11, 21.12, 21.13 and 21.14):

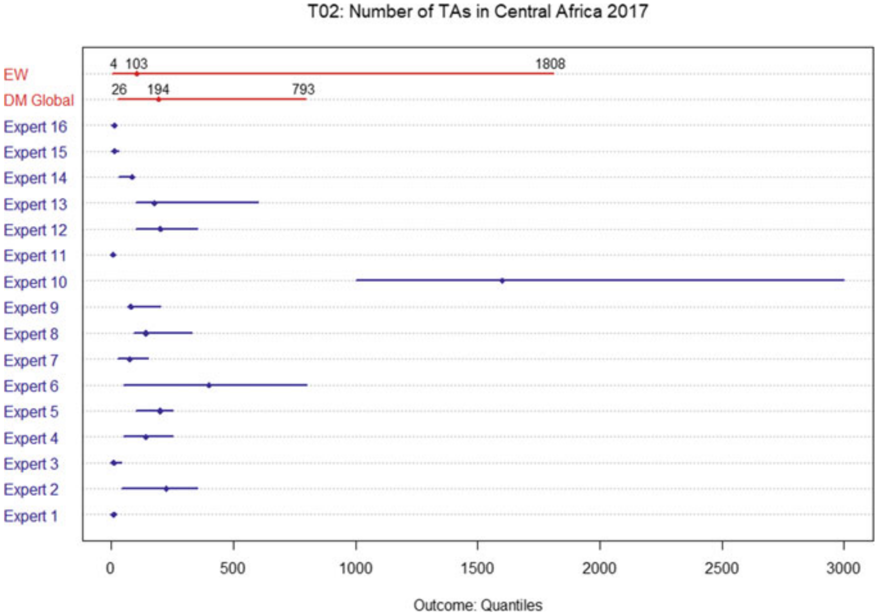


Fig. 21.9 Target question on number of terrorist attacks in Central Africa, 2017 (T02)

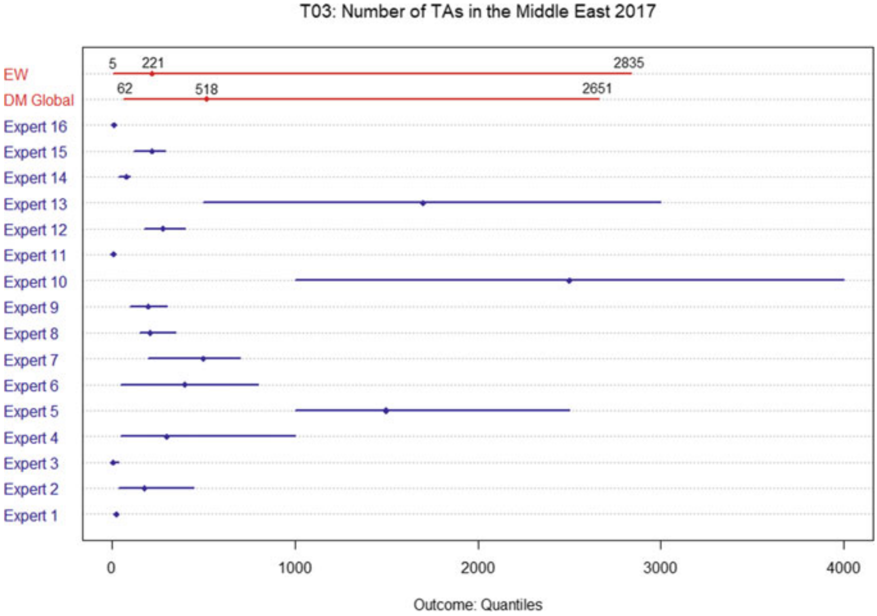


Fig. 21.10 Target question on number of terrorist attacks in the Middle East, 2017 (T03)

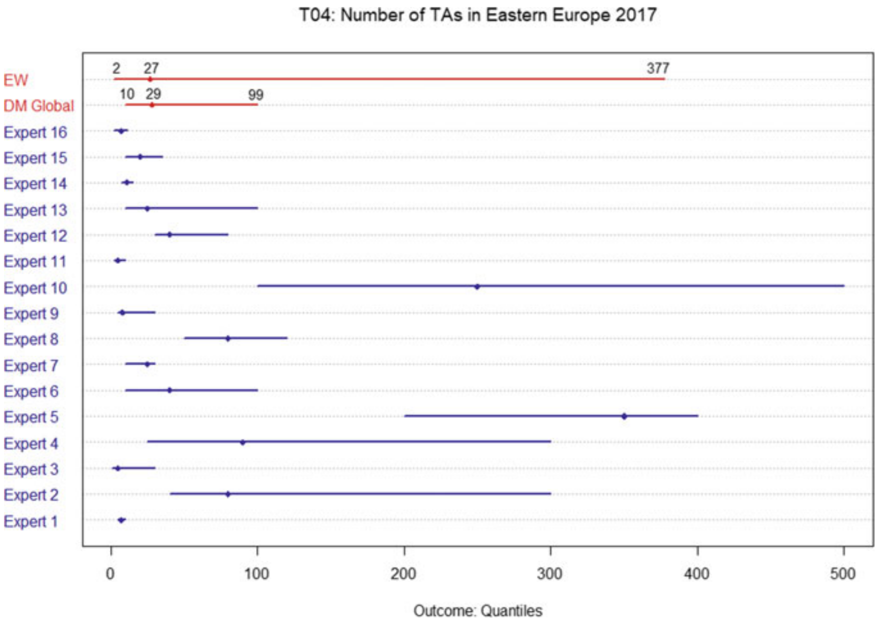


Fig. 21.11 Target question on number of terrorist attacks in Eastern Europe, 2017 (T04)

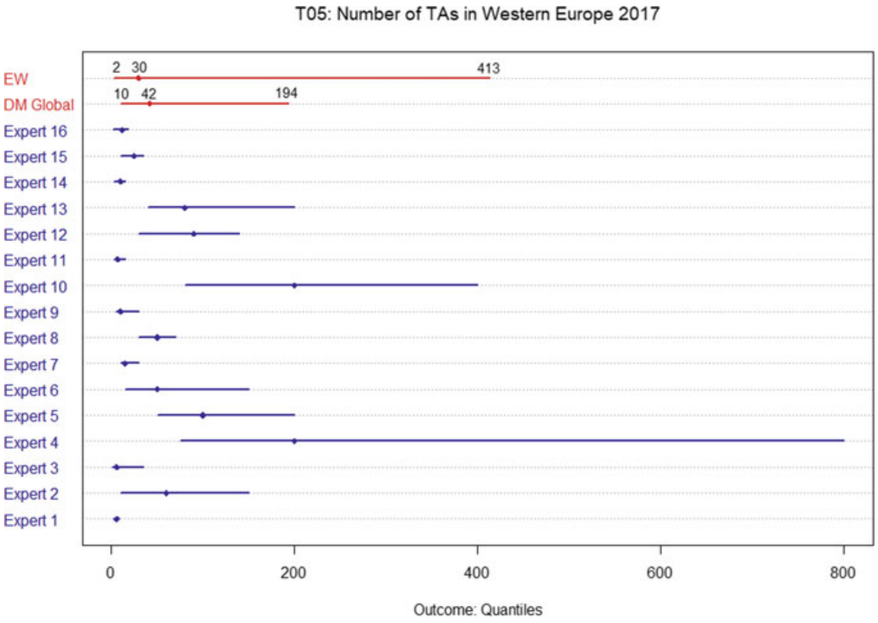


Fig. 21.12 Target question on number of terrorist attacks in Western Europe, 2017 (T05)

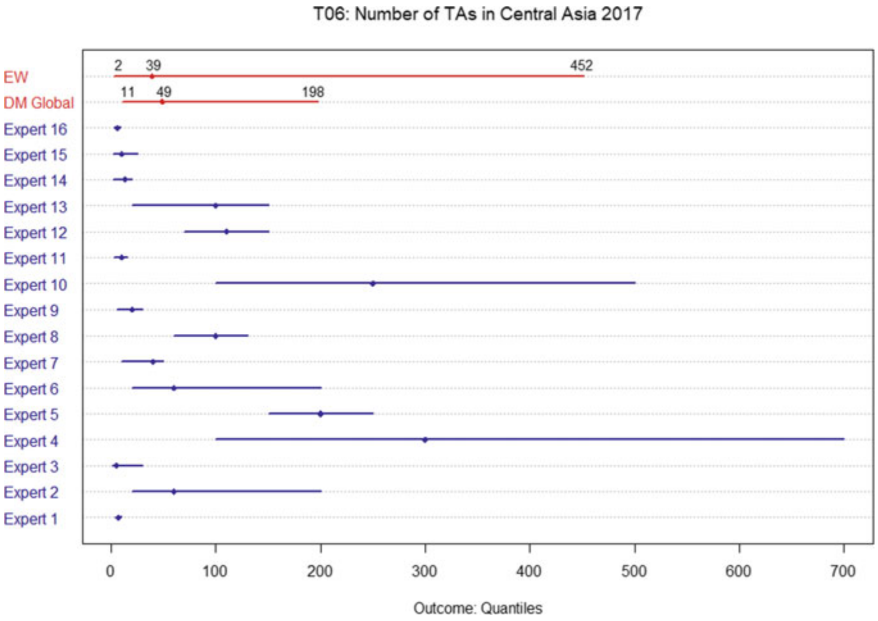


Fig. 21.13 Target question on number of terrorist attacks in Central Asia, 2017 (T06)

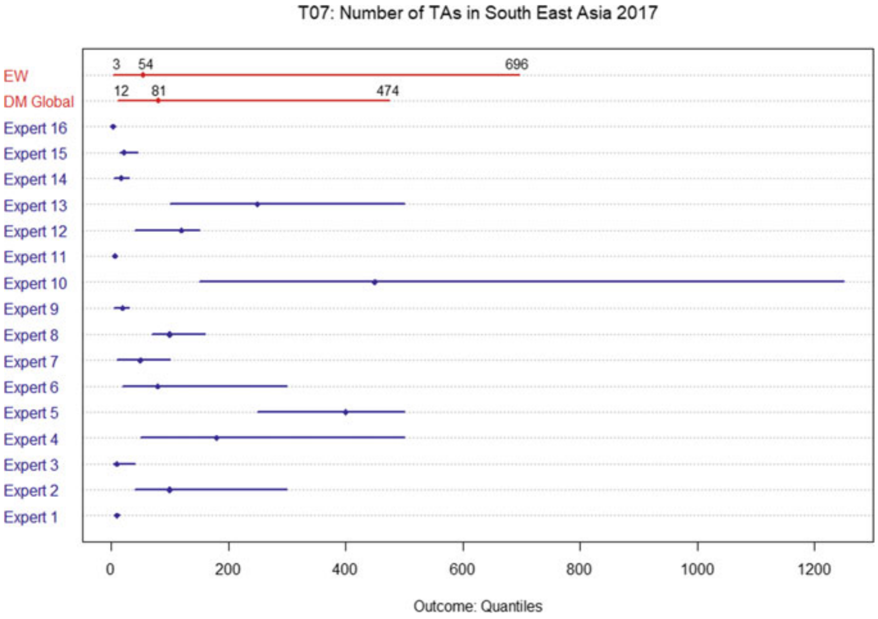


Fig. 21.14 Target question on number of terrorist attacks in South East Asia, 2017 (T07)

References

Abdelgawad, M., & Fayek, A. R. (2010). Risk management in the construction industry using combined fuzzy FMEA and fuzzy AHP. *Journal of Construction Engineering and Management*, *136*(9), 1028–1036.

Aven, T. (2012). The risk concept-historical and recent development trends. *Reliability Engineering & System Safety*, *99*, 33–44.

Aven, T., & Guikema, S. (2015). On the concept and definition of terrorism risk. *Risk Analysis*, *35*(12), 2162–2171.

Aven, T., & Krohn, B. S. (2014). A new perspective on how to understand, assess and manage risk and the unforeseen. *Reliability Engineering & System Safety*, *121*, 1–10.

Banks, D., Petralia, F., & Wang, S. (2011). Adversarial risk analysis: Borel games. *Applied Stochastic Models in Business and Industry*, *27*(2), 72–86.

Barnett, J. (2018). Global environmental change I: Climate resilient peace? *Progress in Human Geography*. https://doi.org/10.1177/0309132518798077.

Brown, G. G., & Cox, L. A, Jr. (2011). How probabilistic risk assessment can mislead terrorism risk analysts. *Risk Analysis*, *31*(2), 196–204.

Burke, M., Hsiang, S. M., & Miguel, E. (2011). Climate and conflict. National Bureau of Economic Research Working Paper Series: Working Paper 20598. http://www.nber.org/papers/w20598.

Clauset, A., Young, M., & Skrede-Gleditsch, K. (2007). On the frequency of severe terrorist events. *Journal of Conflict Resolution*, *51*(1), 58–87.

Clauset, A., & Woodard, R. (2013). Estimating the historical and future probabilities of large terrorist events. *Annals of Applied Statistics*, *7*(4), 1838–1865.

Cederman, L. E., & Weidmann, N. B. (2017). Predicting armed conflict: Time to adjust our expectations? *Science*, *355*(6324), 474–476.

Clemen, R. T., & Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis*, *192*, 187–203.

Clemen, R. T., & Winkler, R. L. (2007). Aggregating probability distributions. In W. Edwards, R. F. Miles, & D. Von Winterfeldt (Eds.), *Advances in Decision Analysis: From Foundations to Applications* (pp. 154–176). Cambridge: Cambridge University Press.

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

Cox, L. A, Jr. (2009). Improving risk-based decision making for terrorism applications. *Risk Analysis*, *29*(3), 336–341.

De Mey, J. (2003). The aftermath of September 11: the impact on and systemic risk to the insurance industry. *The Geneva Papers on Risk and Insurance-Issues and Practice*, *28*(1), 65–70.

Dias, L. C., Morton, A., & Quigley, J. (2018). Elicitation: The Science and Art of Structuring Judgement. Springer International Series in Operations Research and Management Science, 261, Cham.

Enders, W., & Sandler, T. (2009). *The political economy of terrorism*. Cambridge: Cambridge University Press.

Ezell, B. C., Bennett, S. P., Von Winterfeldt, D., Sokolowski, J., & Collins, A. J. (2010). Probabilistic risk analysis and terrorism risk. *Risk Analysis*, *30*(4), 575–589.

Fares, H., & Zayed, T. (2010). Hierarchical fuzzy expert system for risk of failure of water mains. *Journal of Pipeline Systems Engineering and Practice*, *1*(1), 53–62.

Garrick, J. B. (2002). Perspectives on the use of risk assessment to address terrorism. *Risk Analysis*, *22*(3), 421–423.

Global Terrorism Database. (2016). *National Consortium for the Study of Terrorism and Responses to Terrorism (START)*. Retrieved from https://www.start.umd.edu/gtd.

Hadjimichael, M. (2009). A fuzzy expert system for aviation risk assessment. *Expert Systems with Applications*, *36*(3), 6512–6519.

Haimes, Y. Y. (2009). On the complex definition of risk: A systems-based approach. *Risk Analysis*, *29*(12), 1647–1654.

Hamm, M. S., & Spaaij, R. (2017). *The age of lone wolf terrorism*. New York: Columbia University Press.

Hendrix, C., & Brinkman, H. J. (2013) Food insecurity and conflict dynamics: Causal linkages and complex feedbacks. *Stability: International Journal of Security and Development, 2*(2), 1–18.

Idrus, A., Nuruddin, M. F., & Rohman, M. A. (2011). Development of project cost contingency estimation model using risk analysis and fuzzy expert system. *Expert Systems with Applications*, *38*(3), 1501–1508.

Kaplan, S., & Garrick, B. J. (1981). On the quantitative definition of risk. *Risk Analysis*, *1*(1), 11–27.

Kadane, J. B., & Larkey, P. D. (1982). Subjective probability and the theory of games. *Management Science*, *28*(2), 113–120.

Keats J (2019) Jargon watch: The rising danger of stochastic terrorism. Wired, February 2019.

Merrick, J. R., & McLay, L. A. (2010). Is screening cargo containers for smuggled nuclear threats worthwhile? *Decision Analysis*, *7*(2), 155–171.

Michel-Kerjan, E., & Pedell, B. (2006). How does the corporate world cope with mega-terrorism? puzzling evidence from terrorism insurance markets. *Journal of Applied Corporate Finance*, *18*(4), 61–75.

Parnell, G. S., Smith, C. M., & Moxley, F. I. (2010). Intelligent adversary risk analysis: A bioterrorism risk management model. *Risk Analysis*, *30*(1), 32–48.

Paté-Cornell, E., & Guikema, S. (2002). Probabilistic modeling of terrorist threats: A systems analysis approach to setting priorities among countermeasures. *Military Operations Research*, 5-23.

Pinker, E. J. (2007). An analysis of short-term responses to threats of terrorism. *Management Science*, *53*(6), 865–880.

Quigley, J., Colson, A., Aspinall, W., & Cooke, R. M. (2018). Elicitation in the classical model. In: L. C. Dias, A. Morton, & J. Quigley (eds) Elicitation: The science and art of structuring judgement. 15–36. Springer International Series in Operations Research and Management Science, 261, Cham.

Richardson, L. F. (1948). Variation of the frequency of fatal quarrels with magnitude. *Journal of the American Statistical Association*, *43*(244), 523–546.

Rios, J., & Rios Insua, D. (2012). Adversarial risk analysis for counterterrorism modeling. *Risk Analysis*, *32*(5), 894–915.

Roponen, J., & Salo, A. (2015). Adversarial risk analysis for enhancing combat simulation models. *Journal of Military Studies*, *6*(2), 82–103.

Salehyan, I. (2008). From climate change to conflict? No consensus yet. *Journal of Peace Research*, *45*(3), 315–326.

Spagat, M., Johnson, N. F., & van Weezel, S. (2018). Fundamental patterns and predictions of event size distributions in modern wars and terrorist campaigns. *PloS One*, *13*(10), 1–13.

Subrahmanian, V. S., & Kumar, S. (2017). Predicting human behavior: The next frontiers. *Science*, *355*(6324), 489–489.

Theisen, O. M., Gleditsch, N. P., & Buhaug, H. (2013). Is climate change a driver of armed conflict? *Climatic Change*, *117*(3), 613–625.

Werner, C. (2017). Geopolitical risk assessment: A structured expert judgement elicitation - Elicitation Protocol and Results. *Data Set University of Strathclyde, UK, dx.* https://doi.org/10.15129/a0029a02-4283-4bc8-a671-ee5dbe680dfa.

Werner, C., Bedford, T., Cooke, R. M., Hanea, A. M., & Morales Nápoles, O. (2017). Expert judgement for dependence in probabilistic modelling: A systematic literature review and future research directions. *European Journal of Operational Research*, *258*(3), 801–819.

Werner, C., Bedford, T., & Quigley, J. (2018). Sequential refined partitioning for probabilistic dependence assessment. *Risk Analysis*, *38*(12), 2683–2702.

Willis, H. H., LaTourrette, T., Kelly, T. K., Hickey, S., & Neill, S. (2007). Terrorism risk modeling for intelligence analysis and infrastructure protection. RAND Corporation Technical Report, Vol. 386, Santa Barbara.

Woo, G. (2017). *Understanding the principles of terrorism risk modeling from the attack in Westminster*. London: Risk Management Solutions Discussion Paper.

Woo, G. (2002). Quantitative terrorism risk assessment. *The Journal of Risk Finance*, *4*(1), 7–14.

Chapter 22
Expert Judgement in Terrorism Risk Assessment

Gordon Woo

Abstract Since 9/11, the probabilistic risk assessment of losses from terrorism has formed a quantitative basis for informed terrorism risk management. An irreducible element is the elicitation of expert judgement. In any application domain, the reliance on expert judgement can be minimized through the establishment of core conceptual principles, such as economic game theory and adversarial risk analysis, which govern the risk phenomena under consideration. For non-state threat actors, such as the Jihadi groups, Al Qaeda and ISIS, their limited logistical resources compared with western counter-terrorism intelligence and law enforcement capacity, greatly constrain the spectrum of their operations, which can be modelled quite reliably in a probabilistic manner. However, state-sponsored terrorism poses a much more severe challenge, especially in connection with the use of weapons of mass destruction, such as nuclear and chemical weapons. In this paper, the fundamental principles of terrorism risk assessment are reviewed, and the use of expert judgement is illustrated in relation to state-sponsored nuclear and chemical weapon deployment.

22.1 Introduction

Terrorism is asymmetric warfare between opponents of contrasting military capability. The German general, Helmuth von Moltke, openly declared that 'in war, everything is uncertain'. Famously, he wrote that no plan of operation extends with certainty beyond the first encounter with the enemy's main strength. This has become a universally accepted tenet of warfare. In contrast with the deterministic game of chess, the Prussian military invented board games with dice to introduce an aleatory element.

The outcome of chess tournaments is open to speculation and wagering. But imagine the challenge of trying to forecast the outcome of a chess match between two grandmasters, where some moves were decided by the throw of dice. Knowledge

G. Woo (✉)
RMS, 30 Monument Street, London EC3R 8NB, UK
e-mail: Gordon.Woo@rms.com

A. M. Hanea et al. (eds.), *Expert Judgement in Risk and Decision Analysis*,
International Series in Operations Research & Management Science 293,
https://doi.org/10.1007/978-3-030-46474-5_22

of the chess playing styles, strengths and tournament records of the adversaries would inform the expert judgement of the forecasters, but this would be tempered by the added aleatory component.

The Roman general and historian, Julius Caesar, noted that in war, events of importance are the result of trivial causes. Sometimes, the outcome of a terrorist attack is as unlikely as throwing a series of sixes. On 16 January 2013, Jihadis armed with light weapons attacked the InAmenas gas plant in Algeria, operated by Statoil. The plant should have been blown up, but for a stray celebratory terrorist bullet from an AK-47 that accidentally cut the plant power supply and shut down operations. Counterfactually, without this remarkable fortune, the gas plant could have been destroyed. Contingencies such as at InAmenas cannot be forecasted, but the likelihood of operational success and failure can be estimated, based on the technical capability of a terrorist organization, and most importantly its modus operandi. Understanding terrorist modus operandi is like knowing the rules of chess.

22.1.1 Dependence on Human Behaviour

Like all human activities, individual idiosyncrasies of human behaviour, (such as firing an AK-47 in the air), will manifest themselves in the actions of terrorists, but there are some important over-riding factors that govern terrorist behaviour to a considerable extent. In the case of Jihadis, Islamic law is a powerful controlling influence on their terrorist actions. In Arabic, the word for rationality is *aqlaniyyah*, which is an expression of the total basis upon which a person acts (Rauf 2015). For Muslims, this basis must be derived from the ethics, philosophy and traditions of the Islamic religion.

It is often noted that, prior to launching a terrorist attack, Jihadis will immerse themselves in readings from the Qur'an, with the firm assurance of paradise for those who are martyred. In the Qur'an (9:111), it states: '*Allah hath purchased of the believers their persons and their goods; for theirs in return is the garden of Paradise. They fight in His cause, and slay and are slain*'.

One of the resolute long-term ambitions of Jihadis is to bring about an Islamic state. Because such a state would not espouse the same values as a liberal democracy, attempts to coerce western nations through violence lead to acts of terrorism. The characteristics of such terrorism depend much less on individual human behaviour than on a common general religious belief system, and so are far more predictable.

Furthermore, just as the threat of legal sanction constrains the behaviour of criminals, so law enforcement services and security forces constrain the behaviour of terrorists. These are especially tough constraints within the well-funded English-speaking Five Eyes security alliance of USA, UK, Canada, Australia and New Zealand. In these countries, there are tight restrictions on access to bomb-making material, and elaborate plots are very likely to be disrupted.

Besides classified information on the terrorist threat, there is also classified information on counter-terrorism activities. Some information of this kind can be privately accessed through attending annual closed intelligence and terrorism meetings, or discussions under the Chatham House Rule. Other important sensitive information has been publicly disclosed in large volumes by the NSA whistleblower Edward Snowden in June 2013 (Harding 2014). This unauthorized disclosure confirms that the principal agent for counter-terrorism control is massive electronic surveillance and acquisition of communications meta-data, involving multiple contact chaining of terrorist suspects. The details of this surveillance were hitherto classified, but nonetheless have been deliberately leaked into the public domain and so can inform terrorism risk assessment.

We shall argue for and illustrate the potential of structured expert judgement procedures in the development of terrorism risk assessment, noting several areas in which it can be usefully applied. The confidential nature of the context means that our discussion will be general without detailed SEJ case studies.

22.2 Principles of Terrorism Risk Modelling

Terrorist resources of finance, manpower and weaponry are much less than that are available to nation states, so they have to be deployed in an optimally effective and efficient manner. Essentially, excessive effort should not be expended in the short term to achieve their long-term objectives. Extravagant use of resources can doom a terrorist organization to oblivion. The general principle of least action is a guiding principle of the fundamental way that the universe works. This has been expressed in a contemporary fashion by Coopersmith (2017) in the title of her book: ‘the lazy universe’. Terrorists are lazy in the sense that they are work-averse; there is no point in doing more work than is necessary to advance their goals. Attack strategies of nation states may involve wanton expenditure of multi-billion dollar armament budgets, but terrorists cannot afford profligacy. Terrorists need to be frugal with their resources; achieving high leverage, which is the ratio of attack impact to cost. This is exemplified by 9/11: the leverage for this Al Qaeda attack was approximately 100,000, which is the ratio of the economic loss impact of $50 billion to the comparatively modest operational cost of $500,000. The requirement of high leverage is a major input factor in terrorist attack modelling.

Terrorism is the language of being noticed. This can be achieved in the simplest way through a knife attack in a location with high name recognition. In U.K., where access to firearms and bomb-making ingredients is restricted, knife attacks have higher leverage. London Bridge, a popular landmark with high name recognition, was the location of terrorist knife attacks on 3 June 2017 and 29 November 2019. In both cases, fake suicide bomb belts were worn by the Jihadis. Their terrorist goals were well met without the actual need to make real suicide bomb belts, which might have been well beyond their resources and capabilities, and even patience.

In the case of the London Bridge attacks, the targets were defenceless civilians. These were the softest of targets. More generally, terrorists may decide to attack harder targets which have security weaknesses. There is little to be gained by attacking well-defended targets, when there are vulnerable targets available. Defenders will seek to reduce their vulnerability by improving security in various affordable and practical ways.

The adversarial nature of terrorism and political violence is captured within the methodology of game theory, which addresses the strategic interactions between opposing groups. The behavioural aspects of these interactions are accounted for in behavioural game theory (Camerer 2003), and also adversarial game theory (e.g. Rios et al. 2012; Rios Insua et al. 2009; Banks et al. 2015).

22.2.1 Target Substitution

A direct application of game theoretic principles is in the terrorist substitution of targets according to security levels. A common misjudgement about terrorist targeting is that everything is a potential target. This misjudgement arises from the impression that the mind of a suicide terrorist is irrational and that a Jihadi martyr is deranged. However, a decision by a Jihadi to kill himself in the course of a terrorist attack is not irrational within the religious system of belief that paradise awaits a martyr. This is a modern twenty-first century version of Pascal's wager, which is a probabilistic cost-benefit argument for believing in God, despite doubt and scepticism over the existence of God.

The concept of terrorist target substitution applies at all spatial geographical scales: national, city, and building level. At a national level, British Jihadis angered at U.S. foreign policy, may be deterred by U.S. border security from attacking the U.S. homeland, and choose to attack U.K. instead. This is what happened in the London transport bombings of 7 July 2005. At a city level, when there was a police cordon around London, IRA bombers turned around and drove north to bomb England's second city, Manchester, instead. At a building level, Chechen black widows have switched building targets in central Moscow at the last moment if extra security was observed at the original target.

The principle of terrorist target substation underlies the widespread concern over the multiplicity of soft targets in western countries. The more obviously attractive targets are hardened commensurately with their perceived value to a terrorist organization. It is no longer possible to drive vehicles within close bombing distance of the most attractive urban bombing targets, such as principal government buildings. Accordingly, instead of bombing the U.K. parliament, which has long been a Jihadi aspiration, the soft London underground was targeted in 2005. Progressively, since 9/11, the security community has diminished the range of vulnerable targets that might be of interest to terrorists in their attack planning. As a consequence, there has been a progressive reduction in the range of targets against which plots have been organized. As vulnerable targets have been hardened, they have been substituted by softer targets.

22.2.2 Terrorist Weaponry

Terrorists tend to be work-averse and follow the path of least resistance in their choice of weaponry. Off-the-shelf light military weapons, such as guns, are a common choice for terrorists in countries with ready firearm access. In June 2019, Ashiqul Alam, a young Jihadi from Queens, was arrested for plotting to attack Times Square in New York City, using guns and hand grenades to kill police officers and civilians. Individual lone actors like him would be capable of making a significant impact just using conventional weapons. This would not be the case if more ambitious weaponry were considered, or if there were experimentation with advanced weapons of mass destruction. Indeed, even if Ahiqul Alam had been a member of a large terrorist cell, the possibility of deploying a sophisticated, innovative and dangerous weapon would have been extremely remote.

One class of weapons which is coming within reach of work-averse terrorists are Unmanned Aircraft Systems, commonly known as drones. The technology of drones is advancing rapidly. Terrorists need not have the technical capability to construct drones; they have become relatively easy to acquire and operate. On 4 August 2018, two drones equipped with a kilogram of plastic explosives were used in an assassination attempt on President Nicolas Maduro of Venezuela. Powerful smart drones are now a viable attractive option for transporting and delivering payloads ranging from small packages, such as with the Venezuela attack, to heavy cargo, with weight measured in hundreds of kilograms.

A drone would be capable of transporting an improvised explosive device, but this would be less impactful than delivering a chemical or biological payload into a crowded space. Such an attack might have serious lethality consequences, as terrorist organizations well understand, even if they lack operational capability. ISIS propaganda posters have depicted a drone attack on the Eiffel Tower in Paris and in Manhattan.

The terrorist interest in exploiting drone technology is manifested in the Middle East from ISIS drone raids in Iraq and attacks on Saudi targets from the Houthi Islamic militia in Yemen. Terrorists have always been eager to learn from battlefield experience of weaponry. The military battlefield is a traditional testing ground for new terrorist weapons. Drones have been used on the battlefield and what is used on the battlefield will eventually be adapted for terrorist usage. Indeed, terrorist plots have been thwarted that could have involved drone technology. In Manchester, England, an ISIS supporter was developing a drone with the intention of launching a drone attack on an army barracks. However, for a lone actor, there remain significant technical challenges and obstacles in the adaptation of drones for killing people.

22.2.3 Severity of Weapon Attack Modes

The terrorist payoff from an attack depends on the severity of the weapon attack mode. In a tough counter-terrorism environment, the more ambitious a weapon that is selected, the more time, logistical resources and personnel that will be required to achieve operational functionality. The probability distribution of weapon attack modes has a long severity tail.

For the IRA in their terrorist campaign to bring about a united Ireland, the killing of British soldiers and Ulster constabulary was self-legitimatized by the armed struggle. However, the murder of civilians was disfavoured for political and religious reasons, on both sides of the Irish border. By contrast, Jihadis have absolutely no qualms about mass murder, indeed they have an explicit intent to kill civilians.

On 22 May 2017, a Libyan refugee brought up in Manchester, Salman Abedi, detonated a backpack bomb at the entrance to the Manchester Arena concert hall, after a sell-out concert by the American superstar Ariana Grande. Twenty two of her fans, mostly young girls, were killed. On the morning after the terrorist attack, the UK Prime Minister, Theresa May, declared: '*It is now beyond doubt that the people of Manchester and of this country have fallen victim to a callous terrorist attack, an attack that targeted some of the youngest people in our society with cold calculation*'. The Prime Minister added that, '*Although it is not the first time Manchester has suffered in this way, it is the worst attack the city has experienced and the worst-ever to hit the north of England.*'

Included amongst the terrorist outrages suffered by Manchester was the bombing of the Arndale shopping centre on 15 June 1996. Human lives ultimately matter more to society than a shopping mall. Destroyed buildings can be rebuilt in a way that lives cannot. Part of the cold calculation of Salman Abedi was to choose the optimal target for his terrorist attack: a suicide bomber can only die once. Unlike the IRA bombers, who had multiple opportunities for attacking different targets, and ensured they had escape plans for any operation, suicide bombers have just a single opportunity. So the targeting has to be optimal.

For Islamists 'who love death as you love life', society's pain is the terrorist's gain. The greater the pain of bereavement, the greater is the terrorist's sense of gain. The Islamist predilection for killing in gruesome and barbaric ways causes maximal hurt and distress to the western countries attacked. Terrorism is the ultimate devilish act of *Schadenfreude*: rejoicing in the misfortune and suffering of others. The German philosopher, Arthur Schopenhauer, would have recognized his terminology as characterizing the vengeful mindset of Jihadis.

There have been a number of backpack terrorist bombings against the western alliance since 9/11. Although there have been quite a few Jihadi car bomb plots since then, there has yet to be a successful Jihadi car bomb attack against the western alliance. The nearest miss was the Times Square SUV bomb plot by Faisal Shahzad on 1 May 2010, which failed for technical bomb-making reasons. He slipped through the counter-terrorism net, but the great majority of plots are interdicted by counter-terrorism forces.

Before any massive Jihadi bomb of 2 tons or more is detonated in a major western city, there should be some preparatory warning by way of the prior occurrence of a lesser size vehicle bomb plot, possibly as part of a multiple target bombing attack. Indeed, the vehicle plots which have been interdicted since 9/11 have all been car bomb plots. There have been no truck bomb plots. In the IRA terrorist campaign for a United Ireland, there was a gradual severity progression in the size of plots, ranging from a small 100 lb car bomb in 1972 to a 3000 lb truck bomb which caused massive damage in Manchester on 15 June 1996.

22.3 CBRN Attacks

The same development time principle for conventional weapons applies to Chemical–Biological–Radiological–Nuclear (CBRN) attacks, which remain an aspiration of Jihadis, but not yet a practical reality. Before any massive CBRN attack, some precursory lesser attack may provide an early warning indicator of increasing terrorist capability and progression on the demanding technical learning curve.

As the anthrax letter scare in Autumn 2001 demonstrated, even a small quantity of anthrax can cause mass terror. The perpetrator of this attack was a bioweapons expert, Bruce Ivins, at the US Army Medical Research Institute of Infectious Diseases. No non-state organization had this dangerous and potent anthrax capability. More generally, only nation states have the technical capability to launch significant chemical, biological or nuclear attacks.

If a terrorist cell has accumulated even a modest quantity of a highly toxic substance, there would be very strong counter-terrorism pressure to deploy it rather than to delay an attack by months to acquire much more. Public fear and mass media coverage would result from even a small CBRN terrorist attack. The law of diminishing returns would apply to the prospective terrorist gain from a more ambitious attack. Operational research methods can quantify the balance between the risk of arrest and the reward of a more potent weapon. Since 9/11, denial of safe terrorist havens for laboratory R&D has meant that not even a minor Jihadi CBRN attack has been witnessed, and there is scant evidence of experimentation and preparation of toxic material.

22.3.1 State-Sponsored Chemical Attacks

In Syria, the Assad regime has used both the nerve agent sarin and chlorine gas as chemical weapons against opponents of the regime. Only nation states have stockpiles of chemical weapons, and these are typically covert in deference to the Chemical Weapons Convention. Here the focus is on state-sponsored terrorism in a foreign country. A notable example of this occurred on 4 March 2018, when a military grade VX nerve agent was deployed on the streets of Salisbury, England.

The target of this chemical poison attack was Sergei Skripal, a former Russian military intelligence officer and MI6 agent. The highest concentration of nerve agent was discovered on the front door of his house. His daughter Yulia who was visiting him from Russia was also contaminated with the lethal nerve agent. The VX nerve agent used was identified by chemical weapons experts at the UK Defence, Science and Technology lab at Porton Down as originating from a group of nerve agents known as Novichok. These agents were developed in an attempt to circumvent the Chemical Weapons Convention, and engineered to be undetectable by standard equipment. Novichok consists of two separate components that, when mixed, become an active nerve agent, and can be easily deployed using an aerosol, spray, liquid or wipe.

Novichok is not a weapon that can be manufactured by non-state terrorists. It requires the highest-grade state laboratories and expertise. Russia has previously produced this agent; indeed, *Novichok* is a Russian word for 'newcomer'. The likely production facility used to manufacture the agent is in Sarov, a closed town in Russia. As is routine with state-sponsored attacks, Russia categorically denied any involvement, even though only Russia had both the capability and the cogent motive for this chemical attack. Indeed, it is known that a list of around a hundred Russian enemies of the motherland has been drawn up by the Kremlin, and they are deemed to be legitimate targets. The British ambassador to the UN, Jonathan Allen, concluded that it was highly likely that Russia was responsible. Nikki Haley, the US ambassador to the UN, called for immediate action against Russia. The timing of the attack seems to have been chosen two weeks before the 18 May Russian election to boost support for Putin as a tough president.

Prior to collapsing in a catatonic state on a park bench on the afternoon of 4 March 2018, the Skripals had visited the nearby Mill pub and Zizzi Italian restaurant in the centre of Salisbury. Public Health England (PHE) issued advice for those who also had visited these establishments to wash their clothes and belongings, and seal off anything that could not be manually cleaned. However, Dr Vil Mirzayanov, a former Soviet Union chemical weapons scientist who developed Novichok, insisted this was insufficient, asserting that Novichok is so powerful that extremely small doses could remain a danger to public health for years. According to him, hundreds of people could be at risk of suffering possible long-term consequences including headaches and loss of coordination (Deardon and Sharman 2018).

To corroborate this fear, there are suggestions that US veterans Gulf War illness, the symptoms of which are long-lasting, may be related to exposure to low-dose Iraqi chemical warfare agents in the 1991 Gulf War (American Heart Association 2010). Because of the limited long-term experience data on such low-dose nerve gas exposure, opinions are divided over Gulf War illness, and also the outcome of Novichok exposure. Dr Jenny Harries, southern region director at Public Health England noted that PHE had been working very closely with the police and national experts on chemical weapons and that their risk assessment was based on knowledge of the chemical used. Her advice remained that the risk to the general public was low. The advice might have been clarified to state explicitly that the potential adverse outcomes from allowing the public access to potentially hazardous areas were sufficiently unlikely as not to warrant mandatory exclusion orders.

What is the probability distribution of the number of people who are liable to suffer long-term health problems in the years ahead?

Such an important question is all the more challenging for probing the frontier of scientific knowledge. Dr. Jenny Harries had stated that the advice given was based on knowledge of Novichok. Informal elicitation of expert judgement may work quite well when the elicitation covers the existing domain of knowledge. However, where the time frame for elicitation is beyond practical experience or reasonable extrapolation, a carefully structured and professionally facilitated approach would be preferable.

22.3.2 Bio-Terrorism

Non-state threat actors do not have the technical capability and laboratory facilities to develop biological weapons. However, they can act as human agents to spread a natural contagion. To compound the pervasive political conflict in the Middle East there is the terrorism risk associated with the deliberate malicious spread of a pandemic in western countries. The use of biological weapons by terrorists has a long history, and has an extensive literature. Ever since 9/11, the threat of Al Qaeda using biological weapons has been taken very seriously. Indeed, for counter-terrorism response, it has been the Pentagon that has funded research into the development of vaccines for plague and Ebola and other pathogens that might be weaponized by terrorists.

Biological weapons are attractive to terrorists drawn to becoming bio-martyrs. The millenarian sect Aum Shinrikyo sent a medical team to the Congo in 1993 to investigate the prospects for weaponizing Ebola. This proved too difficult, because Ebola was not highly contagious. Two years later, they launched a sarin gas attack on the Tokyo subway.

With the deployment of any terrorist weapon, the three factors that need to be taken into consideration to gauge the threat are (1) intent; (2) capability; (3) opportunity. The intent by ISIS and other terrorist groups to use infectious disease as a biological disease is clear from their communications. Their capability to develop their own pathogens is minimal. However, if a lethal and transmissible infectious disease were to emerge, terrorist groups would have ample opportunity of spreading the disease wilfully at public gatherings, or on public transportation. Infectious disease propagates along social networks. Terrorists who spread disease maliciously become supernodes in these social networks. The epidemiological consequence of supernodes is to amplify the effective degree of contagiousness of a virus.

The nexus between political conflict and a global pandemic provides a worrying route to disaster. If an epidemic were to emerge in one of the numerous developing regions in a state of political unrest, civil strife or anarchy, the absence of disease surveillance and fragile public health system could well allow the contagion to become established there and then spread abroad to other continents via refugees with little constraint.

Accordingly, a major global pandemic is a systemic financial risk, being coupled with supply chain breakdowns and business disruption, potentially aggravated by the chaos and disorder of political conflict. In 2014, the emerging Ebola crisis might not have been contained if there had been a civil war in West Africa. Counterfactually, the political situation in Sierra Leone and Liberia might have been as unstable as in the

1990s, when there were civil wars in both countries. In 2015, when a million Syrian refugees migrated to Europe, an emerging pandemic disaster might have arisen had there been a more transmissible mutation of the camel-borne Middle East Respiratory Syndrome (MERS). Amongst these refugees, ISIS supporters would have acted as malicious superspreaders of the disease.

A counterfactual question for expert elicitation is as follows: *In 2015, what was the probability distribution for the number of fatalities from MERS?* To address this question, carefully structured elicitation is required, where the facilitator decomposes it into separate contingencies:

(a) What was the probability of MERS mutating in 2015 to become much more contagious between humans? Much is known about the virology of MERS, its spread within camel populations in the Middle East, and transmission from camels to humans. However, there is substantial uncertainty over the likelihood of a dangerous mutation.
(b) Given that there was a dangerous mutation, what was the joint probability distribution of MERS lethality and contagiousness?
(c) For each realization of lethality and contagiousness, what was the impact of ISIS in maliciously spreading the contagion?
(d) Given the impact of ISIS, what was the probability distribution for the number of MERS fatalities in 2015?

22.4 Subjective Expert Judgement Elicitation Methods

The preceding review of terrorism risk provides the technical subject matter background for a discussion of the role of the elicitation of subjective expert judgement. Terrorism is a pervasive risk that needs to be managed by many professional groups: military, police, government, corporations, insurers etc. As discussed above, the military has their own traditional procedures for dealing with threats, which tend to be suited to their own special skills, experience and training, and not to invoke the methods of quantitative risk assessment. The same holds for the police and law enforcement services, who may not even be familiar with qualitative threat matrices. War gaming and battle simulation incorporate some of the basic features of threat assessment and stochastic modelling, without the formal mathematical apparatus of quantitative analysis.

The most promising areas of application involve potential financial risk associated with acts of terrorism. The risk of insolvency is regulated by financial authorities, and corporations need to be able to quantify extreme tail risks, including terrorism risk. In connection with terrorism risk insurance, since 2002, RMS has conducted group elicitation meetings annually in London and Washington DC with leading global terrorism experts, such as Bruce Hoffman and Rohan Gunaratna, with extensive knowledge of terrorism. The classical method of group elicitation was adopted.

Group elicitation meetings are particularly effective, in comparison with individual elicitation methods, because they allow the sharing of information that may be known only to a subset of the experts in attendance. Apart from confidential information that is not in the public domain, there is restricted classified information that is disseminated only on a need-to-know basis. And even where terrorism information is available from open source material, such material may not necessarily be easy to find, and so may not be familiar to all experts.

These group elicitation meetings have been successful in as much as the terrorism experts have turned out to be well calibrated against what actually transpired. This may be explained by the robustness of the principles governing terrorism risk, which are universal in their domain of applicability. It should be noted that other methods could be tried for eliciting expert judgement from a group of experts, e.g. the Sheffield Elicitation Framework (SHELF) developed by Tony O'Hagan.

There are numerous methods for aggregating expert opinions. The first chapter of this book includes a review. Axiomatic approaches aim to establish an aggregation rule from axioms that the rule should satisfy. Ad hoc approaches have no axiomatic basis, but are proposed with some ex-post justification. One approach that might work well in a terrorism context is a consensus method whereby experts are allowed to interact with each other (Nau 2002) and share information. This is one mode of behavioural aggregation, aimed at generating a greater degree of agreement.

22.5 Terrorism and Political Risk

Terrorism is one manifestation of political conflict. Terrorist campaigns constitute a form of asymmetric warfare, where the terrorist forces are generally far smaller than those of the nation states which they are attacking. A possible exception to the limited capability of terrorist groups is where they are sponsored by a nation state, which provides them with military, economic and technical resources for their terrorist campaigns. Such states include regimes in Iran, North Korea, Somalia, etc. that might be classified by some political risk commentators as failing states.

Whereas terrorism risk is generally bounded by the limited resources of terrorist groups, and persistent counter-terrorism pressure, state-sponsored terrorism risk is limited essentially by international diplomatic pressure, backed up by the threat of direct military conflict. Inevitably, there is a degree of expert judgement in making any risk forecast in the context of military conflict. There are superior methods for eliciting this expert judgement. Important lessons were learned following the intelligence debacle surrounding the 2003 war in Iraq War, where no evidence of weapons of mass destruction could be found, yet senior US intelligence officials remained adamant that Saddam Hussein definitely possessed such powerful weapons.

The massive intelligence failure associated with Iraq War led to a re-evaluation of intelligence assessment methods in Washington, and the establishment in 2006 of the Intelligence Advanced Research Projects Activity (IARPA). The scientific process of randomized control trials can discriminate those with particularly good judgement on political events. Superforecasters can be identified who have special skill in forecasting, as can be measured through a Brier score. It is not necessary to have years of intelligence experience to be good at forecasting political events. Indeed, many who do have such experience are rather indifferent or poor forecasters. Superforecasters have been identified as having some special traits (Tetlock and Gardner 2016). They are typically numerate, with a technical knowledge of Bayes theorem, even if they may not explicitly make their forecasts doing any actual Bayes theorem calculations. Rather, they edge towards the truth by implicitly following the Bayes principle of updating according to the weight of evidence using their own sense of intuition. For any political conflict risk assessment, explicit use of Bayesian methods, including the construction of Bayesian Belief Networks (BBN), would optimize the forecasts made through progressive updating.

22.5.1 The Trump Card

To paraphrase the German general, Helmuth von Moltke, quoted at the start, 'In the Trump White House, everything is uncertain'. The Prussian military invented board games with dice to introduce an aleatory element. For a board game to begin to represent the challenge of dealing with the Trump White House, the rules of the game themselves would need to include an aleatory element. Imagine playing a game of chess, where the number of squares a piece could move was decided by a dice throw. The game of bridge is the quintessential skilful game of chance where the calculation of probabilities is a decisive advantage in playing strategy. But imagine the chaotic implications in playing a game of bridge where any card could be converted to the trump suit on the throw of dice.

All during the Cold War, the possibility existed of a suitcase nuclear device being planted in Manhattan by an operative of the Soviet Union or other hostile foreign government. Such a risk has always been dealt with capably and effectively by the CIA, who were confident of tracking the flow of communications between the sizeable team planning and executing such a major state-sponsored terrorist attack, and nullifying any plot.

The threat of a nuclear weapon state-sponsored terrorist plot against the US has been a serious cause for concern since 9/11. The Al Qaeda leader, Ayman Al Zawahiri, would have absolutely no qualms in deploying such a fearsome weapon. Since 9/11, until the inauguration of President Trump in January 2017, the most likely source of weapons of mass destruction for a terrorist attack against the US homeland was a rogue state. This threat was of course the rationale for the 2003 war in Iraq to depose Saddam Hussein. The risk of North Korea passing over a nuclear device to a terrorist organization for deployment in the USA has been the subject of numerous political think-tank studies (Bunn et al. 2016), incorporating the elicitation of expert judgement on the nuclear threat over a ten-year time horizon. The hostile intent of the North Korean regime is evident from the proliferation of sophisticated cyber attacks by the notorious Lazarus group, which earns a substantial amount of foreign exchange for the Pyongyang regime. However, looking back on these expert judgements on North Korean state-sponsored terrorism, they have turned out to be excessively pessimistic.

The North Korean prolific testing in 2017 of inter-continental ballistic missiles capable of reaching the USA, materially changed the threat of a state-sponsored attack on USA using weapons of mass destruction. If there had been any external attempt to depose the North Korean leader, the response most likely would have been a military attack on South Korea, or on Guam, Hawaii, or the US mainland, rather than a state-sponsored terrorist attack on the US homeland. In the Autumn of 2017, probabilistic risk analyses were undertaken on behalf of US life and health insurers for the potential number of US casualties in Guam in the event of a nuclear strike.

From the North Korean perspective, the belligerence and volatility of President Trump were also a game-changer. The longstanding cautious western policy of strategic patience reinforced the optimality of Kim Jong-Un's strategy of nuclear weapon development. This was a rational response, geared to maintaining Kim's long-term position as the Supreme Leader of North Korea. However, the abandonment of this policy of strategic patience by President Trump in favour of abrasive aggressive confrontation made it rational for Kim to follow the path of dialogue. This path led inexorably to the Singapore summit meeting on 12 June 2018. Irrespective of the slowness in achieving the agreed objective of denuclearization of the Korean peninsula, the likelihood of North Korea supplying a terrorist organization with a nuclear weapon is greatly reduced, provided the USA keeps to its summit obligations.

22.5.2 Trump Betting

This volatility at the heart of Washington decision-making has been a profitable opportunity for the betting markets. President Trump's rise to power was the biggest non-sports event in betting history. One prominent Irish bookmaker, Paddy Power, hired a head of Trump Betting, whose task was to monitor the administration, updating odds and providing bets.

A parallel book of bets has been kept on Kim Jong-Un, the Supreme Leader of North Korea. Amongst these bets have been wagers on his life coming to an end; being removed from office; being overthrown in a coup or resigning. Such political bets are reminiscent of the exploratory terrorism betting market that DARPA piloted in 2003, before it was shut down and castigated as immoral by congress. Any odds offered on the assassination of any named person might be an illegal inducement for someone to place a bet and then carry out the assassination. Terrorist attacking for financial gain is, however, part of the threat landscape. A popular leading German football team, Borussia Dortmund, was targeted with a bomb attack on 11 April 2017 by a financial trader who hoped to profit from puts he placed on the club's stock price. He left deceptive notes suggesting this was a Jihadi attack.

On 8 August 2017, President Donald Trump warned (CNBC 2017) that threats from North Korea 'will be met with fire and fury like the world has never seen'. Irish bookmaker Paddy Power responded by slashing the odds on the possibility of a cataclysmic conflict in 2017 from 500/1 to 100/1. Bets on a statue of Trump being erected in North Korea in 2017 had odds of 66/1, while the likelihood of Kim Jong-un staying on as North Korean leader beyond 2031 were put at 4/7 (CITYAM 2017).

One of the purposes of expert elicitation is to facilitate smarter practical decision-making under uncertainty. If Kim Jong-Un had seen the latter odds of his staying in power for at least another 14 years, or himself commissioned an expert elicitation, he would have realized it was advisable to meet with President Trump. If his policy is America First, President Kim's policy is self-survival, and his policy choices would be those that gave the young dictator a high chance (>90%) of reaching the age of fifty in his presidential office.

In the aftermath of the Singapore Summit, the odds of Kim Jong-Un's survival would have been greatly boosted. This may be inferred from the comparatively short odds of 10:1 soon quoted by PaddyPower on North Korea hosting the Olympic Games before the end of 2040. No country can host the Olympic Games without massive infrastructure expenditure. These short odds reflect the plausibility and promise of major inward investment in the coming two decades, coinciding with potential denuclearization of the Korean peninsula.

In September 2017, the 2024 and 2028 Olympics were awarded to Paris and Los Angeles, respectively, after Tokyo in 2020. In that September, if there had been an expert elicitation on the Olympic Games venues in 2032, 2036 and 2040, the odds of North Korea being selected would have been those for a rank outsider—on economic and infrastructure grounds alone. But as perceived in the immediate aftermath of the Singapore Summit, the odds of the infrastructure investment and development being sufficient by 2032 for North Korea to host the Olympics might be as good as 5:1. Assuming five cities bid for each of the 2032, 2036 and 2040 Games, and that Pyongyang, North Korea, bids each time, the chance of winning one of the awards is about one-half. This yields the overall odds of North Korea hosting the Olympic Games before the end of 2040 at about 10:1, as quoted by PaddyPower after the Singapore Summit.

22.5.3 Expert Political Judgement on the Middle East

The Trump Presidency challenge for the US State Department has been immense and unprecedented. In an interview with the LA Times (2017), Nicholas Burns, a senior State Department official noted that Trump's policy in his first year of office was a radical departure from every president since WWII. The most recent example of US isolation came with Trump's decision to formally recognize Jerusalem as the capital of Israel, reversing decades of international consensus. On Monday, 14 May 2018, the US embassy in Jerusalem was opened, amidst mass protests on the Gaza–Israel border.

The impact on Middle Eastern terrorism of this breach of international consensus is potentially one of the most significant questions on terrorism risk. It seems very unlikely that any formal attempt was made within the White House to gauge the terrorism costs of recognizing Jerusalem as the capital of Israel. This was an uncosted campaign promise. Ex-post, risk stakeholders representing US interests and citizens both at home and abroad must have been assessing potential terrorism consequences. This is a clear threat: on 15 January 2019, the Islamist militant group Al-Shabaab attacked a hotel and office complex in Nairobi, claiming that it was a response to the US recognition of Jerusalem as the capital of Israel.

Group decision conferencing is the traditional framework for terrorism assessment. However, it would be interesting to compare this with a calibrated expert judgement approach. As a reminder of the practical importance of such an exercise in the context of the US recognition of Jerusalem as the capital of Israel, it should not be forgotten that the first attack on the World Trade Center in Manhattan on 26 February 1993 was perpetrated by Ramzi Yousef, who was motivated by the cause of Palestine.

A characteristic of terrorism risk is that the spectrum of expertise is very broad, covering those who have a deep knowledge of terrorist modus operandi and history, such as the 1993 WTC attack. There are experts who have known key members of terrorist organizations; those who may have been members or sympathizers in the past; those who have worked in the intelligence or security services; and those who know or have interviewed currently active terrorists. Just as criminologists interview criminals in prison, terrorism analysts also interview terrorists in prison. The opportunities expanded with the Islamist threat. Between 2002 and 2016, with the rise of militant Islam, the proportion of Muslims in the UK prison population doubled. Williams (2018) noted that more than 40% of the prisoners in the high-security prison he worked in were Muslim.

Open source information, such as provided by Jihadi online publications, also provide valuable insight for terrorism experts, who can infer recommended attack strategies, and the principal drivers of terrorism risk. The large variability in the breadth and depth of terrorism expertise argues against any elicitation procedure that weights experts equally, or treats as equal the opinions of participants in a group decision conference.

22.5.4 Iran

The Joint Comprehensive Plan of Action (JCPOA) is an international agreement on the nuclear programme of Iran eventually reached in Vienna on 14 July 2015, between Iran, the P5+1 (the five permanent members of the United Nations Security Council—China, France, Russia, UK, USA—plus Germany) and the European Union.

President Trump's intense dislike of JCPOA, negotiated during the Obama presidency, presented some major challenges for political pundits forming their expert judgements on the Iranian response to the US withdrawal from JCPOA. Three principal policy options were open to Iran (dispute, leave or continue), and Iranian officials would have been able to offer estimates of the likelihood that each would have been pursued. Under the Chatham House rule, these chances could be obtained through the participation of knowledgeable Iranian officials. Indeed, at a London lecture at Chatham House itself, a question was raised as to what the most likely option might be. Ayatollah Ali Khamenei, previously issued a fatwa against the development of nuclear weapons. This religious ruling would suggest that nuclear terrorism would not be an outcome, whichever option was taken. However, this dogmatic position may be over-ruled by pragmatic Iranian politicians.

Of particular methodological interest is how a formal elicitation of Iranian political pundits would fare by comparison with those with real inside knowledge from Tehran. Unlike elicitations relating to natural or environmental hazards, the answers would actually be known to insiders. Questions where the answers are known might be usefully employed as calibration seeds for a structured elicitation using Cooke's method (Cooke 1991).

The opportunity has not yet arisen for a practical application of Cooke's method to an actual real-time political risk crisis. This exercise might avoid the systematic groupthink associated with traditional decision conferencing, which is liable to be distorted in favour of those who are the most opinionated, have the most forceful personalities, and speak loudest; traits not entirely disassociated from the Trump White House. But whatever the approach taken to elicit expert judgement, Sunstein (2019) draws a lesson from counterfactual analysis that small shifts or nudges can produce massive political changes, such as the 1979 Iranian revolution, which was unforeseen, like the Arab Spring.

References

American Heart Association. (2010). Low-dose exposure to chemical warfare agent may result in long-term heart damage. *Science News*. https://www.sciencedaily.com/releases.htm.

Awan, A. N. (2016). The impact of evolving Jihadist narratives on radicalization in the west. In S. Staffell & A. N. Awan (Ed.), *Jihadism Transformed*. London: Hurst & Co.

Banks, D. L., Aliaga, M. R., Insua, D. R. (2015). *Adversarial risk analysis*. CRC Press.

Bunn, M., Malin, M. B., Roth, M., Tobey, W. H. (2016). *Preventing nuclear terrorism*. Harvard University: Belfer Center for science and international affairs report.

Camerer, C. (2003). *Behavioral game theory*. Princeton, NJ: Princeton University Press.

CNBC. (2017). Trump's fire and fury speech. https://www.cnbc.com/2017/08/08/trump-warns-north-korea-threats-will-be-met-with-fire-and-fury.html.

CITYAM. (2017) http://www.cityam.com/270141/paddy-power-slashes-odds-world-ending-year-north-korea.

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

Coopersmith, J. (2017). *The lazy universe*. Oxford: Oxford University Press.

Deardon, L., & Shardon, J. (2018). Russian spy attack. *Independent*.

Gallagher, C. (2018). *Telling it as it wasn't*. University of Chicago Press.

Harding, L. (2014). *The Snowden files*. London: Guardian Books.

LA Times. (2017, December 26). Trump claims he's boosting U.S. influence, but many leaders see America in retreat. http://www.latimes.com/nation/la-fg-trump-us-influence-20171226-story.html.

Nau, R. F. (2002). The aggregation of imprecise probabilities. *Journal of Statistical Planning and Inference, 105*(1), 265–282.

McNeilly, M. (2001). *Sun Tzu and the art of modern warfare*. Oxford: Oxford University Press.

Rauf, F. A. (2015). *Defining Islamic statehood*. Basingstoke: Palgrave Macmillan.

Rios, J., & Rios, Insua D. (2012). Adversarial risk analysis for counterterrorism modelling. *Risk Analysis, 32*(5), 894–915.

Rios, Insua D., Rios, J., & Banks, D. (2009). Adversarial risk analysis. *Journal of the American Statistical Association, 104*(486), 841–854.

Silber, M. (2012). *The Al Qaeda factor*. Philadelphia: University of Pennsylvania Press.

Sunstein, C. R. (2019). *How change happens*. Cambridge, MA: MIT Press.

Tetlock, P., & Gardner, D. (2016). *Superforecasting*. London: Random House books.

Williams, R. (2018). Muslims leaving prison talk about the layers of their lives. *Horizons* 36.

Werner, C. (2019). This volume.

Woo, G. (2011). *Calculating catastrophe.* Imperial College Press.

Woo, G. (2015). *Principles of terrorism risk modelling from Charlie Hebdo.* Ankara: Defence Against Terrorism Review.

Chapter 23
Decision-Making in Early Internationalization: A Structured Expert Judgement Approach

Michał Zdziarski, Gabriela F. Nane, Grzegorz Król, Katarzyna Kowalczyk, and Anna O. Kuźmińska

Abstract The aim of this chapter is to show how a structured approach to elicit expert judgement (SEJ) can guide the practice of early internationalization. We applied SEJ to forecast some critical issues upon which an innovative start-up wished to base their decision of whether to expand their initial operations in Poland and Czech Republic to Brazil. Sixteen participants of an Executive MBA program acted as experts and underwent the procedure for eliciting their judgements. The performance of experts was quantified in terms of statistical accuracy and informativeness, which were combined to provide a performance-based weight for each expert according to Classical Model. The combination of weighted expert judgements led to improved statistical accuracy and informativeness of the forecast. The procedure demonstrates how entrepreneurs can take advantage of expert knowledge in deciding about risky endeavours when lacking their own experiences and reliable data that can guide their choices.

Keywords Structured expert judgement · Internationalization · Location choice · Forecasting · International new venture

23.1 Introduction

How can international new ventures take advantage of external expertise in their initial location choice decisions? This question is quite fundamental, as new ventures lack the resources to mitigate risks of internationalization, and decisions on location choice largely condition their future fortunes. Earlier literature established that the success of international new ventures largely depends on the unusual composition of competencies and experiences from different national markets in the possession of

M. Zdziarski (✉) · G. Król · K. Kowalczyk · A. O. Kuźmińska
Faculty of Management, Warsaw University, Warsaw, Poland
e-mail: m.zdziarski@uw.edu.pl

G. F. Nane
Delft Institute of Applied Mathematics, Delft University of Technology, Delft, Netherlands

A. M. Hanea et al. (eds.), *Expert Judgement in Risk and Decision Analysis*,
International Series in Operations Research & Management Science 293,
https://doi.org/10.1007/978-3-030-46474-5_23

entrepreneurs and the core management team (Phillips McDougall et al. 1994). As it is unusual to have a relevant constellation of experience agglomerated in the top management team of a new venture, the problem of knowledge sourcing arises for companies that need to internationalize. Using the cumulated knowledge of external experts provides the means to overcome this problem. The need to source knowledge from outside seems especially relevant for new ventures located in the Central and Eastern Europe (CEE) due to very limited chances of having entrepreneurs with relevant personal experience from earlier internationalization projects. It is well established in the literature that outward FDIs from this region were scant before transition, and the process of internationalization have gradually started to emerge after the fall of the Berlin Wall (Ferencikova and Hluskova 2015; Wilinski 2013).

The necessity to internationalize may be present due to the nature of the industry, or limited opportunities to grow business in the home market. If the necessity is there, and entrepreneurs do not have sufficient competences and knowledge, what options do they have? One solution can be to accept affordable risks of losses and experiment with the internationalization process, applying effectuation logic (Sarasvathy 2001). The effectuation process of decision-making assumes a limited set of resources, such as financial resources, knowledge or managerial time is available and so the attention is concentrated on choosing between the possible effects of applying resources to alternative internationalization projects. This approach seems most suitable when the decision in not precisely specified due to ambiguous and rapidly changing goals and values. An alternative approach proposed by Sarasvathy (2001) is a process that applies causation logic. The causation process assumes a particular result, such as expanding into a location of choice, and focuses on the best means available to achieve that result. In the case of international expansion, this approach assumes that the choice of location can be made by entrepreneurs, and only particular modes of expansion require further inquiry. In reality, entrepreneurs are forced to choose among many locations due to the scarcity of resources and the management attention that they can give to an international expansion project at an early stage of company development (Nummela et al. 2014). The studies determining how new ventures are making strategic decisions usually examine two types of approaches: effectuation and causation (Nummela et al. 2014; Kalinic et al. 2014), however they can also be based on the entrepreneurs' idiosyncratic prior knowledge and their prior social and business ties (Evers and O'Gorman 2011). Lower probabilities of survival rates of international new ventures as compared to other internationalizing companies (Mudambi and Zahra 2007) raise the question whether the use of effectuation and causation logics aiming to find creative solutions in the absence of knowledge and expertise are indeed the best possible routines in the initial phase of internationalization. In this paper, we propose that rather to accept affordable loses and aim to improve their decision logic, entrepreneurs may elicit the expertise from outside of their team.

The use of external advisors can prove to be a must when a new venture considers a location in a distant and largely unknown country. A distant location, such as one on another continent, is perceived as a risk increasing choice in international business literature (Zdziarski et al. 2017). Assessing a case where risks are very high, and internal knowledge is limited, requires appropriate methods of eliciting the knowledge from experts. In this paper, we demonstrate how the method of structured expert judgement (SEJ) can improve the reliability of information upon which entrepreneurs make their location choice decision. In our study, we focus our attention on the location for the first deliberate foreign investment decision of a firm that is considering expanding globally. Our particular interest in this study is on the use of external forms of assistance, such as experts' advice. We apply the method known as the Classical Model (Cooke 1991) of eliciting experts' knowledge in a structured manner in controlled setting in which Executive MBA students from the International Management Centre at the University of Warsaw acted as experts. The Cooke's method is widely applied to elicit expertise needed in technical projects where risks are high, and little or no prior data is available. Our unique contribution presented in this chapter consists of demonstrating how this state-of-the-art decision support technique under uncertainty can be applied to guide business decision on foreign location choice. To our best knowledge, structured expert judgement hasn't been used in the strategic business decision-making process so far. We also aim to contribute to literature on decision-making and risk mitigation in early internationalization by focusing on knowledge sourcing from external experts.

The chapter is structured as follows. We start with introducing the context of the foreign market location choice at international new ventures operating in emerging markets. Later, we present the structured expert judgement methods with particular focus on Cooke's Classical Model. We describe the research study and show how the application of SEJ improves the reliability of forecasts in key areas as defined by the entrepreneur in the process of decision-making about location choice. We conclude with a discussion on the possibilities to improve practices in location choice decisions, as well as the advantages and limitations of the presented method of decision support.

23.2 Literature Review

23.2.1 Foreign Market Location Choice in International New Ventures

Foreign market location choice in international expansion is among the classic subjects of inquiry in the field of International Business (IB), and its predecessors in international trade and capital theories (Kim and Aguilera 2016). The inquiry on foreign location choice is a part of a broader attempt to explain the logic of a firm's internationalization that includes research on the selection of an entry mode, sequence of internationalization and the related concepts of liabilities of foreignness and outsidership (Johanson and Vahlne 1977, 2011). In the seminal paper, Dunning (2009) argues that "more attention needs to be given to the importance of location per se as a variable affecting the global competitiveness of firms". We follow this call to increase the research attention in a specific context of the location choice process in the case of a small, entrepreneurial firm that has the potential to become an international new venture.

International new ventures were identified as a new phenomenon in the last decade of the XX century as “a business organization that, from inception, seeks to derive a significant competitive advantage from the use of resources and the sale of outputs in multiple countries” (Oviatt and McDougall 1994). In the light of progressing globalization and increasing competition from abroad, a small business must be interested in internationalization, as this is one of the ways to counter the growing competition (Kubickova and Peprny 2011). Supporting this view, a study of 126 CEOs and top managers responsible for their companies’ internationalization indicated that they perceived non-internationalization as bearing higher risk than concentrating exclusively on the home market (Kraus et al. 2015).

Since these firms do not possess abundant resources that are at the disposal of multinational corporations, the consequences of selecting a wrong location to expand bring even more critical risks for their survival and future prospects. Entrepreneurs and managers of international new ventures are often unexperienced and despite their mindset for international expansion, they possess limited knowledge and competences (Crick 2009). Past research confirms that risks from global expansion materialize for many rapidly internationalizing firms, which often do not perform well after initial investments (Barringer and Greening 1998; Bell et al. 2004)

New theoretical approaches like the LLL (linking, learning, leveraging) model of internationalization (Mathews 2006), springboard perspective (Luo and Tung 2007) or adventurous internationalization (Zdziarski et al. 2017) are helpful in explaining the logic of internationalization of large corporations from emerging markets. However, the explanatory power of many IB theories is fairly limited in its application to small, entrepreneurial and international new ventures (Phillips McDougall et al. 1994; Coviello 2006). The unique character of these firms justifies the exploration of new theoretical propositions and decision routines that can guide both the theory and the practice of international entrepreneurship. In particular, it should help to explain internationalization from less developed, emerging economies (Bruton et al. 2008).

Usually, research on the antecedents of the location choice regresses the probability of investing in a given location on a set of independent variables that are expected to influence the profitability of an internationalization project. These variables explaining the probability of selection typically include some measures of local market potential, cost of production, cost of transportation, taxes and the general business environment in a given location (Cheng and Kwan 2000). Some researchers have also given attention to the legal form, or the mode of entrance. For example, in their study, Agarwal and Ramaswami (1992) found that small firms with limited multinational experience preferred entry into foreign markets through a joint venture. Physical distance is also taken into account; however, differences should be marked between distance-creating factors like culture and language, as well as distance-bridging factors like international travel (Ellis 2007) and the Internet, including the presence and intensity of absolute and comparative advantages (Franco et al. 2008).

Indeed, research confirmed that various forms of distance (cultural, geographic, political and economic) are strong predictors of risk perceptions in internationalization decisions, markedly exceeding the role of market-entry mode (Kraus et al. 2015).

For some time now, we have been also observing a more embedded network perspective on the location choice of multinational enterprises (Cantwell 2009; Johanson and Vahlne 2011; Xia et al. 2014). In consequence, the network-based relational variables are increasingly prevailing in explaining the selection of a country for international expansion. This is reflected in a recent critical review of location choice research in the field of IB from 1975 to 2015, which has identified the following determinants of location choice: experiential learning, top management's or firm's background and networks, customer relationship, industry characteristic, inter-regional ties, macroeconomic environment, distance between home and the host country, availability of natural resources and agglomeration (Jain et al. 2016). As a result of the review, the authors have proposed a two-stage decision model in which the determinants were grouped into two higher level constructs: those that facilitate resource deployment internationally for exploitation or exploration, and those which enable to evaluate the attractiveness of a host country for resource deployment (Jain et al. 2016).

Internationalization is often perceived as a gradual process in which firms accumulate knowledge over time, or as a learning process based on trial and error (Blomstermo et al. 2004). The fact that decision-makers and firms learn in the internationalization process implies that the first decision on location bears the most severe risks and the highest liabilities for a firm. This belief can be found in the early IB literature: "The first foreign investment decision is, to a large extent, a trip to the unknown. It is an innovation and the development of a new dimension as well as a major breakthrough in the normal course of events" (Aharoni 1966). In our study, we focus our attention on the location for the first deliberate foreign investment decision of a firm that is considering expanding globally. Our particular interest in this study is on the use of external forms of assistance, such as experts' advice that proved beneficial for the entrepreneurs in four of the five cases included in the study of the internationalization of small firms (Barringer and Greening 1998). Experts help to limit uncertainty and risks, such as in the case of investing in a distant location, by providing relevant information upon which a decision-maker decides about the future project. However, since experts are used for advice on uncertain future events and states, they often do differ in their judgements. In such a case, the entrepreneur may be often left with an uneasy choice of which expert advice to follow, and which to ignore. In the absence of own expertise, he or she can also use some form of averaging the conflicting forecasts. The work on improving the assessment methods under uncertainty resulted in the development of standard procedures that prove to outperform either simple averaging or random choice of an expert in the majority of analysed cases, such as the Cooke's method (Cooke 1991) that we use for this study and present below.

23.3 Materials and Methods

23.3.1 Structured Expert Judgement Use in the Uncertainty Quantification

The evaluation of risk is an assessment of the uncertainty, and, in the absence of data, experts' knowledge can provide proper risk quantifications. The Classical Model (CM) or the Cooke's method (Cooke 1991) is one of the best-known methods of eliciting experts' knowledge in a structured manner. CM has been used in numerous applications from various sectors, e.g. nuclear applications, chemical and gas industry, water pollutions, occupational, health, aerospace, banking, volcanoes and dams (Cooke and Goossens 2008; Colson and Cooke 2017).

We emphasize here the distinction between problems of managerial and scientific uncertainty; therefore, we distinguish between indecision, ambiguity and uncertainty (Liesch et al. 2014). The issue of indecision refers to finding the best solution given the circumstances and it is seen as the stakeholder's or problem owner's task. The issue of ambiguity is in the responsibility of the analyst to make sure that the stated problem is clear. The issue of uncertainty refers to quantifying the existing uncertainties, either from data or from experts. It is the analyst's responsibility to account for uncertainties resulting from data and it is the experts' responsibility to account for uncertainties when data is lacking or is inappropriate.

CM employs a protocol in which experts are asked to assess their uncertainties by stating quantiles for the distributions of various uncertain quantities. The standard approach is to ask experts for the 5%, 50% and 95% quantile. The 5% quantile is the value stated by the expert for which she/he thinks there is a 5% chance that the true value is below the stated value. It is regarded as the lower bound of expert's credible interval. Similarly, the 95% quantile represents the upper bound of the credible interval, denoting a value for which there is 5% chance that the true value lies above the 95% quantile. We interpret the expert best estimate as the median or the 50% quantile.

The protocol distinguishes two types of questions: the questions of interest and the calibration or seed questions. The calibration questions are questions for which the true value (or realization) is known to the analyst but not to the experts. The role of the calibration questions is threefold. Firstly, they support the objective quantification of experts' performance with respect to statistical accuracy and information. Secondly, they enable a performance-based combination of experts. Finally, they allow for the evaluation and validation of the performance-based combination of experts (Cooke and Goossens 2008). The calibration questions and hence the calibration score provide the means to prove that "heuristics can be accurate in the face of uncertainty" (Loock and Hinnen 2015).

The performance-based weighting has been shown to outperform the equal weighting of experts in all but one of the 33 CM studies and when performing in-sample analysis (Colson and Cooke 2017). Similarly, it has been shown that in

60 of the 63 considered professional studies, performance-based weighting outperformed equal weighting (see Chap. 10, this volume). Furthermore, performance-based weighting of experts has been shown to outperform the equal weighting via out-of-sample validation, in 26 out of 33 CM studies (Colson and Cooke 2017).

The method aims at a rational consensus rather than a census or a political consensus (Cooke and Goossens 2008). The rational consensus emerges as group decision processes, where "the group agrees on a method according to which a representation of uncertainty will be generated for the purposes for which the panel has convened, without knowing the result of this method" (Cooke and Goossens 2008). Therefore, unlike other expert judgement methods such as Delphi and Sheffield method, CM does not require each expert to adopt the results as her/his own degree of belief. The rational consensus implies that the experts agree with the scientific method of assessing the performance and combining expert opinion.

Rational consensus invokes four necessary conditions: accountability, empirical control, neutrality and fairness. The accountability assumption ensures that the method is based on a fully tractable process, in which experts' assessments are not publicly linked, but are available to peer review and must be reproducible. Secondly, experts' assessments are subject to empirical control. The neutrality ensures that experts are encouraged to state their true beliefs. Fairness entails that experts are regarded equal prior to objectively evaluating their assessments.

Along with CM, different models and methods that help to quantify uncertainty attracted quite a lot of attention in recent years (Bolger and Wright 2017). EKE consists of a set of techniques and methods, including the Delphi and Sheffield method, that helps to elicit the knowledge of experts. Furthermore, expert assessment is an established methodology to obtain information about relationships that are difficult to observe directly (Uusitalo et al. 2015).

23.3.2 Empirical Setting and the Expert Elicitation

23.3.2.1 The Context of the Study

We performed SEJ for an existing company that was pondering over the area of future market expansion. Sat Agro is a Polish start-up company providing applications that translate satellite maps into programs guiding precision fertilization. The company developed from a scientific collaboration of Przemysław Żelazowski and Kazimierz Stopa having institutional affiliations at University of Warsaw, Polish Academy of Science and Oxford University. In 2016 they registered the company Sat Agro and were joined by another partner and board member Urszula Starakiewicz-Krawczyk. During the first year of their activity, Sat Agro internationalized its operations to the Czech Republic based on a client's request. The initial internationalization was dome without seriously considering this move as the neighbouring country of Czech Republic was considered to be close and well known to entrepreneurs, and thus bearing no serious risks. The company currently considers the further expansion to other international markets, possibly to another continent. Such a move has many unknowns and requires more careful managerial consideration of choices, according to the opinion of decision-makers we interviewed at the beginning of the study.

23.3.2.2 The Process of the Study

The process of the study was performed in two stages: first, the company specified potential markets and criteria for consideration as well as information that would help them to make an informed choice. The founders considered several potential markets for international expansion, including France, Australia, US, China, Russia, Ukraine, Brazil, and the southern African region. Executive MBA students participating in the International Business course were assigned these markets—one for each group with a task to recommend a decision if the company should go for an international expansion project in the market that they have been analysing. The students presented their reports during a 4-h workshop with Sat Agro entrepreneurs commenting on each presentation. In a summary of the session, the entrepreneurs explained that, based on their updated knowledge from the teams' presentations, their preferred choice for the market to focus on was Brazil. For their final location investment choice, they believed several further uncertainty areas needed to be considered to assess what they could expect in near future. As the company expressed an interest in the Brazilian market, the second stage of the study that is of core interest for this chapter focused on this country. In the second stage, we applied Cooke's method to elicit expertise from a group of Executive MBA students having more diverse experiences and competencies with internationalization projects than the entrepreneurs themselves.

23.3.3 Method

23.3.3.1 Participants

Sixteen Polish participants (9 male; 7 female) of the Executive Master of Business Administration (MBA) course participated in the study. The participants were in the middle of senior executive positions in a variety of organizations, including banks, multinational and Polish enterprises, as well as public administration, i.e. in the Ministry for Economic Development, the Ministry of Foreign Affairs, or the Chief Pharmaceutical Inspectorate. The Executive MBA is a flagship executive education program at the University of Warsaw, and the first program of this type was established in Poland at the beginning of the transition to market economy in 1991. Since then, 23 cohorts of students, i.e. almost 1000 people, graduated from the program.

23.3.3.2 Procedure

The study took place after a regular class. The participants were informed about ïts purpose and asked for the consent to participate. Next, the introduction to the SEJ method was given, followed by a dry-run of the CM methodology. The dry-run exercise was used to acquaint the experts with CM and included 3 calibration weather-related questions. During a short break, the experts' statistical accuracy or calibration score and information score for the dry-run were computed and their assessments aggregated using the performance-based weighting scheme. The experts were informed afterwards about the results and the manner in which their assessments are evaluated in the CM was emphasized. After making sure that all experts clearly understood the procedure, the formal elicitation was conducted. All participants received the elicitation forms containing the calibration questions and questions of interest. The elicitation was conducted for all the participants at once, so that it was ensured that the participants did not have contact with each other and made their assessments independently. After the study, the experts were thanked for their participation.

23.3.3.3 EJ Elicitation Protocol

We adopted the formalized procedure for eliciting expert judgements, based on the Classical Model for structured expert judgement (Cooke 1991). All participants completed questionnaires consisting of 18 questions—12 calibration questions and 6 questions of interest. The questions were prepared based on the interview conducted by the authors with one of the founders of the Sat Agro company. The owner was asked about the factors that they take into account when deciding on the internationalization strategy as well as the foreign markets that they consider for potential expansion. The owner was also asked to justify their decision to explore further opportunities in the Brazilian market, which they chose in the first phase of the project. One of the arguments in favour of Brazil was the lack/small number of competitors, while, i.e. in the United States, the market was congested, and barriers of entry would be higher. Regulations in the Brazilian market were not as strict as in the other countries under consideration.

The interview enabled the identification of key criteria that the entrepreneurs would focus on when evaluating their final location choice decision. Accordingly, the questions of interest enquired about the prediction of various Brazilian market scores in 2020. We asked six questions regarding the Corruption Perceptions Index (CPI), the Global Innovation Index, the Global Competitiveness Index (GCI), the Country Risk Index, the World Justice Project (WJP) Rule of Law Index, as well as the forecasted number of paid users. The relevance and content of the items used in this task were verified by peer judges prior to the study.

23.4 Results

First, we analysed the experts' assessments for the calibration questions with respect to two performance measures, the calibration score or statistical accuracy and the information score. The analysis has been performed using the Excalibur software, which has been developed at Delft University of Technology. Table 23.1 presents the performance scores for each expert, as well as their combined score and the weights resulting from these scores.

The calibration score is computed from the 12 calibration questions and denotes the statistical accuracy with respect to the true values of the calibration questions. It ranges between zero and one, where a high score denotes a better statistical accuracy. We note that the most statistically accurate expert is Exp1, with a calibration score of 0.046. Nonetheless, the calibration scores are quite low. The information score denotes how informative the experts' assessments are. The information score reflects the experts' uncertainty; therefore, a low information score denotes a high uncertainty, whereas a high information score denotes a low uncertainty. The information score in Table 23.1 is obtained by averaging the information scores of the 12 calibration questions. Similarly to the calibration score, the higher the information score, the more informative the expert is. We observe that the information score ranges from 0.97 to 2.662, where 2.662 denotes a high information score.

Table 23.1 Experts' performance scores

Expert	Calibration	Information	Information all questions	Combined score	Weight
Exp1	0.04663	1.166	1.037	0.05438	0.6554
Exp2	6.20E−06	1.778	1.805	1.10E−05	0.000133
Exp3	1.42E−06	2.662	2.61	3.78E−06	4.55E−05
Exp4	0.000344	1.538	1.385	0.000529	0.006372
Exp5	5.59E−07	2.071	1.921	1.16E−06	1.40E−05
Exp6	3.97E−08	1.596	1.548	6.33E−08	7.62E−07
Exp7	2.55E−05	1.802	1.756	4.59E−05	0.000553
Exp8	5.59E−07	1.936	1.902	1.08E−06	1.30E−05
Exp9	3.19E−05	0.9768	0.9359	3.12E−05	0.000376
Exp10	1.42E−06	1.832	1.774	2.60E−06	3.13E−05
Exp11	4.69E−06	2.028	1.904	9.52E−06	0.000115
Exp12	1.35E−06	2.4	2.324	3.24E−06	3.90E−05
Exp13	1.37E−05	2.224	2.086	3.05E−05	0.000367
Exp14	0.01639	1.704	1.569	0.02793	0.3366
Exp15	3.49E−09	2.517	2.414	8.77E−09	1.06E−07
Exp16	5.59E−07	2.343	2.315	1.31E−06	1.58E−05

Even though the true value is not known for the questions of interest, the information score can still be computed. An average of all the information scores of all questions in the study, and therefore both calibration questions and the questions of interest, is provided in ‘Information all questions’ in Table 23.1. It is interesting to investigate the differences between the two information scores, as it reflects on the differences between the uncertainties in the calibration questions and the questions of interest. For experts with a lower information score for all questions, such as Exp1, Exp4, Exp10, etc., it denotes a higher uncertainty in the questions of interest than in the calibration questions. For Exp2, the information score in the questions of interest is higher than the information score for the calibration questions.

Ideally, we would like the experts to be highly informative and, more importantly, highly calibrated or more statistically accurate. A higher calibration score is preferred to a higher information score, since high information with poor calibration denotes overconfidence. This is observable, for example, for experts with very high information scores but very low calibration scores. The combined score captures this preference, and we observe that Exp1, though not as informative as other experts, has the best-combined score, as a reward for being the highest calibrated expert.

The normalized weights of the experts are computed by dividing the expert’s combined score by the sum of all experts’ combined score. Given the highest combined score of Exp1, it is straightforward that Exp1 also receives the highest weight. The second highest weight is received by Exp14 and all other experts receive a very low weight.

These weights are referred to as performance-based weights, since they are computed based on the two performance measures. The performance-based weights allow for the aggregation of experts into the so-called decision-maker (DM) for the questions of interest. It is the DM’s assessments that are usually reported as a conclusion of the study. Furthermore, the DM can be regarded as any other expert and hence can have its performance evaluated with respect to the calibration and information score obtained from DM’s assessments for the calibration questions.

Another method of aggregating the experts’ assessments is equal weighting, where each expert, regardless of their assessments, receives equal weight. In our study, since there are 16 experts, every expert receives the equal weight of 0.0625. We will denote by ‘Performance DM’ the DM obtained by aggregating the experts using performance-based weights and ‘Equal DM’ the DM obtained by weighting the experts equally. The results of the two DM’s are presented below.

Table 23.2 presents the results for the two DM. First of all, we notice a calibration score of 0.446 for the performance-based DM. This reflects a good statistical accuracy, which is much higher than the calibration scores of each expert. It shows that

Table 23.2 Performance measures for a performance-based and equal-based decision-maker (DM)

DM	Calibration	Information	Information all
Performance DM	0.446	1.039	0.895
Equal DM	0.298	0.476	0.424

Table 23.3 Experts' performance scores and optimized DM

Expert	Calibration	Information	Information all questions	Combined score	Weight
Exp1	0.04663	1.166	1.037	0.05438	0.6607
Exp2	6.20E−06	1.778	1.805	1.10E−05	0
Exp3	1.42E−06	2.662	2.61	3.78E−06	0
Exp4	0.000344	1.538	1.385	0.000529	0
Exp5	5.59E−07	2.071	1.921	1.16E−06	0
Exp6	3.97E−08	1.596	1.548	6.33E−08	0
Exp7	2.55E−05	1.802	1.756	4.59E−05	0
Exp8	5.59E−07	1.936	1.902	1.08E−06	0
Exp9	3.19E−05	0.9768	0.9359	3.12E−05	0
Exp10	1.42E−06	1.832	1.774	2.60E−06	0
Exp11	4.69E−06	2.028	1.904	9.52E−06	0
Exp12	1.35E−06	2.4	2.324	3.24E−06	0
Exp13	1.37E−05	2.224	2.086	3.05E−05	0
Exp14	0.01639	1.704	1.569	0.02793	0.3393
Exp15	3.49E−09	2.517	2.414	8.77E−09	0
Exp16	5.59E−07	2.343	2.315	1.31E−06	0
DM_opt	0.446	1.067	0.9381		

the DM has improved significantly its statistical accuracy compared to the statistical accuracy of all experts. Moreover, its calibration score is also higher than the calibration score of the equal-based DM. Finally, the information scores display a much better performance for the performance-based DM than for the equal-based DM.

We can attempt to improve DM's performance by excluding some experts with very low calibration scores. The optimized combination of experts leads to a weighting scheme that is different from the one in Table 23.1. Table 23.3 shows the results of performing an optimization analysis, as well as the performance scores of the optimized DM.

We notice that only two experts get non-zero weight in the optimized combination of experts. Nonetheless, given the very low weights of other experts, the weights do not differ much from the weights in Table 23.1. Furthermore, we note that the calibration score is the same as for the non-optimized DM, whereas the information scores are slightly higher.

The final results regard the questions of interest, namely, the DM's resulting quantiles. Table 23.4 contains this information.

The first question of interest helps to assess the anticipated corruption level in Brazil. We have informed the experts on the standard measure of corruption perceptions provided annually by Transparency International—the CPI index. The CPI ranges from 0 (highly corrupt) to 100 (very clean). Between the years 2012 and 2015, the level of the index ranged between 38 and 43. The experts were asked to elicit the

Table 23.4 DM's answers for the questions of interest

Question	5%	50%	95%
I1	25.12	35	44.8
I2	30.87	37.35	44.8
I3	3.03	4.04	4.98
I4	55.02	65	74.86
I5	0.40	0.52	0.6
I6	5.39	42.61	100

CPI index in 2020. As we can read from Table 23.4, DM expects the corruption to increase in the next few years to the anticipated level of 35 CPI, which is the value assigned to 50% quantile that best represents experts' opinion. The entrepreneurs can also be assured by this table that DM expects less than a 5% chance that CPI will decrease below 25.12, which would denote a substantial increase in the corruption levels; similarly, a 5% chance is assumed for the index to be above 44,8. This will imply a very small increase as compared to the years of 2012 and 2014 in which Transparency International CPI scores for Brazil were 43.

In question two, we were concerned about the innovation capacity of Brazil for which the experts were asked to estimate changes in the Global Innovation Index (GII), This index is based on, among others, human capital and research, infrastructure, scientific outputs, creative outputs. It ranges from 0 (very bad) to 100 (very good). We have used a similar format as the one reported for question 1. The experts were given information about the Global Innovation Index for Brazil in 2012 and 2014, which ranged from 34.95 to 36.33. Their task was to respond to the following question: What will the Global Innovation Index be in 2020? The results from Table 23.4 indicate that the best DM estimate is that the level of index moderately increases to 37.35. It is worth noting that, unlike in the case of question 1, the best estimate is closer to the range of historical values, suggesting one should only expect a moderate and positive change in respect to the innovation capacity—the factor that the entrepreneurs thought is important in their knowledge intensive industry.

The third question concerned the Global Competitiveness Index (GCI) for Brazil in 2020. The index is provided by the World Economic Forum every year in the Global Competitiveness Report. The Global Competitiveness Index (GCI) accounts for factors that determine the level of productivity and economy, but also institution and policies; its scores range from 1 (the lowest GCI) to 7 (the highest GCI). The Global Competitiveness Index 2016–2017 for Brazil was 4.06. According to the answers provided by the optimal performance-based DM, the estimated GCI in 2020 is 4.04, which denotes a conservative approach to the current GCI. The experts' combined assessments lead to confidence intervals of [3.03; 4.98] to capture the uncertainty around the estimate.

Question number four involved the Country Risk Index (CRI), calculated based on the business risk rating, the country risk rating and the political risk rating. The index ranges from 1 (very risky) to 100 (not risky at all). In 2014, the Country Risk Index for Brazil was 69 and in 2015, it was 67. The experts needed to estimate the Country Risk Index for Brazil in 2020. The DM's estimated the index to be at 65, which denotes a slight decrease compared to the values in 2014 and 2015. The uncertainty inherited from the experts' distributions is nonetheless quite large. This shows a high variance among the experts' assessments, which denotes a disagreement among the experts' assessments.

The fifth question regarded the WJP Rule of Law Index. The WJP Rule of Law Index 2016 presents a portrait of the rule of law in each country by providing scores and rankings organized around eights factors: constraints on government powers, absence of corruption, open government, fundamental rights, order and security, regulatory enforcement, civil justice, and criminal justice. The ninth factor—informal justice—is measured but not included in the aggregated scores and rankings. The scores range from 0 to 1 (with 1 indicating the strongest adherence to the rule of law). In 2015, The WJP Rule of Law Index in Brazil was 0.56. The question that the experts needed to answer was: What will be the WJP Rule of Law Index in Brazil in 2020? Once more, the DM's solution shows that the index is forecasted to slightly decrease, denoting a slight improvement of the Brazilian market with respect to the Law Index. The confidence intervals are relatively smaller when compared to other confidence intervals, suggesting a reduced uncertainty and more agreement among the experts' assessments.

Finally, the experts were asked to provide uncertain assessments for the number of paid users in the Brazilian market. The question was as follows: SatAgro had 23 paid users in 2016 and was monitoring 31,000 ha of land in Poland and The Czech Republic. If the company decides to expand to the Brazilian market and offer their services there, how many paid users will the company have in the Brazilian market 3 years after the internationalization in Brazil? The DM's best estimate is around 42 paid users. Nonetheless, the number of paid users can vary between 5 and 100, denoting a high uncertainty.

23.5 Conclusions and Discussion

This chapter details an application of a well-established decision support methodology in a new context—that of strategic managerial decision-making on international expansion of a small, entrepreneurial firm. The aim of the paper was to demonstrate how international new ventures might benefit from using external advice of experts while taking a risky decision about their initial foreign investment to a distant location. In a controlled setting, we engaged Executive MBA students as experts. We applied the Classical Model for Structured Expert Judgement to elicit their expertise on the internationalization project. The expert panel enabled us to provide forecasts in six areas identified by the entrepreneur as critical in the process of finalizing the decision whether to invest in Brazil. We collaborated with an existing, innovative Polish company SatAgro, which was at the stage of selecting from among different alternatives for its international expansion. To assist the company in making its risky decision, we engaged the participants of the Executive MBA program to first gather information about the potential locations defined by the firm, and then based on its interest in Brazil, to elicit future states in areas where the firm wishes to know more to ground its investment decision.

An initial investment in a distant location is a type of decision in which uncertainty and risks are very high. If entrepreneurs do not possess the required competences and direct experience with the market they consider for an expansion, like in the described case, they may take the advantage of reaching out for expertise. However, one can elicit expertise in several ways. An entrepreneur will often take into account advice from a single expert who seems to have business credentials and expertise. If accessible, a student of a prestigious Executive MBA program can likely be approached as an advisor. Such students need to have several years of managerial experience before being admitted to the program, and many of them had come across internationalization projects in their prior managerial careers. The result of our study should bring attention to the fact that an expert having sound business acumen, and perhaps even some international experience, does not necessarily offer a sound advice with respect to uncertainty quantification. In fact, quite the opposite proves to be true in our research—the assessments of our experts were poorly calibrated, and often also overconfident as indicated by the information scores. These results reflect the poor performance of individual experts as assessors of uncertainty. If the entrepreneur bases his or her decision on advice from a single expert, randomly chosen from our sample, he or she will be misguided by the poor assessments of uncertainty of an individual.

Nonetheless, it is remarkable that the combination of experts based on their performance leads to a decision-maker that is much more statistically accurate as well as more informative. Even in the situation when each individual expert was poorly calibrated as assessed by the seed questions, we were able to combine their expertise and greatly improve the calibration scores—from 0.04 for the best calibrated individual expert to 0.446 for the performance-based decision-maker. Notably, performance-based weighting also works much better from a simple combination of experts based on equal weights, which results in almost a half of the statistical accuracy and more than half of the informativeness that can be achieved in the case of more optimal combinations.

Concluding, our study clearly demonstrates that engaging a panel of experts in a structured elicitation process with the application of the Classical Model offers a much better alternative to either using advice from individual experts or simply averaging expert judgements from a group. The likely improvements in both statistical accuracy and informativeness are indeed impressive and reassure that using the Classical Method enables a big improvement in the reliability of information upon which the decision is made.

Finally, the present study has limitations that need to be pointed out. Since we cannot expect that the company will soon expand to the Brazilian market, we are unable to check if the predictions of the judges are correct. That does not diminish the value of the method, but indicates the path for future studies—we would like to perform a study in which we could check the correctness of the experts versus the empirical results, which requires the study to be extended in time for the overall period of the forecast. Furthermore, only a few experts had experience in internationalization, and only to markets other than Brazil. Thirdly, in the Classical Model, the experts are interviewed separately, whereas in our adopted version—we conducted our study for all of the participants simultaneously. This is not an unusual practice, as some researchers are conducting elicitations in a workshop format (Hanea et al. 2018). The Classical Model emphasized the importance and necessity of the motivation and rationales behind experts' assessments that provide additional information beyond the numerical judgements. Due to the time and cost constraints of conducting a more elaborated study, we were not able to include additional questions on the rationale in the present study.

The process of an interactive support provided by the students of the Executive MBA program to an innovative start-up on its way to becoming an international new venture that we described in this chapter is a good example of action research. The early proponent of action learning approach, Kurt Lewin has famously said: "There is nothing as practical as a good theory" (1951). Our study demonstrated that the practice of internationalization in small, entrepreneurial firms can be guided by a notable contribution of Cooke's Classical Model to applied mathematics and the decision-making theory.

References

Agarwal, S., & Ramaswami, S. (1992). Choice of foreign market entry mode: Impact of ownership, location and internalization factors. *Journal of International Business Studies, 23*(1), 1–27. https://doi.org/10.1057/palgrave.jibs.8490257.

Aharoni, Y. (1966). The foreign investment decision process. *The International Executive, 8*(4), 13–14. https://doi.org/10.1002/tie.5060080407.

Barringer, B. R., & Greening, D. W. (1998). Small business growth through geographic expansion: A comparative case study. *Journal of Business Venturing, 13*(6), 467–492. https://doi.org/10.1016/S0883-9026(97)00038-4.

Bell, J., Crick, D., & Young, S. (2004). Small firm internationalization and business strategy. *International Small Business Journal, 22*(1), 23–56. https://doi.org/10.1177/0266242604039479.

Blomstermo, A., Eriksson, K., Lindstrand, A., & Sharma, D. D. (2004). The perceived usefulness of network experiential knowledge in the internationalizing firm. *Journal of International Management, 10*(3), 355–373. https://doi.org/10.1016/j.intman.2004.05.004.

Bolger, F., & Wright, G. (2017). Use of expert knowledge to anticipate the future: Issues, analysis and directions. *International Journal of Forecasting, 33*(1), 230–243. https://doi.org/10.1016/j.ijforecast.2016.11.001.

Bruton, G. D., Ahlstrom, D., & Obloj, K. (2008). Entrepreneurship in emerging economies: Where are we today and where should the research go in the future. *Entrepreneurship Theory and Practice, 32*(1), 1–14. https://doi.org/10.1111/j.1540-6520.2007.00213.x.

Cantwell, J. (2009). Location and the multinational enterprise. *Journal of International Business Studies, 40*(1), 35–41. https://doi.org/10.1057/jibs.2008.82.

Cheng, L. K., & Kwan, Y. K. (2000). What are the determinants of the location of foreign direct investment? The Chinese experience. *Journal of International Economics, 51*(2), 379–400. https://doi.org/10.1016/S0022-1996(99)00032-X.

Colson, A. R., & Cooke, R. M. (2017). Cross validation for the classical model of structured expert judgment. *Reliability Engineering and System Safety, 163,* 109–120.

Cooke, R. M. (1991). *Experts in uncertainty: Opinion and subjective probability in science*. New York: Oxford University Press.

Cooke, R., & Goossens, L. (2008). TU Delft expert judgment data base. *Reliability Engineering and Systems Safety, 93*(5), 657–674.

Coviello, N. E. (2006). The network dynamics of international new ventures. *Journal of International Business Studies, 37*(5), 713–731. https://doi.org/10.1057/palgrave.jibs.8400219.

Crick, D. (2009). The internationalisation of born global and international new venture SMEs. *International Marketing Review, 26*(4/5), 453–476. https://doi.org/10.1108/02651330910971986.

Dunning, J. H. (2009). Location and the multinational enterprise: A neglected factor? *Journal of International Business Studies, 40*(1), 5–19.

Ellis, P. D. (2007). Paths to foreign markets: Does distance to market affect firm internationalisation? *International Business Review, 16*(5), 573–593. https://doi.org/10.1016/j.ibusrev.2007.06.001.

Evers, N., & O'Gorman, C. (2011). Improvised internationalization in new ventures: The role of prior knowledge and networks. *Entrepreneurship & Regional Development, 23*(7–8), 549–574. https://doi.org/10.1080/08985621003690299.

Ferencikova, S., & Hluskova, T. (2015). Internationalization of Central and Eastern European companies—Theory and its implications in the Slovak IT sector. *Journal of East European Management Studies, 20*(4), 415–434.

Franco, C., Rentocchin, F., & Marzetti, G. V. (2008). Why do firms invest abroad? An analysis of the motives underlying foreign direct investments. *The IUP Journal of International Business Law, 9*(1 and 2), 42–65. https://doi.org/10.2139/ssrn.1283573.

Hanea, A., McBride, M., Burgman, M., & Wintle, B. (2018). Classical meets modern in the IDEA protocol for structured expert judgement. *Journal of Risk Research, 21*(4), 417–433.

Jain, N. K., Kothari, T., & Kumar, V. (2016). Location choice research: Proposing new agenda. *Management International Review, 56*(3), 303–324. https://doi.org/10.1007/s11575-015-0271-6.

Johanson, J., & Vahlne, J. E. (1977). The internationalization process of the firm—a model of knowledge development and increasing foreign market commitments. *Journal of International Business Studies, 8*(1), 23–32.

Johanson, J., & Vahlne, J. E. (2011). Markets as networks: implications for strategy-making. *Journal of the Academy of Marketing Science, 39*(4), 484–491. https://doi.org/10.1007/s11747-010-0235-0.

Kalinic, I., Sarasvathy, S. D., & Forza, C. (2014). 'Expect the unexpected': Implications of effectual logic on the internationalization process. *International Business Review, 23*(3), 635–647. https://doi.org/10.1016/j.ibusrev.2013.11.004.

Kim, J. U., & Aguilera, R. V. (2016). Foreign location choice: Review and extensions. *International Journal of Management Reviews, 18*(2), 133–159. https://doi.org/10.1111/ijmr.12064.

Kraus, S., Ambos, T. C., Eggers, F., & Cesinger, B. (2015). Distance and perception of risk in internationalization decisions. *Journal of Business Research, 68*(7), 1501–1505.

Kubickova, L., & Peprny, A. (2011). The internationalization of small and medium-sized enterprises in the viticulture. *Agricultural Economics—Czech Republic, 57*(7), 331–339.

Lewin, K. (1951). *Field Theory in Social Science*. New York: Harper and Row.

Liesch, P. W., Welch, L. S., & Buckley, P. J. (2014) Risk and uncertainty in internationalisation and international entrepreneurship studies. In *The Multinational Enterprise and the Emergence of the Global Factory*. London: Palgrave Macmillan.

Loock, M., & Hinnen, G. (2015). Heuristics in organizations: A review and a research agenda. *Journal of Management Research, 68*(9), 2027–2036.

Luo, Y., & Tung, R. L. (2007). International expansion of emerging market enterprises: A springboard perspective. *Journal of International Business Studies, 38*(4), 481–498.

Mathews, J. A. (2006). Dragon multinationals: New players in 21st century globalization. *Asia Pacific Journal of Management, 23*(1), 5–27. https://doi.org/10.1007/s10490-006-6113-0.

Mudambi, R., & Zahra, S. (2007). The survival of international new ventures. *Journal of International Business Studies, 38*(2), 333–352. https://doi.org/10.1057/palgrave.jibs.8400264.

Nummela, N., Saarenketo, S., Jokela, P., & Loane, S. (2014). Strategic decision-making of a born global: a comparative study from three small open economies. *Management International Review, 54*(4), 527–550. https://doi.org/10.1007/s11575-014-0211-x.

Oviatt, B. M., & McDougall, P. P. (1994). Toward a theory of international new ventures. *Journal of International Business Studies, 25*(1), 45–64.

Phillips McDougall, P., Shane, S., & Oviatt, B. M. (1994). Explaining the formation of international new ventures: The limits of theories from international business research. *Special International Issue, 9*(6), 469–487. https://doi.org/10.1016/0883-9026(94)90017-5.

Sarasvathy, S. D. (2001). Causation and effectuation: Toward a theoretical shift from economic inevitability to entrepreneurial contingency. *The Academy of Management Review, 26*(2), 243–263.

Uusitalo, L., Lehikoinen, A., Helle, I., & Myrberg, K. (2015). An overview of methods to evaluate uncertainty of deterministic models in decision support. *Environmental Modelling and Software, 63,* 24–31. https://doi.org/10.1016/j.envsoft.2014.09.017.

Wilinski, W. (2013). Internationalization of Central and Eastern European countries and their firms in the global crisis. In M. A. Marinov & S. T. Marinova (Eds.), *Emerging Economies and Firms in the Global Crisis.* London: Palgrave Macmillan.

Xia, J., Ma, X., Lu, J. W., & Yiu, D. W. (2014). Outward foreign direct investment by emerging market firms: A resource dependence logic. *Strategic Management Journal, 35*(9), 1343–1363. https://doi.org/10.1002/smj.2157.

Zdziarski, M., Światowiec-Szczepańska, J., Troilo, M., & Małys, Ł. (2017). Adventurous foreign direct investment. *Journal of Management and Business Administration. Central Europe, 2,* 117–138. https://doi.org/10.7206/jmba.ce.2450-7814.197.