International Series in
Operations Research & Management Science

Luis C. Dias
Alec Morton
John Quigley *Editors*

Elicitation

The Science and Art of Structuring Judgement

EXTRAS ONLINE

Springer

International Series in Operations Research & Management Science

Volume 261

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

Luis C. Dias • Alec Morton • John Quigley
Editors

Elicitation

The Science and Art of Structuring Judgement

Springer

Editors
Luis C. Dias
Faculty of Economics
CeBER and INESC Coimbra
University of Coimbra
Coimbra, Portugal

Alec Morton
Department of Management Science
University of Strathclyde
Scotland, UK

John Quigley
Department of Management Science
University of Strathclyde
Scotland, UK

ISSN 0884-8289 ISSN 2214-7934 (electronic)
International Series in Operations Research & Management Science
ISBN 978-3-319-65051-7 ISBN 978-3-319-65052-4 (eBook)
https://doi.org/10.1007/978-3-319-65052-4

Library of Congress Control Number: 2017955545

Printed on acid-free paper

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

Contents

About the Editors

Luis C. Dias obtained a degree in Informatics Engineering from the School of Science and Technology, a Ph.D. in management, and a habilitation in decision aiding science at the University of Coimbra. He is currently associate professor and vice-dean for research at the Faculty of Economics of the University of Coimbra (FEUC), where he has been teaching courses on decision analysis, operational research, informatics, and related areas. He held temporary invited positions at the Paris Dauphine University and the University of Vienna. Luis is also a researcher at the CeBER and INESC Coimbra R&D centres, a member of the coordination board of the University of Coimbra's Energy for Sustainability Initiative, and currently the vice-president of APDIO, the Portuguese Association of Operational Research. He is on the editorial board of the *EURO Journal on Decision Processes* and *Omega*. His research interests include multicriteria decision analysis, performance assessment, group decision and negotiation support, decision support systems, and applications in the areas of energy and environment.

Alec Morton has degrees from the University of Manchester and the University of Strathclyde. He has worked for Singapore Airlines and the National University of Singapore and London School of Economics; has held visiting positions at Carnegie Mellon University in Pittsburgh, Aalto University in Helsinki, and the University of Science and Technology of China (USTC) in Hefei; and has been on secondment at the National Audit Office. His main interests are in decision analysis and health economics. Alec has been active in the INFORMS Decision Analysis Society and the OR Society. He is on the editorial board of *Decision Analysis* and is an associate editor for the *EURO Journal on Decision Processes*, the *Transactions of the Institute of Industrial Engineers*, and the *OR Spectrum*. His research has won awards from the International Society for Pharmacoeconomics and Outcomes Research and the Society for Risk Analysis and the publication award from the INFORMS Decision Analysis Society.

John Quigley has a bachelor of Mathematics in Actuarial Science from the University of Waterloo, Canada, and a Ph.D. in management science from the University of Strathclyde, where he is currently professor. He is an industrial statistician with extensive experience in elicitation of expert judgment to support model development and quantification through subjective probability distributions, having worked closely over the past 25 years with various engineering organizations on problems concerned with risk and reliability. John has been involved in consultancy and applied research projects with, for example, Aero Engine Controls, Rolls-Royce, Airborne Systems, BAE Systems, and the Ministry of Defense (MOD). His collaborative work on Bayesian model development as part of the Reliability Enhancement Methodology and Modelling (REMM) project is included in the industry standard for reliability growth analysis methods. John is a tutor for the European Food Safety Agency (EFSA) on expert knowledge elicitation (EKE) as well as an associate of the Society of Actuaries, a chartered statistician, and a member of the Safety and Reliability Society.

Chapter 1
Elicitation: State of the Art and Science

Luis C. Dias, Alec Morton, and John Quigley

> *I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of Science, whatever the matter may be.*
>
> Lord Kelvin (1891, pp. 72–72)

Abstract This book is about elicitation, which may be defined as the *facilitation of the quantitative expression of subjective judgement*, whether about matters of fact or matters of value. To motivate, we review case studies from human health (swine flu); provision of public services (airport location); natural hazards (assessment of the risk of earthquakes) and environmental protection (in the case of radioactive waste) where elicitation was or could have been profitably used to inform decisions. It is often argued that uncertainties are too deep or human values are too profound for quantitative thinking to be applicable: we argue on the contrary (drawing again on cases) that it is impossible to think about important problems without dealing with problems of "how big" and "how much". We provide an overview of chapters in the book, which, we argue, shows that there is a huge body of knowledge and expertise about how to elicit both probabilities and preferences in important social problems, and conclude with future trends that make the subject of this book (in our view) particularly timely.

L.C. Dias (✉)
Faculty of Economics, CeBER and INESC Coimbra, University of Coimbra,
Av Dias da Silva 165, 3004-512, Coimbra, Portugal
e-mail: lmcdias@fe.uc.pt

A. Morton
University of Strathclyde, 16 Richmond St, Glasgow, G1 1XQ, UK

J. Quigley
University of Strathclyde, Glasgow, UK
e-mail: j.quigley@strath.ac.uk

L.C. Dias et al. (eds.), *Elicitation*, International Series in Operations Research & Management Science 261, https://doi.org/10.1007/978-3-319-65052-4_1

1.1 Conceptual Background

A useful definition of a problem is that it is a situation where there is a current state, and a desired state, and they are not the same. Most people are familiar with this sort of situation and many day to day problems can be dealt with by largely subconscious or automatic processes (the coffee is too bitter, so I add sugar; the water is too cold, so I turn the tap). But some problems (I want to take up a new hobby, perhaps a new sport, a new language, or a new instrument) require reflection: I have to reflect what goals I want to achieve and whether the actions I have at my disposal will help me achieve them. In such cases I have to build a mental model of my problem to organise my thoughts and help me choose wisely. Other problems, even more complex, involve the significant others in my life (where should we go on holiday?; should we move to a new city, or new country, to take that new job?): in these cases, the model I build should be a shared one, so as to ensure that all those involved in the problem understand what they are getting into. At a higher level still, society has to take important decisions about responses to threats to our environmental and economic wellbeing and security: in a democracy these decisions should take account of the views of the public in some organised fashion.

Tackling problems at multiple levels therefore, requires models of value and models of (our knowledge of) the world around us which may be to a greater or lesser extent implicit or explicit depending on the background nature of the problem. These models can also be more or less complex. For example, in the event of an uncontrolled emission of radioactive materials from a nuclear plant into the atmosphere, the core decision may look as shown in Table 1.1, where the rows are the choice of actions, the columns are scenarios which may be realised and the consequences in the cells are the outcomes experienced by humanity. This is a very simple model. At the other end of the scale, there are much more complex models (e.g. Geldermann et al. 2009). Such systems may allow (probabilistic) forecasting of wind direction, and model the dispersion of radionuclides and the consequent damage to human health. These more complex models may require drawing on extensive amounts of data and cutting-edge science.

Table 1.1 A simple model of a decision in nuclear emergency management

		Scenarios	
		Wind blow seawards and radioactive material is dispersed over the ocean	Wind blow landwards and radionuclides are dispersed over land
Actions	Evacuate nearby town	Unnecessary evacuation with result cost and hardship	Population are moved out of the path of harm's way
	Do not evacuate but encourage people to shelter indoors	Damage and inconvenience both minimised	Population are exposed to potentially hazardous levels of radiation

To whom should we go when we wish to deliberate on these models? In the case of models of the world around us, it seems reasonable to privilege experts—those who have relevant knowledge about the subject matter—above lay members of the public. However, identifying these experts may not be straightforward. Senior professors may have long since ceased to keep up with the research literature and be primarily expert at obtaining research funding, and managing grants. Industry experts may be blinkered by social norms and conflicts of interest, especially if they depend for employment or consulting income on other powerful stakeholders. Moreover, if we want to make a genuinely informed decision, we want experts who are able, not just to offer an opinion, but to give us an assessment of how much confidence we can have in their assessment. This requires a cognitive ability which is entirely distinct from actual knowledge.

We may seem to be on safer ground when it comes to models of value. In these cases, surely the person to engage with is the decision maker. Yet this is not a particularly helpful observation. In many situations there is no single unitary decision maker. Even if one person has to sign on the dotted line, the agreement of many people is required if the decision is to be real—is actually to result in action and change. What is more, research strongly suggests that even when making consequential decisions people do not know the goals that they have—even if asked to spontaneously list their goals, there are many other not less important goals which they also recognise as being relevant to their choices (Bond et al. 2008). Therefore, it is wise to engage in reflective dialogue with their friends and partners about significant choices—even if the decision falls to you alone.

This book is a book about elicitation, which may be defined as the *facilitation of the quantitative expression of subjective judgement*, whether about matters of fact or matters of value. Why should anyone want to express their judgements quantitatively, or to help others to do so? So far, we have stressed the role of models in underpinning decision making. But these models are often—and always in the case of models which are exclusively mental models—qualitative in nature.

We believe that people should be encouraged to express their judgements quantitatively as a way of making their thoughts precise, and ensuring that they are testable against the evidence from the real world. Statements like "This year there will probably be a lot of rain in Glasgow" or "Artistic self-realisation is more important to me than money" are hopelessly vague: "This year there is a 50% chance of more than 1100mm of rain in Glasgow" and "I would be prepared to take a pay cut of up to £7K per annum to free up a day a week for my theatre workshop" can be tested against the actual realised weather and my actual choices respectively.

For important decisions, this clarity is critical, we believe, if we are to have high quality, transparent engagement of experts and stakeholders; if we seriously care about having high quality deliberative dialogue. It is not that words and qualitative reasoning are not important. However, significant decisions inevitably involve weighing competing risks and values and questions of relative magnitude inevitably arise. The only way to communicate clearly about relative magnitude is through the use of numbers. For such decisions, words and numbers are jointly necessary and indeed, complementary.

It is true that people are not (yet) accustomed to use numbers to express their judgements of fact or value. For some people this is difficult or uncomfortable; others have an ideological objection to it, as they view quantification as having a technocratic flavour. Yet we believe that the difficulties are overstated. As the chapters in this book show, there are many ways to enable people to express their quantitative judgements, which can be customised to quite different cognitive styles and tastes. Many of the elicitation methods we review involve asking respondents purely qualitative questions: the numbers are, so to speak, "backed out" from their answers.

Our purpose in this section has been to present our motivating philosophy and the conceptual underpinnings of the current volume. In the remainder of this chapter, we discuss in more detail the need for, and barriers to, using elicitation of probabilities and preferences to support decision making, outline the chapters of the book, and in conclusion, present some common themes and ways forward.

1.2 The Need for and Barriers to Elicitation

> For the purposes of sharpening assumptions and distinguishing them, nothing beats an exercise in probability. (Neustadt and Fineberg 1983, p. 118)

> Values are what we care about. As such values should be the driving force for our decision making . . . But that is not the way it is. It is not even close to the way it is. Keeney (1992)

1.2.1 The Need for Elicitation of Judgement

In this subsection, our aim is to reflect on the need for elicitation. We do so by considering cases where elicitation was or could have been profitably used. These in-depth cases will give a sense of the breadth of potential application across time and across domains. Specifically we deal with four areas of applications, which are depicted in Table 1.2 below: they cover human health (swine flu); provision of public services (airport location); natural hazards (assessment of the risk of earthquakes) and environmental protection (in the case of radioactive waste). Although all cases involve both uncertainties about matters of fact (probabilities) and conflicts about values (preferences), two case studies are better used to highlight the former, and the other two, the latter.

Table 1.2 Four case studies which illustrate the potential for structured elicitation

	Assessment of probabilities	Assessment of preferences
Historic (1960s/70s)	Case 1. Swine flu	Case 2. Airport location
Recent	Case 3. Assessment of risk of earthquake	Case 4. Radioactive waste

1.2.1.1 Case 1. Swine Flu

In early February 1977, then US Secretary of Health, Education and Welfare Joseph A Califano Jr. was confronted with the decision to release stocks of influenza vaccine; he had been in post for 2 weeks. The vaccine had been used in autumn 1976 to begin immunizing the nation against swine flu, a strain of the H1N1 influenza, and possibly prevent an epidemic on the scale of the Spanish flu which caused the death of 3–6% of the world's population in 1919. The vaccine had been withheld due to possible but not certain links with Guillain-Barre Syndrome, which is an often paralyzing and sometimes fatal side effect. This unenviable time pressured task of decision making under uncertainty concerned trading between risks, where traditional "scientific" evidence from controlled lab based experiments did not exist and as such must rely on expertise. Today, the outbreak is most remembered for an unnecessary mass immunisation that cost $135 million (Harrell 2009). The virus resulted in one fatality while side-effects from the vaccine are thought to have caused 25 deaths due to Guillain-Barre syndrome (Roan 2009). There is no guarantee that decision making under such circumstances will result in the best outcome post-hoc, however better processes for working with expert judgement seem to have been needed.

Much has been written on this outbreak with the most in-depth critique of the decision making process "The Epidemic That Never Was: Policy Making & The Swine Flu Affair" (Neustadt and Fineberg 1983), published after the event with the aim of learning lessons for the future. While there are a number of confounding issues that led to the decision to attempt to vaccinate the entire US population a key shortcoming identified in the process was the lack of probability assessments, explicitly identifying the need for experts to quantify their uncertainty in terms of probability, exposing their judgment for comparison with one another.

1.2.1.2 Case 2. Airport Location

A perennial issue in UK politics over the last several decades has been airport capacity planning in the crowded South-East of the country around London. An instructive episode in this history is the Roskill Commission (Hall 1980) appointed by the UK government in 1968, and which reported in 1971. The centerpiece of the Roskill Commission's Report was a highly detailed economic cost-benefit analysis ("without doubt the largest and most complex of its kind attempted anywhere"—Hall 1980, p 32) which involved calculations and monetisation not only of capital investment and passenger time, but also noise impacts, agricultural impacts and the like. The Commission's calculations pointed to a site—Cublington—between London and Birmingham as the best choice. However the publication of the report and the substantive recommendation of Cublington generated a storm of controversy. One commission member wrote an impassioned note of dissent suggesting that the Commission's entire methodology had been misguided as it completely ignored the overriding importance of preserving open countryside. Academic commentators

such as Mishan (1970) ("What is wrong with Roskill?") and Self (1970) ("Nonsense on stilts") piled into the discussion with trenchantly expressed take-downs of the study methodology. An important theme of the Mishan and Self critiques is that the Roskill Commission calculations embed disputable and critically important assumptions about social values, such as equity. In large part because of a disconnect between the values embedded in the cost-benefit analysis and political and popular perceptions, the Roskill recommendation of Cublington was ultimately rejected and the government chose to explore the option of building an airport at Foulness.

The experience of the Roskill Commission is a reminder that complex decisions are "wicked problems" (Rittel and Webber 1973) and feature conflicting stakeholders, with multiple, competing, objectives. Effective analyses have to grapple with these features of the problem context rather than wish them away. It is interesting to contrast the mode of analysis of the Roskill Commission with the Multicriteria Decision Analysis (MCDA) described in Keeney (1992) for the location of the new Mexico City airport. This very early decision analysis (originally reported in 1972) nevertheless features the use of computerised sensitivity analysis to explore and communicate the model, in order to assist decision makers to reflect on their value judgements.

1.2.1.3 Case 3. Assessment of the Risk of Earthquake

In early April 2009, an earthquake struck L'Aquila Italy killing 309 people. Six scientists and one government official who participated in Italy's National Commission for the Forecast and Prevention of Major Risks 6 days prior to the earthquake were sentenced to 6 years in prison in October 2012 for manslaughter. The prosecution argued that the expert advice from the Commission resulted in 30 people deciding to stay indoors which resulted in their death. The case led to outrage from many in the scientific community who argued that earthquakes cannot be predicted with certainty, so the trial was seen by some as an attack on science. The prosecuting attorney Fabio Picuti was not criticizing the experts on these grounds, rather on a lack of evaluation of the degree of risk present in L'Aquila; the presiding judge Marco Billi ruled the analysis was superficial. An appeal in November 2014 resulted in all six scientists being acquitted and the government official having his jail sentence reduced to 2 years, on the grounds that only the government official was responsible for the communication of the risk assessment that led to the death of the 30 individuals. For details see Nature (2011) and Science (2012, 2014).

A further criticism of the L'Aquila risk assessment identified by Alessandro Martelli and Lalliana Mualchin who were respectively the President and General Secretary of the International Seismic Safety Organisation (ISSO) concerned the dangers of the lack of independence amongst expert judgments (Martelli and Mualchin 2012). This tragedy highlights a need for transparent, rigorous and widely accepted processes for assessing uncertain events.

1.2.1.4 Case 4. Radioactive Waste Management

In 2003 the UK government set up a Committee on Radioactive Waste Management (CoRWM) to address the problem of what to do with the UK's inventory of radioactive waste. This problem was not new, but acquired new urgency as the current fleet of nuclear reactors was coming to the end of its life and government wanted to commission new nuclear power plants in order to ensure continuity of generating capacity. However, previous efforts to arrive at a solution—involving "deep disposal" of waste stocks in a deep underground repository had left a legacy of popular distrust of the nuclear industry and of the government. CoRWM was asked to take a new alternative approach—open and participative, and capable of inspiring public confidence.

Early on, CoRWM decided that they would systematically involve a broad range of stakeholders and conduct as much as possible of their business in public. However, a challenge was how to reconcile this with the need to actually reach a decision which all members of the committee (who brought a diverse range of views) could actually sign up to. One of the strengths of the CoRWM process was their use of a systematic MCDA as a core (though not the only) component of their deliberative strategy (Morton et al. 2009). The MCDA model provided a transparent basis through which different concerns—for example about safety, or about the need to avoid a burden on future generations—and stakeholder perspectives could be discussed and weighed up against each other. The MCDA also played a key role in communicating the rationale for the decision in the final report (CoRWM 2006). Thus, CoRWM provides a good example of how explicit elicitation and modelling of value tradeoffs can play an important role in supporting complex societal decisions.

1.2.2 Why do People Resist Expressing Their Uncertainty and Values Quantitatively?

It is sometimes argued that attempting to employ analytic methods in situations which are characterised by uncertainty and conflict over objectives reflects a technocratic arrogance in the face of a fundamentally uncertain, unpredictable world and/or a profane disregard for the role of human values in decision making. As examples of the former, *Black Swans* (Taleb 2007) and *Perfect Storms* (Junger 1997) are two metaphors used to describe rare events about which there is "deep uncertainty" which is impossible to quantify. As an example of the latter, consider Tetlock's (2003) discussion of "sacred values": "A sacred value can be defined as any value that a moral community implicitly or explicitly treats as possessing infinite or transcendental significance that precludes comparisons, trade-offs, or indeed any other mingling with bounded or secular values" (Tetlock et al. 2000, p. 853).

We believe that appeals to "deep uncertainty" or "sacred values" often reflect lazy, superficial thinking about both possible future events and human objectives.

Again, we frame our discussion through four case studies: Deepwater Horizon and the Fukushima nuclear disaster for deep uncertainty; the approval of new drugs and the concept of capability in military planning.

1.2.2.1 Deep Uncertainty Case 1: Deepwater Horizon

In April 2010, a geyser of seawater erupted onto the BP Deepwater Horizon rig located in the Gulf of Mexico resulting in the largest offshore oil spill in US history; eleven platform workers were killed and seventeen injured. The National Academy of Engineering and National Research Council (2010) argued early indications of the problem existed from several repeated tests of well integrity. Bea (2010) attributes the cause stemming from the failure of multiple processes, systems and equipment. While this event may appear as a Black Swan as we have never experienced such an event before, it was not beyond the boundaries of reasoned imagination (Paté-Cornell 2012), as early warning signals were present. To model this is possible: we would require an assessment not only of each event but the dependency between events, where all the events which precipitated the disaster are made more likely through a certain management style.

1.2.2.2 Deep Uncertainty Case 2: The Fukushima Disaster

In March 2011, an earthquake in Japan resulted in the release of seismic energy into a place of convergent boundaries of tectonic plates, i.e. a subduction zone, causing a tsunami that reached 14 m. The Fukushima Daiichi nuclear reactors which were designed for a maximum wave height of 5.7 m, were affected by the tsunami, resulting in nuclear meltdowns and release of radioactive material. The plants design was deemed safe as the likelihood of a wave in excess of 6 m was less than 0.01 in the next 50 years, although historical evidence of such extreme waves existed albeit from the 9th and seventeenth Century (Paté-Cornell 2012). Moreover, while the buildings were designed to withstand a tsunami, the plants backup generators were not (Masys 2012). This event illustrates how analogous events for which data exists could inform the identification of events and the assessment of the associated uncertainty on events.

1.2.2.3 Sacred Values Case 1: the Approval of New Drugs

Some of the hardest values to think about systematically are values which relate to one's own quality of life and, ultimately one's own mortality. However, since most of our healthcare is provided by third parties, either governments or insurance funds, there is a need to make tradeoffs since not all medical technologies, which influence one's health and survival are affordable. One tool for structuring such

tradeoffs is the Quality Adjusted Life Year or QALY (Pliskin et al. 1980). QALYs provide a numerical assessment of health benefit which integrates quality of life and survival. Roughly speaking, QALYs are calculated via a quality of life score, which reflects different dimensions of quality of life such as level of pain, mental distress or mobility, multiplied by length of life. Over the last two or three decades QALYs (and their variants) have become widely used and accepted in many jurisdictions (Drummond et al. 2015), with the precise parameters used to calculate the QALYs being elicited from local populations to reflect local preferences. The success of the QALY in ensuring that public spending on medicines is in line with social values shows the potential of a simple, yet theoretically robust concept in making previously taboo tradeoffs discussable in the public sphere.

1.2.2.4 Sacred Values Case 2: The Concept of "Capability" in Military Planning

A common way in which values become sacred in organisational management is where values are specified at an insufficiently strategic level. Protection of existing programmes becomes identified with loyalty to one's division of the organisation and accepting reallocations becomes identified with surrender. Addressing such issues requires creating an overarching conceptual framework in which the contribution of individual programmes to the common good can be traced and articulated. In businesses, profitability often provides this framework but in other sorts of organisations, the path to constructing such a framework might be less obvious. A good example of such a framework in a non-business setting is the idea of "military capability" which has been recently popular in countries such as the UK, US, Australia and Finland (Anteroinen 2012). The idea in such frameworks is to substitute arguments between individual services about how many ships, tanks, or planes with arguments about how to deliver particular capabilities: for example a monitoring capability may be delivered by human reconnaissance, UAVs or satellites. Once this substitution has been made, it is possible to reframe decisions away from being about which branch of the service suffers and towards what constitutes the best way of delivering the ability to meet national military needs.

Reviewing the above cases, we freely admit that eliciting probabilistic or tradeoff information may be difficult: but we argue that the proper response is not to declare that the problems are somehow too profound for quantitative thinking to be useful, but rather to think carefully and creatively about what the difficulties are and how to tackle them. With this motivation, the rest of the book represents a sourcebook of methods and concepts for doing this.

1.3 Overview of the Book

The idea for this book originated in the COST Action "Expert Judgment Network: Bridging the Gap Between Scientific Uncertainty and Evidence-Based Decision Making",[1] noting the importance of using sound elicitation processes when building models to inform decision making. Elicitation may be needed to populate models of uncertainty, interacting with subject experts, but it may also be needed to set up models of preferences, interacting with experts, decision makers, and other stakeholders. In both cases, it is important that analysts and experts follow a process that allows them to think clearly about numbers, whether they concern probabilities or they concern the importance of attributes, for instance. Hence, this book covers elicitation processes having in mind both probabilities and preferences.

A first major group of chapters in this book (Chapters 2 to 9) focusses on processes to elicit uncertainty from experts. Chapter 2: "Elicitation in the Classical Model", by Quigley, Colson, Aspinall and Cooke, presents the Classical Method for aggregating judgements from multiple experts concerning a probability distribution; the method uses mathematical aggregation based on the performance of experts. In Chapter 3: "Validation in the Classical Model", Cooke discusses the issue of validation: what constitutes good uncertainty assessment and how can this be measured. This chapter addresses in particular the Classical Method, for which many studies have already been carried out. Chapter 4: "SHELF: The Sheffield Elicitation Framework", by Gosling, presents the Sheffield elicitation framework, also to elicit probability distributions, covering its foundations, its extensions, and its applications. In contrast to the Classical method, a behavioural aggregation method is proposed in this chapter. Chapter 5: "IDEA for Uncertainty Quantification", by Hanea, Burgman and Hemming, outlines a protocol named IDEA, which is a mixed approach combining behavioural and mathematical aggregation techniques that can be used instead of, for example, the well-known Delphi protocol. The two ensuing chapters present approaches based on the principles of Bayesian updating of probability distributions. Chapter 6: "Elicitation and Calibration: A Bayesian Perspective", by Hartley and French, discusses how one might use a full Bayesian model to combine the judgements of multiple experts into a posterior distribution, considering prior experts' judgements as data. In Chapter 7: "A Methodology of Constructing Subjective Probability Distributions with Data", Quigley and Walls present an approach to represent expert uncertainty through analogies with existing empirical data so reducing the burden of quantification on experts.

Chapters 8 and 9 address important issues that are of relevance for different probability elicitation approaches. Chapter 8: "Eliciting Multivariate Uncertainty from Experts: Considerations and Approaches along the Expert Judgement Process", by Werner, Hanea and Morales-Nápoles, discusses the main elements of structured expert judgement processes for dependence elicitation, when eliciting

[1] http://www.expertsinuncertainty.net/.

multivariate distributions using either pooling or Bayesian approaches. Chapter 9: "Combining Judgements from Correlated Experts", by Wilson and Farrow, discuss how mathematical methods for expert judgement aggregation, whether opinion pooling or Bayesian methods, can incorporate correlations between experts; they also consider behavioural approaches and the potential effects of correlated experts in this context.

A second major group of chapters in this book (Chapters 10 to 14) focusses on processes to elicit preferences from stakeholders or decision makers. The elicitation processes covered here consider situations in which a decision maker (or a group) needs to make a decision. The purpose of the elicitation is then to build a model of the preferences of the decision maker (and often, preferences of other stakeholders) that helps this decision maker in making sound and informed decisions. The first two chapters on preference elicitation deal with problems under uncertainty. Chapter 10: "Utility Elicitation", by Gonzalez-Ortega, Radovic, and Ríos Insua, presents the classical decision analysis paradigm of utility theory, to elicit models of preferences that can be combined with models of uncertainty. They cover the case of preferences concerning one attribute, multiple attributes, and the preferences of adversaries. In Chapter 11: "Elicitation in Target-Oriented Utility", Bordley presents a different perspective on utility elicitation based on targets, which allows using probability elicitation methods to elicit utility functions.

Chapters 12 to 14 address the elicitation of preferences independently of, or in the absence of, any uncertainty elicitation. These chapters concern preferences over multiple attributes or evaluation criteria. In Chapter 12: "Multiattribute value elicitation", Morton presents the multiattribute value theory, from its foundations to process aspects important in practice. Chapter 13: "Disaggregation Approach to Value Elicitation", by Matsatsinis, Grigoroudis and Siskos, is based on the same theoretical grounds, but introduces a new outlook on a very different elicitation paradigm, which involves eliciting preferences indirectly, "disaggregating" comparisons that a decision maker is able to make at an holistic level into the components of a multiattribute value function. In Chapter 14: "Eliciting Multi-Criteria Preferences: ELECTRE Models", Dias and Mousseau discuss the elicitation of an outranking-based preference model, considering in particular ELECTRE methods, which are based on principles that are different from the value measurement framework of the preceding chapters.

Chapters 15 and 16 are about cross-cutting issues that are relevant for elicitation of uncertainties as well as for elicitation of preferences. In both cases, the experts or stakeholders involved can incur into biases leading to answers that, upon reflection, they would wish to revise. In Chapter 15: "Individual and Group Biases in Value and Uncertainty Judgments", Montibeller and von Winterfeldt overview the biases that individuals and groups are subject to, and also what might be done to reduce the occurrence of such biases. Chapter 16: "The Selection of Experts for (Probabilistic) Expert Knowledge Elicitation", by Bolger, addresses another issue present in any elicitation process involving expert judgement, which is the selection of the experts. This chapter presents a structured process having in mind mainly probability elicitation, but it is also relevant for preference elicitation.

The last group of chapters illustrates how some of the approaches presented in this book can be, and are being, applied in practice. In Chapter 17: "Eliciting Probabilistic Judgments for Integrating Decision Support Systems", Barons, Wright and Smith describe an integrating decision support system for probabilistic judgement elicitation (under a Bayesian approach), and they illustrate its potential on a food security case in the UK. Chapter 18: "Expert Elicitation to Inform Health Technology Assessment", by Soares and Bojke, illustrates expert elicitation in health care decision making, discussing two examples of formal elicitation to inform Health Technology Assessments in the UK. Chapter 19: "Expert Judgment Based Nuclear Threat Assessment for Vessels Arriving in the US", by Merrick and Albert, demonstrates an expert judgment based method using pairwise comparisons and parameter estimation to elicit nuclear threat risks concerning the security of US ports. In Chapter 20: "Risk Assessment using Group Elicitation: Case Study on Start-up of a New Logistics System", Porthin, Rosqvist and Kunttu present a risk assessment concerning a new logistics system for a pulp and paper manufacturer in the Nordic countries, using a computer system designed to support decision-making (i.e. that can also be used to elicit preferences). Chapter 21: "Group Decision Support for Crop Planning: A Case Study to Guide the Process of Preferences Elicitation" differs from the previous applications in that it deals with elicitation of preferences rather than uncertainties, but it also considers a situation involving a group of individuals. In this chapter, Delias, Grigoroudis and Matsatsinis present a case study of applying a multicriteria disaggregation approach to elicit a model of preferences for crop planning in a Greek region.

1.4 Conclusions and Future Directions

As we have emphasised in our initial section, we see the aim of elicitation as being to facilitate the quantitative expression of subject judgement, not as an end in itself, but to facilitate high quality dialogue and reflection about important decisions. As our overview of the chapters shows, there is now a vibrant applied discipline which draws on a rigorous and well-developed theoretic base, and which has provided us with a toolbox of techniques, each custom-developed to meet the needs of particular sorts of problems and the preferred cognitive styles of different people.

The philosophy of this book is that quantification, through elicitation, is a way to refine and clarify the mental models which people inevitably use in thinking about complex problems. Quantification enables clearer communication about these models between people—all sorts of people—but also sharpens the predictions which these models make about the world and enable them to be tested empirically. A book on this subject is (in our view) particularly timely because of the following trends in the world.

- *Increasing range of choice*. Our experience of the world is increasingly mediated through digital technologies which are global in reach. This means that we are routinely confronted with choices broader than ever before. A trivial example is that we are now able to download virtually any book in print from Amazon onto

our Kindle reader; our choices about our education, career, potential political or religious beliefs, choice of life partner, etc., have been similarly broadened. We need aids which will help us organise and make sense of these complex choices, and which will enable us to weigh, select and aggregate, and ultimately make better and more life-enhancing decisions.

- *Increasing availability of data*. As more transactions are conducted online and as the cost of data storage and processing drops, businesses and governments have been increasingly able to collect, process, and make available large volumes of data relating to their activity. Unfortunately, gathering data does not itself bring insight, and in the absence of a strong research design, making inferences about what caused what, and the generalising from then and there to now and here can be extremely difficult. To build meaning from this data requires somehow infusing the data with expert judgement.
- *Increasing demands for accountability*. During the debate prior to the 2016 Brexit referendum, the UK politician Michael Gove remarked that the "people in this country have had enough of experts" (Financial Times 2016). Whatever one's view on the substantive issue, the outcome of that referendum clearly validated Gove's claim: the UK voting public did not trust the experts who predicted that leaving the EU would be a disaster, or the elites who purported to take decisions in their best interests. A technological society cannot survive without experts, or without political officeholders, but lay people may reasonably demand confidence that expert judgements—and claims to expertise—are as open to scrutiny and testing as possible, and that the values which inform public decisions are subject to open and transparent debate.

The tools in this book have a vital role meeting all of these challenges: empowering purposeful decisions in the face of these overwhelming choices; making sense of vast, complex and ill-structured datasets; and building bridges between experts and elites on the one hand, and (perhaps rightly) suspicious lay people on the other.

Elicitation is a young technology. Other quantitative technologies—counting, measurement of physical dimensions—have been around for millennia. Yet other quantitative technologies—cost accounting for example—have become well established in the space of a few decades when it became clear that they met a need of a modern complex industrial society. There is the potential for elicitation of value and uncertainty to have a no less central role in the society of the future. We hope that this book will give the reader some ideas as to how that might come about.

References

Anteroinen J (2012) Integration of existing military capability models into the Comprehensive Capability Meta-model. Paper presented at the SysCon, Vancouver

Bea R (2010) Failures of the deepwater horizon semi-submersible drilling unit. Deepwater Horizon Study Group, University of California, Berkeley. https://ccrm.berkeley.edu/pdfs_papers/bea_pdfs/DeepWaterBobBeaPrelimAnalyses-rev5-2.pdf. Accessed 08 June 2017

Bond SD, Carlson KA, Keeney RL (2008) Generating objectives: can decision makers articulate what they want? Manag Sci 54(1):56–70
CoRWM (2006) Managing our radioactive waste safely. https://www.gov.uk/government/publications/managing-our-radioactive-waste-safely-corwm-doc-700. Accessed 08 June 2017
Drummond MF, Sculpher MJ, Claxton K, Stoddart GL, Torrance GW (2015) Methods for the economic evaluation of health care programmes, 4th edn. Oxford University Press, Oxford
Harrell E (2009) How to deal with swine flu: heeding the mistakes of 1976. Time Magazine http://content.time.com/time/health/article/0,8599,1894129,00.html. Accessed 12 June 2017
Financial Times (2016) Britain has had enough of experts, says Gove https://www.ft.com/content/3be49734-29cb-11e6-83e4-abc22d5d108c?mhq5j=e1. Accessed 8 June 2017
Geldermann J, Bertsch V, Treitz M, French S, Papamichail KN, Hamalainen RP (2009) Multi-criteria decision support and evaluation of strategies for nuclear remediation management. Omega Int J Manage S 37(1):238–251
Hall P (1980) Great planning disasters. University of California Press, Berkley
Junger S (1997) The perfect storm. Norton, New York
Keeney RL (1992) Value-focussed thinking: a path to creative decision making. Harvard University Press, Cambridge
Kelvin L (1891) Popular lectures and addresses. MacMillan, London and New York
Martelli A and Mualchin L (2012) Indictment and conviction of members of the Italian "Commissione Grandi Rischi" (open letter to the President of Italy). www.cngeologi.it/wp-content/uploads/2012/10/CoverletterandStatementISSO1.pdf. Accessed 08 June 2017
Masys AJ (2012) Black swans to grey swans: revealing the uncertainty. Disaster Prev Manag 21(3):320–335
Mishan EJ (1970) What is wrong with Roskill. J Transp Econ Policy 4(3):221–234
Morton A, Airoldi M, Phillips LD (2009) Nuclear risk management on stage: a decision analysis perspective on the UK's committee on radioactive waste management. Risk Anal 29:764–779
National Academy of Engineering and National Research Council (2010) Interim report on causes of the Deepwater Horizon oil rig blowout and ways to prevent such events. www.nap.edu/catalog/13047.html. Accessed 08 June 2017
Nature (2011) Scientists on trial: at fault? www.nature.com/news/2011/110914/full/477264a.html. Accessed 08 June 2017
Neustadt RE, Fineberg HV (1983) The epidemic that never was. Vintage, New York
Paté-Cornell E (2012) On "Black Swans" and "Perfect Storms": risk analysis and management when statistics are not enough. Risk Anal 32(11):1823–1833
Pliskin JS, Shepard DS, Weinstein MC (1980) Utility functions for life years and health status. Oper Res 28(1):206–224
Rittel HWJ, Webber MM (1973) Dilemmas in a general theory of planning. Policy Sci 4(2):155–169
Roan S (2009) Swine flu debacle of 1976 is recalled. Los Angeles Times. http://articles.latimes.com/2009/apr/27/science/sci-swine-history27?pg=1. Accessed 12 June 2017
Science (2012) Earthquake experts convicted of manslaughter. www.sciencemag.org/news/2012/10/earthquake-experts-convicted-manslaughter. Accessed 08 June 2017
Science (2014) Updated: appeals court overturns manslaughter convictions of six earthquake scientists www.sciencemag.org/news/2014/11/updated-appeals-court-overturns-manslaughter-convictions-six-earthquake-scientists. Accessed 08 June 2017
Self P (1970) "Nonsense on stilts": cost-benefit analysis and the Roskill commission Polit Q 41(3):249–260
Taleb NH (2007) The black swan: the impact of the highly improbable. Random house, New York
Tetlock PE (2003) Thinking the unthinkable: sacred values and taboo cognitions. Trends Cogn Sci 7(7):320–324
Tetlock PE, Kristel OV, Elson SB, Green MC, Lerner JS (2000) The psychology of the unthinkable: taboo trade-offs, forbidden base rates, and heretical counterfactuals. J Pers Soc Psychol 78(5):853–870

Chapter 2
Elicitation in the Classical Model

John Quigley, Abigail Colson, Willy Aspinall, and Roger M. Cooke

Abstract The Classical Model (CM) is a performance-based approach for mathematically aggregating judgements from multiple experts, when reasoning about target questions under uncertainty. Individual expert performance is assessed against a set of seed questions, items from their field, for which the analyst knows or will know the true values, but the experts do not; the experts are, however, expected to provide accurate and informative distributional judgements that capture these values reliably. Performance is measured according to metrics for each expert's statistical accuracy and informativeness, and the two metrics are convolved to determine a weight for each expert, with which to modulate their contribution when pooling them together for a final combined assessment of the desired target values. This chapter provides mathematical and practical details of the CM, including describing the method for measuring expert performance and discussing approaches for devising good seed questions.

2.1 Introduction

The Classical Model (CM) (Cooke 1991) was designed to function within a science-based quantitative uncertainty analysis. From the wide differences in experts' assessments and performance as uncertainty assessors, the CM must forge a rational consensus of 5–20 experts who each quantify their uncertainty on, typically, 20–30 uncertain items—all within the time and resource constraints of the study. The

J. Quigley (✉) • A. Colson
University of Strathclyde, Glasgow, UK
e-mail: j.quigley@strath.ac.uk

W. Aspinall
University of Bristol, Bristol, UK

Aspinall and Associates, Tisbury, UK

R.M. Cooke
Resources for the Future, 1616 P Street NW, Washington, DC, 20036, USA

Department of Mathematics, Delft University of Technology, Mekelweg 2, 2628 CD, Delft, Netherlands

L.C. Dias et al. (eds.), *Elicitation*, International Series in Operations Research & Management Science 261, https://doi.org/10.1007/978-3-319-65052-4_2

feature that makes this possible is empirical validation, which enables performance-based combination of the assessments. The goal of the CM is to sensibly combine the experts' disparate assessments of uncertainty, not to convince one person, be it expert, analyst, or problem owner, that his/her personal beliefs have been faithfully represented. For the results to qualify as science and appeal to rational consensus, the element of empirical control is essential. This is achieved by asking experts to quantify their uncertainty on a set of items from the experts' field, called calibration or seed variables, for which the true values are known post hoc. Details on performance measures and combination algorithms are described in Sect. 2.2.

The use of seed variables and performance measures transforms expert uncertainty quantification into an activity with all the hallmarks of scientific data acquisition. Not only is the analyst tasked to find plausible seed variables, (s)he must be able to explain and defend the elicitation method, performance measures, combination algorithms and indeed the entire purpose of external validation of subjective probabilities. Familiarity with the scientific background of the study, with foundational literature on the representation of uncertainty and with some previous applications is a requisite. In Sect. 2.3 we provide a detailed discussion on finding seed variables, which is illustrated with real applications.

To date, well over 200 professional expert panels have been elicited using seed variables and performance-based combination of expert judgements. The oft heard remonstration that "seed variables were not possible in this case" may really signal the challenge of finding an analyst/facilitator with the necessary skills to conduct a CM elicitation. This chapter describes the CM method and is intended to share experiences in finding and using seed variables, with the goal of lowering the barriers to science-based uncertainty quantification.

2.2 Classical Model Basics

The CM provides an approach for combining subjective probability distributions on quantities of interest through calculating a weighted average across multiple experts. In contrast to behavioral aggregation approaches, see Chapter 4: "SHELF: the Sheffield Elicitation Framework" in this book (Gosling 2018), the experts do not discuss and agree on a final aggregate distribution. Instead, the combined distribution is determined mathematically. The aim is to provide the most dependable probability distribution through assigning greater weight to the judgements of those experts who can be demonstrated to have performed better via the empirical validation stage; higher weight is assigned to experts with better statistical accuracy and informativeness. The weights are based on the theory of proper scoring rules so reward experts who authentically state their own scientific or technical judgement.

The fundamentals of the CM will be described throughout this section. We start with a description of the type and form of questions asked during the elicitation, which is followed by a discussion on statistical accuracy and how it is measured within the CM. The results from a simple Monte Carlo exercise are presented to

show how well the calibration score performs in discriminating between experts. The information score is explained and illustrated with a simple example. Finally, a description of how weights are assessed based on these measures is provided before closing with a reflection on further topics not covered.

2.2.1 Elicitation Questions

The CM considers two types of questions, namely, target variables and seed variables, about which we elicit expert uncertainty. Uncertainties are described by subjective probabilities and should be such as to ensure self-consistent betting behavior, e.g. assigning a probability of 0.5 to an event implies the expert is indifferent to betting on the outcome being realized or a fair coin landing heads. The target variables are the variables of interest for the study. They involve quantities that cannot be perfectly observed, calculated, or predicted with existing knowledge and tools, so they are uncertain. The purpose of the elicitation exercise is to obtain probability distributions to characterise the uncertainty associated with the value of these variables. Unlike other elicitation methods, the CM uses variables from their field for which the experts do not know the true value but the analyst does (or will sometime within the timeframe of the study), which are referred to as seed variables. The experts are not expected to know the values precisely but, as experts in the topic(s), they are expected to provide accurate and informative distributional judgements that capture these values reliably within a suitable narrow credible range. A set of several seed variables, typically eight to twenty in number, are used in a joint hypothesis test to assess formally, and auditably, the performance of the experts in expressing their uncertainty in terms of probability.

Seed variables should be defined so that they trigger the same judgement heuristics as the target variables. Thus, the seed variables should be representative of the target variables, so that performance on seed variables can be reasonably assumed to map over to performance on target variables. This presumption needs to be carefully scrutinized in each elicitation case. We will revisit this in detail in Sect. 2.3.

Unlike in a behavioural aggregation elicitation, during a CM elicitation session the expert will not discuss their assessments with other experts, although there can be a pre-elicitation workshop with interaction of experts prior to seeing the questions. Ideally, experts are unaware if they are assessing a seed variable or a target variable. This is often not feasible in practice, however, as the experts may be able to identify the subset of questions for which realizations are possible (i.e., the seed variables).

Experts are presented with a series of questions and asked to provide specified quantiles from their subjective probability distributions concerning the values of each variable. Typically, experts will be asked to provide their 5th, 50th and 95th quantiles, although many studies also ask for the 25th and 75th quantiles as well. In essence, an expert's quantiles reflect a point of indifference between two bets,

one based on the realization of the quantity of interest and the other based on a realization from a mechanism where the probability is known. As such, the expert's x^{th} quantile corresponds to a value for which the expert is indifferent to betting on the quantity of interest being less than or a lottery with a probability of winning being $^{x}/_{100}$.

As an example of the form of the questions, an aircraft engineer may be asked the following concerning the reliability of a new design.

> From a fleet of 100 new aircraft engines how many will fail before 1000 hours of operations?
>
> Please provide the 5th, 95th, and 50th percentiles of your estimate.
>
> 5%_____95%_____50%_____

The questions are sequenced to encourage the expert to first assess the range of values, i.e. the 5th and 95th percentiles, then a more central measure, i.e. 50th percentile. This sequencing is recommended as it compensates for a common heuristic, so-called 'anchoring', in which people do not adequately adjust their judgements on either side of their central estimate to reflect accurate upper and lower quantile values. However, the ease for experts of creating a probability distribution around a central value sometimes proves more important than avoiding this possibly troublesome bias. Heuristics and biases cannot be completely removed from an elicitation, but their impact should be minimized to the extent possible.

The performance measurement aspect of the CM allows an analyst to identify which experts are able to provide statistically accurate assessments despite the heuristics and biases in play. See Chapter 15: "Individual and Group Biases in Value and Uncertainty Judgments" of this book (Montibeller and von Winterfeldt 2018) for a full discussion on biases and heuristics.

The number of questions that can be assessed by an expert in a session will depend on the number of mental models they are required to generate in order to respond reliably to the questions, i.e. it depends are how closely linked the questions are. An expert can be expected to answer up to 100 similar or related questions during a session, although most elicitations involve fewer than 40 questions.

2.2.2 Calibration Score

As described in Sect. 2.2.1, the experts provide quantiles on both seed and target variables. The variables are measured on a continuous scale, for example the temperature tomorrow rather than whether it rains or not. The seed variables are used to assess calibration as the realizations are known, and the degree of calibration informs the weight applied to the assessment on the target variables. Figure 2.1 illustrates performance data from an expert on ten seed questions, where each horizontal line represents the number line, the vertical lines are where the expert has assigned their quantiles (for this example we consider only three, i.e. 5th, 50th and 95th), and the X represents the realization.

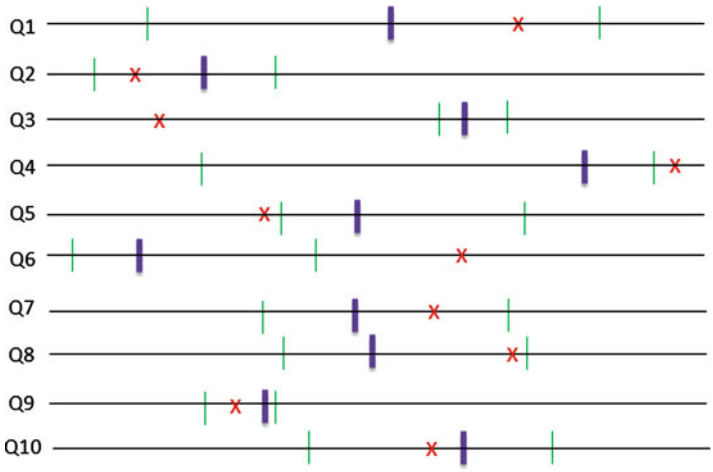


Fig. 2.1 Illustration of an expert's assessment on ten seed variables where the realization is denoted with X, the thin vertical lines represent the 5th and 95th quantiles and the thick vertical lines represent the 50th quantile

Table 2.1 Summary of Performance of Expert on Ten Seed Questions showing higher than expected realizations below the 5th and above the 95th quantile

Quantile	Below 5th	5th to 50th	50th to 95th	Above 95th
Observed proportion of realizations	0.20	0.30	0.30	0.20
Expected proportion of realizations	0.05	0.45	0.45	0.05

From Fig. 2.1 we see that on Q1 the realization was between the 50th and 95th quantile, while of Q2 the realization was between the 5th and 50th quantile. In assessing the accuracy of the experts, we do not distinguish between the proximity of the realization to a quantile, so for example in both Q3 and Q5 the realization is below the 5th quantile, and we are not concerned that it is farther below in Q3 than Q5. We see that 5 of the realizations are above the 50th quantile and 5 below, however there are more realizations than expected in the tails of the distribution, i.e. below the 5th and above the 95th quantile. Table 2.1 provides a summary of the overall performance in terms of the proportion of the realizations observed in the four intervals.

Upon inspecting the data from Table 2.1 we can see that the expert is overconfident: 4 of the 10 realizations were in the tails of the distribution and we would have expected only 1, i.e. 10% of the seed questions. However, we need to measure how extreme the set of realizations is with respect to the expert's distributions, which we do through the Kullback-Leibler (KL) divergence measure, which measures the difference between two probability distributions.

The KL divergence measure is a measure for the difference between two probability distributions. We will use it to measure the difference between the probability distribution specified by the expert and the empirical distribution obtained from the raw frequencies. For our example, this would correspond to comparing the differences between the observed and expected frequencies presented in Table 2.1.

The formula for the divergence measure, denoted by $I(s,p)$, is:

$$I(s,p) = \sum_{i=1}^{n} s_i \ln\left(\frac{s_i}{p_i}\right)$$

where: s_i is the observed proportion of realizations in interval i

p_i is the expected proportion of realizations in interval i

n is the number of intervals

Applying this to our example we obtain:

$$\begin{aligned} I(s,p) &= 0.2\ln\left(\frac{0.2}{0.05}\right) + 0.3\ln\left(\frac{0.3}{0.45}\right) + 0.3\ln\left(\frac{0.3}{0.45}\right) + 0.2\ln\left(\frac{0.2}{0.05}\right) \\ &= 0.31 \end{aligned}$$

If the observed proportions perfectly match the expected proportions then the divergence measure would be 0, as the difference grows so does the measure.

This divergence measure has been extensively studied in mathematical statistics. In particular, if the expert's assessments are statistically accurate, i.e. the long-run observed proportions will equal the expected proportions, then the probability distribution of this measure is related to the χ^2 distribution for large sample sizes. Specifically,

$$\Pr\{2qI(s,p) \le x\} \to \chi^2_{n-1}(x), \quad as \;\; q \to \infty$$

where q is the number of seed questions and $\chi^2_{n-1}(x)$ is the Cumulative Distribution Function (CDF) of the χ^2 distribution with $n-1$ degrees of freedom evaluated at x.

For our example, we have n equal to 4. Figure 2.2 provides an illustration of the Probability Density Function (PDF) of the χ^2_3 distribution. The expert in the example had a divergence measure of 0.31, so $2qI(s,p)$ equals 6.2. The probability of observing a divergence measure in excess of 6.2, i.e. the area under the χ^2_3 curve greater than 6.2, is 0.1. This is the calibration score on which we will compare all experts.

The calibration score used with the CM is the probability of observing a more extreme divergence statistic between specified and observed proportions. The best expert score would be 1, the worst would be 0. The cutoff is imposed by the scoring rule constraint and chosen by optimization, which is explained in Sect. 2.2.5.3.

Figure 2.3 illustrates the relationship between the KL divergence measure and the calibration score for an expert, denoted by C(expert). From the maximum score of 1, where divergence is 0, the score drops quickly as divergence increases. The

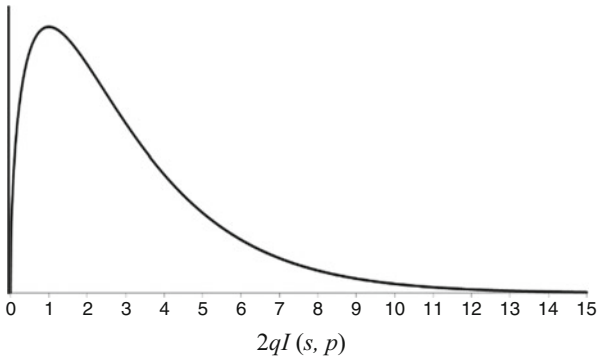


Fig. 2.2 The PDF of the χ^2_3 distribution measuring the variable in the divergence between the experts specified probabilities and observed frequencies

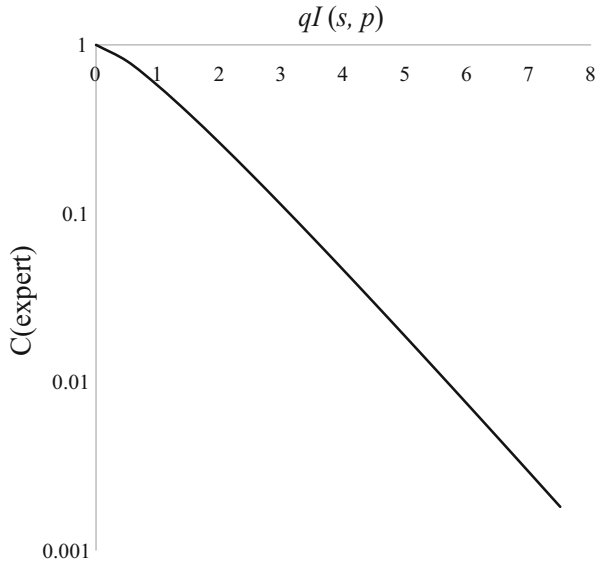


Fig. 2.3 The relationship between the calibration score for an expert, C(expert), and $qI(s,p)$ where q is the number of seed questions and $I(s,p)$ the KL divergence showing a steep drop in the score as divergence increases from 0 is on a log scale)

calibration score is on a log scale, showing an almost linear relationship with $I(s,p)$. Over the range $qI(s,p) \in [1,8]$ we have the approximate relationship of $C(\text{expert}) \approx 1.44e^{-0.9qI(s,p)}$, which we present only to provide an indication on the relationship between these variables. This implies that the logarithm of the ratio of any two calibration scores is approximately $0.9q$ times the difference in their KL divergence measures.

2.2.3 *Distribution and Discrimination of Calibration Score*

The purpose of the calibration score is not to perform hypothesis testing but to discriminate across experts on their level of calibration. We conducted a simple Monte Carlo simulation exercise to assess the difference in calibration score between an expert who was perfectly calibrated with one who was slightly over confident based on ten seed questions.

Specifically, we simulated ten realizations from a multinomial distribution based on a Calibrated Expert (CE) with probabilities (0.05, 0.45, 0.45, 0.05) and ten from an Over Confident Expert (OCE) with probabilities (0.15, 0.35, 0.35, 0.15). For each expert we calculated their calibration score as per Sect. 2.2.2 and evaluated the ratio. We repeated this 10,000 times. The results are presented in Fig. 2.4.

Figure 2.4 has the ratio expressed on a log scale and shows that even with only ten seed questions the calibration score discriminates well between a calibrated and marginally over confident expert. For 75% of the simulations the CE received a higher score (i.e. the ratio was greater than 1), for 50% of the simulations the ratio was in excess of 2.5 and for 10% it was in excess of 87.

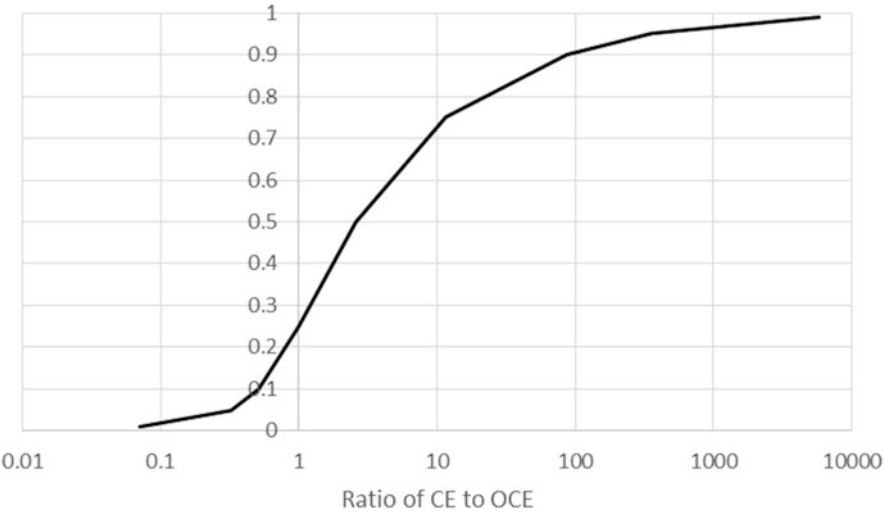


Fig. 2.4 Cumulative Distribution Function of the ratio of the calibration scores of Calibrated Expert (CE) (0.05, 0.45, 0.45, 0.05) to Over Confident Expert (OCE) (0.15, 0.35, 0.35, 0.15) showing the CE receives a higher score (ratio exceeds 1) on 75% of the simulations

2.2.4 Information Score

By providing extremely large quantile intervals, an expert can achieve what appears to be excellent, or even perfect calibration - but they will be totally uninformative by doing so. The ideal expert is both well-calibrated and informative. There are many standard approaches to measure the degree to which a probability distribution is spread out, such as the standard deviation or width of prediction intervals, but these have shortcomings as changes in units of measure (e.g. grams to kilograms) can affect some variables but not others. The CM uses the Kullback-Leibler (KL) divergence measure, as it is scale invariant.

The spread of the experts' distributions is assessed relative to a background range. During the elicitation the expert does not assign a minimum or maximum value, so we need to determine the length of the lower and upper intervals. This is done through the intrinsic range, which is based on the range of judgments on a variable (target or seed) across all experts.

An intrinsic range is determined for each question (seed and target). By default, the intrinsic range overshoots the range spanned by the lowest and highest assessed values by 10%. (The overshoot is chosen by the analyst and affects only the measure of information; a larger overshoot tends to make all information scores similar, a small overshoot boosts the differences.) The informativeness of an expert's probability distribution will be measured using KL divergence measure relative to a uniform distribution applied to the intrinsic range, i.e. the least informative distribution across the collected range of opinion (the loguniform background measure is used for very wide ranges).

Illustrating this with an example, assume we have three experts, each asked to predict the recorded temperature in Toronto at noon on September 15 next year. Each provides their 5th, 50th and 95th quantiles in Celsius. Expert 1 provides (0, 13, 40), Expert 2 provides (0, 37, 40) and Expert 3 provides (10, 13, 20).

The range of elicited quantiles is 40°C, as Expert 1 and 2 had the lowest 5th quantile and coincidentally both had the highest 95th value of 40°C. 10% of the range of assessments is 4°C, which we add and subtract to the range in the experts' assessments to obtain the intrinsic range of (−4°C, 44°C). We compare each expert against the uniform distribution on the intrinsic range, which is illustrated in Fig. 2.5.

Figure 2.5 illustrates the CDF for each expert, assuming a uniform distribution between specific quantiles, and compares this with a uniform distribution across the entire intrinsic range. As the uniform distribution over the intrinsic range is the least informative distribution, we are looking for the distribution that diverges from it the most. Expert 1 appears the least informative of the experts as their CDF is almost the same as the uniform. Expert 3 appears to diverge the most from the uniform, thereby being the most informative.

The KL divergence measure will provide a measure for the degree of divergence. For each interval provided by the expert we assess the probability assigned by the uniform distribution. For expert e we denote the information measure on question i with $I_i(e)$ and calculate it as in the following.

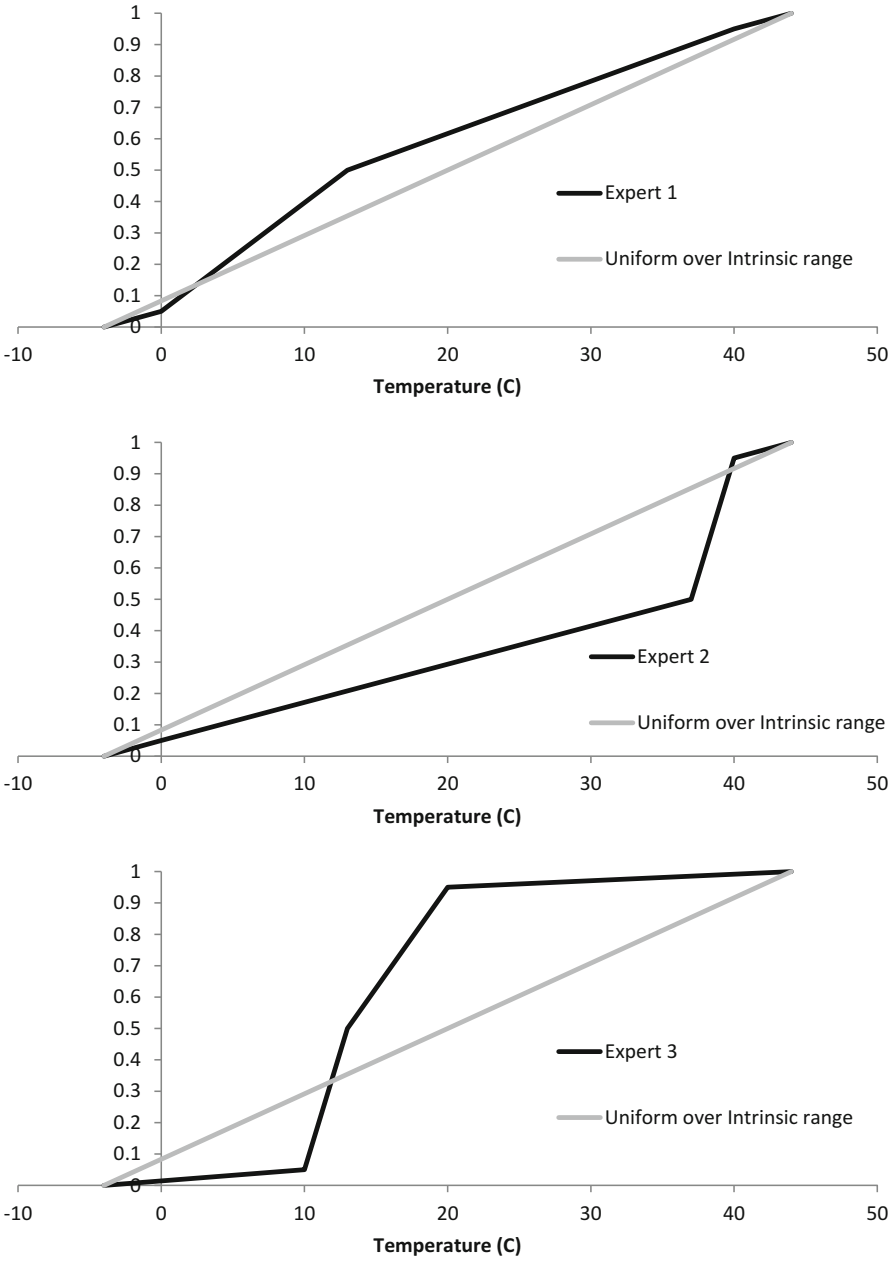


Fig. 2.5 CDF for each expert compared with uniform distribution on intrinsic range showing Expert 1 to be least informative and Expert 3 to be most informative

$$I_i(e) = 0.05 \ln\left(\frac{0.05}{\left(x_{ei1} - x_{i0}\right)/\left(x_{i4} - x_{i0}\right)}\right) + 0.45 \ln\left(\frac{0.45}{\left(x_{ei2} - x_{ei1}\right)/\left(x_{i4} - x_{i0}\right)}\right)$$
$$+ 0.45 \ln\left(\frac{0.45}{\left(x_{ei3} - x_{ei2}\right)/\left(x_{i4} - x_{i0}\right)}\right) + 0.05 \ln\left(\frac{0.05}{\left(x_{i4} - x_{ei3}\right)/\left(x_{i4} - x_{i0}\right)}\right)$$

where x_{i0}, x_{i4} are the lower and upper bound of the intrinsic range for question i and $x_{ei1}, x_{ei2}, x_{ei3}$ are the 5th, 50th and 95th quantiles provided by the expert. Note that we do not use realizations in this formula, so this can be calculated for seed and target questions (although the lower and upper bound of the intrinsic range do depend on the realizations, if they exist).

Illustrating the calculation for Expert 1 consider the following. Note that the intrinsic range is 48°C in length.

$$I\,(\text{Expert 1}) = 0.05 \ln\left(\frac{0.05}{0.08}\right) + 0.45 \ln\left(\frac{0.45}{0.27}\right) + 0.45 \ln\left(\frac{0.45}{0.56}\right) + 0.05 \ln\left(\frac{0.05}{0.08}\right)$$
$$= 0.0770$$

Inspecting the calculation, we see there are four items being summed, one for each interval. The first interval, for which Expert 1 assigned probability 0.05, was 4°C in length so the uniform distribution would assign probability 0.083, i.e. 4/48. Similar assessments are made for each interval. If the expert assigns a probability to an interval greater than the proportion of that interval on the intrinsic range, then the ratio is greater than 1 and the item being summed makes a positive contribution to the summation. Likewise, if the ratio is smaller than 1 then a negative contribution is made. The summation will always result in a non-negative number and will be 0 only if the expert's distribution coincides with the uniform distribution (i.e., the least informative distribution).

Consider the information score for Expert 3, who is more informative than Expert 1.

$$I\,(\text{Expert 3}) = 0.05 \ln\left(\frac{0.05}{0.29}\right) + 0.45 \ln\left(\frac{0.45}{0.06}\right) + 0.45 \ln\left(\frac{0.45}{0.15}\right) + 0.05 \ln\left(\frac{0.05}{0.50}\right)$$
$$= 1.1921$$

Inspecting this calculation, we see that Expert 3 performs very well in the two middle intervals, assigning probability of 0.45 to intervals with length of 6% and 15% of the intrinsic range. While the first and last intervals perform poorer than with Expert 1, these intervals are weighted at only 5% of the sum.

Lastly, Expert 2 received an information score of 0.5951, unsurprisingly this is not as good as Expert 3 but better than Expert 1. It is interesting to note that Expert 2 had provided the same 5th and 95th quantiles as Expert 1 but scored much better due to the difference in the 50th quantile. This resulted in much higher concentration

of probability between the 50th and 95th quantile, where the uniform distribution assigned only 0.06 compared with 0.45.

2.2.5 Weights

The outcome of the CM is a weighted average across all experts' Cumulative Distribution Functions (CDF), which we refer to as the CDF of the Decision Maker (DM).

$$P_{DM}\left(X \leq x\right) = \sum_{i=1}^{n} w_i P_i\left(X \leq x\right)$$

where: $P_i(X \leq x)$ is the CDF provided from expert i.

w_i is the weight assigned to expert i

$P_{DM}(X \leq x)$ is the CDF of the Decision Maker

An expert's weight is derived from the product of their calibration score and information score. We will describe three methods for obtaining the weights, namely, Global, Itemized and Optimized.

2.2.5.1 Global Weights

Global weights are obtained through averaging an expert's information score across all seed questions.

We first calculate the average information score: $I\left(e_i\right) = \frac{\sum_{j=1}^{q} I_j(e_i)}{q}$.

Raw weights are then calculated removing any expert whose calibration score is below acceptable levels, typically 0.01.

$$w_i' = C\left(e_i\right) \times I\left(e_i\right) \times 1_\alpha\left(C\left(e_i\right)\right)$$

where: $1_\alpha(C(e_i))$ is 1 if the calibration score $C(e_i)$, exceeds the cutoff threshold α and 0 otherwise. The use of $1_\alpha(C(e_i))$ is imposed by the requirement that the weights w'_i should be an *asymptotically strictly proper scoring rule*: an expert maximizes his(her) long run expected weight if and only if his(her) quantile assessments correspond to his(her) true beliefs.

Weights are then normalized across all experts.

$$w_i = \frac{w_i'}{\sum_{\forall k} w_k},$$

These weights are applied to all target variable assessments as well as the seed variables, which enables the same performance measures to be calculated to assess

the quality of the Decision Maker (DM). The calibration score will vary much more between experts than the information score and, as such it is the calibration score that will drive the differences in weights.

2.2.5.2 Itemized Weights

Unlike the calibration score which necessarily depends on performance across a set of seed variable assessments, the information score is calculated on a question by question basis and as such can vary as an expert provides a narrower range relative to the other experts for some questions as opposed to others. The itemized weights approach allows the weight to change between questions.

Similar to the Global weights, the raw weight for expert i on question j is calculated as:

$$w'_{ij} = C(e_i) \times I_j(e_i) \times 1_\alpha(C(e_i))$$

As such we have two subscripts, one for the expert and one for the question. If there is little relative variation in the information score between experts across questions, then the assessment will be similar to the Global weights method. Note that the criterion for including an expert's assessments in the combination, i.e.$1_\alpha(C(e_i))$, is not influenced by the information score.

2.2.5.3 Optimized Weights

Requiring the unnormalized weights to be strictly proper scoring rules imposes a cutoff α, but it does not say what the value of α should be. The value of α is chosen to optimize the unnormalized weight of the DM. Starting with $\alpha = 0$, the value is successively raised and the global weight of the DM successively recomputed, whereupon the value α^* is chosen which returns the DM with highest global weight.

Studies have shown that optimally choosing α has a significant impact on the performance of the DM. It is therefore important that the use of a cutoff be explained properly to experts and problem owners. From the above it is clear that α is NOT chosen to accept or reject experts. In most cases α^* bears no resemblance, or probative relationship, to the value 0.05 commonly used in simple hypothesis testing. Unweighted experts are not "bad" and are not "rejected." Unweighting an expert *can* mean that, but it doesn't *necessarily* mean that. Some experts are justified bad. The value of unweighted experts for the study becomes evident if robustness analysis on the choice of experts is performed: successively excluding experts and re-running the analysis typically shows that the DM still performs well after redistributing weight over the non-excluded experts—sometimes the DM's score even improves.

Knowing that the quality of the DM would not change significantly if any given expert were excluded (or hadn't turned up) aids in the acceptance of the study

results. Robustness analysis is routinely performed on both the choice of experts and on the choice of seed variables.

2.2.6 Summary

This section has provided a brief introduction to the CM. We have discussed its purpose and explored its details, focusing on how calibration and information are measured and weights determined. More analysis on the performance data of a given study is possible, which we have not considered here. In particular, the robustness of the results can be explored in more depth. The sensitivity of the performance of the DM to specific seed questions can be assessed, identifying whether an expert has received a particularly high or low weight due to their performance on a single question and assessing the appropriateness of that question. In the following section, we explore seed questions in more depth.

2.3 Finding Seed Variables

Seed variables serve three purposes in the CM. First, they demonstrate that quantifying uncertainty as subjective probability is a science-based activity, as an experts's hypotheses for the set of seed questions are falsifiable based on the observed realizations. Second, they allow for the measurement of an expert's performance as an uncertainty assessor. Third, they enable performance-based combinations of experts' judgments and hopefully validate the legitimacy of that combination.

If a field is scientific, then there are observations and measurements underlying the experts' judgments, and this information can be mined for seed variables. Seed variables should be items for which an expert on the variables of interest can, reasonably, be expected to provide an informed judgement. The thematic link between the seed variables and variables of interest should be strong enough that an expert or problem owner in the field accepts differential weighting of experts—with similar specialist qualifications and knowledge in the topic—based on their varying abilities to quantify uncertainty on the seed questions.

Seed questions are classified in two ways: predictions versus retrodictions, and domain versus adjacent (Cooke and Goossens 1999; European Food Safety Authority 2014). Predictions are questions about future quantities that will be observed or measured within the timeframe of the study, and retrodictions are based on previously collected data. Domain questions are in the same field of expertise or use the same physical dimensions as the variables of interest, whereas adjacent questions are related to but slightly different from the variables of interest. Although there is minimal research on what constitutes a good seed question, it is generally accepted from practical experience in many elicitations that domain predictions are

the ideal, but they can be tough to find in practice. Domain retrodictions or adjacent predictions are the next best target, with adjacent retrodictions falling third.

If a study involves experts from different fields, a seed question could be considered "domain" for one expert and "adjacent" for another. For example, in an expert judgment study on foodborne disease, experts could have backgrounds in epidemiology, public health field work, or food safety. Again, actual experience, and some unpublished tests with sub-groups of a number of panels, suggest that generally a well-calibrated individual is usually well-calibrated on adjacent seeds, and *vice versa* for low weight individuals.

Thus, finding good seed variables is an art. Although they should be closely linked to the variables of interest, seed variables are not a test of the expert's subject matter expertise and they should not be queried about the sort of values that are established 'constants' or well-known to all experts in their professional domain; testing an individual's recall is not appropriate for calibrating skill in judging uncertainty. Experts can become frustrated when they feel like they should know *the answer* to the question because they have come across it before. There are several strategies for finding seed variables. Among the more common are utilizing:

A. Results of future measurements that will be performed within the study's time frame
B. Unpublished measurement results
C. Querying relevant though unfamiliar features of standard databases (e.g., censoring rates)
D. Combining or comparing values from disparate datasets.

Of these strategies, only A provides prediction seed variables; the other three yield retrodictions.

These varieties of seed questions are illustrated below with examples.

A. Results of Future Measurements

Potential Damage from Asian Carps in the Great Lakes

In a study for the U.S. Environmental Protection Agency (EPA), researchers at the University of Notre Dame and Resources for the Future (RFF) used expert elicitation to quantify the future impacts, with uncertainty, of Asian carp (silver and bighead) establishment on the food web of Lake Erie and to evaluate the efficacy of strategies to prevent their establishment (Wittmann et al. 2014; Cooke et al. 2014; Wittmann et al. 2015). These two species have recently dispersed to waterways directly connected to the Great Lakes and may cause substantial ecological and economic damage. Seed variables were based on future measurements about the Great Lakes ecosystem contained in a report released once a year; previous measured values were supplied for convenience. Eleven experts quantified the 5, 50, and 95 percentiles of their subjective probability distributions for fifteen seed variables. Example questions include:

- What was the total harvest (in tons; 1 ton = 1000 kg) of yellow perch in Lake Erie in 2010?
- What was the abundance (number of fish) of walleye in Lake Erie in 2010?

- What percentage of Lake Erie eastern basin lake trout (lean strain) contained round goby in their stomach contents in 2010?

B. Unpublished Measurement Results

Nitrogen Removal in the Chesapeake Bay

Managing nitrogen removal requires an ecosystem-based approach. Data are needed to define Best Management Practices (BMPs) for engineered structures so that natural resource managers can meet prescribed water quality targets and other ecosystem service needs. Given the rapid rate of urbanization, managers must often quickly make decisions on where, which, and how many BMPs to implement within a catchment area. Empirical data to inform decisions about the types and optimal placement of BMPs remain problematic in urbanized areas such as the Chesapeake Bay, and high levels of uncertainty about the effectiveness of BMPs in nitrogen-removal impedes deployment decisions. The U.S. EPA, the University of Maryland, and RFF conducted an expert elicitation to quantify, with uncertainty, the performance of various urban stormwater management structures under a variety of rain events (Koch et al. 2015). Nine experts assessed 5, 50 and 95 percentiles for eleven seed variables based on measured but unpublished values for actual rain events. Data for the seed variables were collected by one of study's collaborators, and the protocol described the conditions and details of the watershed area and rain event in which they were measured, including hydrographs and the precipitation record. Below is one example question regarding the Piedmont watershed.

What is the outgoing total nitrogen load (kg TN) from the sub-watershed over the entire duration of rain event I?

C. Unfamiliar Features of Standard Databases

Breastfeeding and Cognitive Development

The long-term effects of nutritional interventions are notoriously difficult to assess in well-controlled randomized blinded trials. Conventional longitudinal studies frequently underlie policy recommendations, although confounding always poses threats to findings based on such data. Evaluating the long-term effects of breastfeeding exemplifies the challenges posed by using non-randomized longitudinal data sets as key variables such as mother's IQ, mother's income, birth order, and length of time breastfeeding are all highly correlated. Under contract with the Bill and Melinda Gates Foundation, the University of Virginia and Resources for the Future applied structured expert judgment to evaluate the effects of the duration of exclusive and any breastfeeding on cognitive performance, measured as IQ (Colson et al. 2016). Specifically, seven experts were asked to quantify their uncertainty about outcomes of a hypothetical fully randomized trial and on 11 seed variables. Seed questions were based on new calculations with the survey data most frequently used in the breastfeeding and IQ literature. The 5, 25, 50, 75 and 95 percentiles were elicited, and seed questions included the following.

1. In NLSY79-C (National Longitudinal Study of Youth data base) the average Peabody Picture Vocabulary Test (Revised Form L) (PPVT) mean score, among

the children with scores, is 90.660. What is the average among first-born children with at least one PPVT score?
2. In NLSY79-C the average PPVT mean score, among the children with scores, is 90.660. What is the average among first-born children who were ever breastfed?
3. In NLSY79-C, 1706 children have PPVT scores recorded for 1986 and Peabody Individual Assessment Test math scores for 1986. What is the correlation among these scores?
4. In NLSY79-C, 1700 children have PPVT scores recorded for 1986 and Peabody Individual Assessment Test reading recognition scores for 1986. What is the correlation among these scores?
5. In what percentage of the 11,512 records in NLSY79-C is the Peabody Picture Vocabulary Test (PPVT) never reported?
6. In NLSY79-C the average age in weeks when breastfeeding ended is 9.12. What is the average age in weeks when breastfeeding ended among the 1583 only children who were breastfed?
7. In the 2005–06 Demographic Health Survey for India, what is the 75th percentile for duration of breastfeeding (in months), among children who were breastfed and who were not still breastfeeding at the time of the survey? This data excludes children who died while breastfeeding
8. The U.S. Panel Study of Income Dynamics Child Development Supplement (PSID-C) data set has 3563 records. In what percentage of completed records is the sum of Woodcock-Johnson Word Scores and Woodcock-Johnson Applied Problem Scores in 1997 greater than in 2002?

D. Combining or Comparing Datasets

Mortality Impact of Fine Particulate Matter

The Kuwait Oil Fires of 1991 emitted vast quantities of fine particulate matter and related gases to the atmosphere. A team from Harvard University and the TU Delft was asked to estimate the mortality impacts of exposure to oil fire smoke to support the State of Kuwait's environmental reparations claims (Evans et al. 2005; Tuomisto et al. 2005; Wilson et al. 2005). The primary goal of the elicitation was to probabilistically characterize the number of deaths attributable to the oil fires. Six European experts quantified the 5, 25, 50, 75, and 95 percentiles of their subjective probability distributions for 12 seed variables which were found by combining existing datasets. The following are representative:

1. On how many days in 2001 did the daily average PM_{10} concentration exceed 50 $\mu g/m^3$ at one or more of the above London stations (max 365)?
2. On how many days in 2001 did the daily average PM_{10} concentration fall below 30 $\mu g/m^3$ at all of the above London stations (max 365)?
3. On how many days in 1997 did the daily average PM_{10} concentration exceed concentration exceed 50 $\mu g/m^3$ at least one of the above London stations (max 365)?
4. On how many days in 1997 did the daily average PM_{10} concentration fall below 30 $\mu g/m^3$ at all of the above London stations (max 365)?

5. On how many days in 2001 did the daily average PM_{10} concentration exceed 50 $\mu g/m^3$ at least one of the above Athens stations (max 365)?
6. What is the ratio: Number of non-accidental deaths in the week (7 days starting from January 1st) of 2000 with the highest average PM_{10} concentration/Weekly average number non-accidental deaths in 2000.
7. What is the ratio: Number of cardiovascular deaths (ICD10 Cause I) in the week (7 days starting from January 1st) of 2000 with the highest average PM_{10} concentration /Weekly average number cardiovascular deaths in 2000

2.4 Elicitation Styles

In addition to getting experts to quantify their uncertainty as subjective probability distributions, the elicitation process involves providing the experts with some training and, ideally, should also involve a procedure for capturing the qualitative reasoning behind the experts' assessments. Elicitation is typically done by one or two facilitators, with the ideal being one facilitator who is well-versed in elicitation practices and can manage the process—and the participating experts—neutrally and fairly, and a second person with extensive knowledge in the subject area of that particular study. On occasion, the latter may be the problem owner—i.e. the person who commissions the elicitation and uses the findings.

Expert training should include an introduction to structured expert judgment and an explanation of the motivation for using expert judgment in the specific problem at hand. Names and affiliations will be published but not associated with individual assessments. This linkage will be preserved to enable competent peer review if required at some later date. The facilitator should also explain the use of seed variables and the performance measures of the CM. Training includes a discussion of subjective probabilities as a method of quantifying uncertainty, with at least one example discussed in detail. The facilitator should warn the expert, or the group if a one-off "plenary" elicitation is being undertaken, about overconfident responses (i.e., a set of assessments with high information but low statistical accuracy scores) and explain that less informative but statistically accurate assessments are much more useful to the analyst and problem owner than informative but inaccurate assessments. Finally, training should involve walking the expert or panel through a few example questions. Typically, the search for seed variables yields about 15–20 questions of which the best 10 to 15 are used in the actual study, the others can be used for training. With the training questions, the expert provides the specified percentiles from her subjective probability distribution for each question, and the facilitator immediately provides the true value for each of these questions. These questions help the experts understand how to think about their uncertainty as subjective probability and give them a little bit of feedback on their performance before elicitation on the real questions begins.

Throughout the elicitation, in addition to providing values from their subjective probability distributions, experts should also explain their thinking on each question,

particularly the target variables of interest. This can be done through written comments or discussion, with the facilitator or a rapporteur taking notes; some elicitations may be recorded (with the agreement of the participants) and the contributions transcribed later; the latter is a gold standard approach, but adds to costs. Facilitators should ask experts to explain any existing data, assumptions, scenarios, or other information that informed the expert's uncertainty for each question. Capturing this qualitative information is an important part of the elicitation process. The qualitative rationales can help the analyst and problem owner understand the results of the study, and comparing the rationales of different experts can illuminate differences in assessments. Occasionally, through the rationales it emerges that different experts interpreted a question differently, and the facilitator can follow-up with experts to clarify any relevant issues. In most cases, the collected rationales from experts become part of the published record of a study, though they are also kept anonymous.

Following an elicitation session, the facilitator or analyst should provide some feedback to the expert or expert panel. At a minimum, this should include sharing the final decision maker assessments with the experts for their review. If a facilitator took notes on the experts' rationales during the elicitation process, each expert should also have a chance to review her rationale to confirm it is complete and correctly captures her thinking. In some cases, feedback may also include sharing the realizations for the seed questions or giving each expert her individual performance scores.

Experts' subjective probability distributions can be elicited within the CM framework via a group plenary session, in-person in one-on-one interviews, or one-on-one remote sessions. Plenary sessions consist of the experts and facilitator(s) all meeting together for the training, elicitation, and possibly feedback. A plenary elicitation session begins with the motivation for the study and training. Experts then individually work through the elicitation protocol, and the session may conclude with the facilitator feeding back results to the experts. Individually working through the questions in a group setting allows the experts and facilitator to discuss any ambiguity in the questions, ensuring the experts interpret everything in the same way, as far as possible. Plenary elicitations may require less facilitator time, since all the participating experts work through the elicitation protocol simultaneously. Plenary elicitations are logistically more difficult to arrange, though, and may be impossible for geographically dispersed experts, especially if the elicitation needs to be completed within a specific, sometimes urgent, timeframe. Plenary elicitations risk suffering from negative group dynamics, like groupthink or strong personalities having an outsized influence on discussion, and an accomplished facilitator is needed to prevent these issues and individual propensities from unduly influencing the experts' assessments. Finally, plenary sessions allow for less interaction between the facilitator and each individual expert, which can make it harder for the facilitator to accurately capture an expert's rationale, to challenge the expert's thinking, and to make certain the expert accurately interprets the question and understands the relevant uncertainty. This said, as noted above and if appropriate, the proceedings of plenary sessions can be recorded to obtain a complete account of deliberations.

One-on-one or two-on-one in-person interviews involve the facilitator(s) interviewing and eliciting probabilities from each expert individually. The interview may be preceded by a group training session, conducted either in-person or by webinar, or the interview can begin with individual training. In-person expert interviews are easier to schedule than a group elicitation session, but they require more time from the facilitators who must travel to each expert and individually walk them through the elicitation protocol. Individual expert elicitations are not subject to the potentially troublesome group dynamics that can affect plenary sessions. In individual elicitations, however, the experts cannot discuss the questions to make certain everyone understands them to have the same meaning (experience with plenary elicitations suggests this is a common and significant potential shortcoming). Dry-runs to identify and clarify ambiguity or misspecification in the questions are especially important for one-on-one elicitations so that all potential issues are resolved before the first expert is interviewed. If something new emerges after one or more interviews are complete, the facilitator may need to re-elicit probabilities from experts who responded earlier. In a study with one-on-one elicitations, feedback of the results to the experts happens after all the expert interviews are complete.

Increasingly, as online meeting and communication tools improve, there is growing interest in conducting elicitations remotely. When the experts are spread across the globe, remote elicitations are much cheaper and easier to arrange than plenary sessions or in-person interviews. Remote elicitations can use software and online tools that allow the expert and facilitator to communicate by video and share computer screens or documents so that one person can enter values into the elicitation protocol and the other instantly sees them, to enable smooth communication and make sure all data is captured correctly. Training can also be done remotely, although it may not be as effective as in-person training. Experts may feel less invested in an elicitation conducted remotely, they may provide less qualitative information than they would during an in-person conversation, and remote elicitations may be subject to more interruptions than in-person meetings. Additional research is needed to understand if the quality of expert assessments degrades in a remote elicitation. As the demand for structured expert judgment grows and remote meeting tools improve, however, new best practices specific to remote elicitations are needed.

2.5 Discussion

The core principle of the CM that makes it distinctive and unique is external and empirical validation. Each expert is assessed on their ability to express their uncertainty in probabilistic terms; this ability is measured and compared across the pool of participating experts, rewarding the better experts with higher weight in the final assessment. As such, the quality of seed variables is paramount and within this chapter we have provided a detailed discussion of four different approaches to

developing these questions. While finding appropriate seed variables does add to the preparation work and design of an elicitation, the calibration function they provide substantially improves the quality of the ensuing target item combination results, see Chapter 3: "Validation in the Classical Model" of this book (Cooke 2018).

We have described the two key performance measures in the CM, i.e. calibration and information scores, and illustrated with simple examples to show these to behave as intuitively sensible performance measures. There have been, to date, well over 200 professional expert elicitation panels that have successfully—and efficaciously—utilized seed variables to provide rational consensuses on important issues of scientific, engineering and medical concern.

References

Colson AR, Cooke RM, Lutter R (2016) How does breastfeeding affect IQ? Applying the classical model of structured expert judgment. Resources for the future. http://www.rff.org/research/publications/how-does-breastfeeding-affect-iq-applying-classical-model-structured-expert

Cooke RM (2018) Validation in the classical model. In: Dias LC, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgment. Springer, New York. 2018 (Chapter 3 in this book)

Cooke RM, Goossens LJH (1999) "Procedures guide for structured expert judgment." EUR 18820. Delft University of Technology, Delft, The Netherlands. https://cordis.europa.eu/pub/fp5-euratom/docs/eur18820_en.pdf

Cooke RM, Wittmann ME, Lodge DM, Rothlisberger JD, Rutherford ES, Zhang H, Mason DM (2014) Out-of-sample validation for structured expert judgment of Asian carp establishment in Lake Erie. Integr Environ Assess Manag 10(4):522–528. doi:10.1002/ieam.1559

Cooke RM (1991) Experts in uncertainty: opinion and subjective probability in science. Oxford University Press, New York

European Food Safety Authority (2014) Guidance on expert knowledge elicitation in food and feed safety risk assessment. EFSA J 12(6):3734. doi:10.2903/j.efsa.2014.3734

Evans JS, Wilson A, Tuomisto JT, Tainio M, Cooke RM (2005) What risk assessment can tell us about the mortality impacts of the Kuwaiti oil fires. Epidemiology 16(5):S137–S138

Gosling JP (2018) SHELF: the Sheffield elicitation framework. In: Dias LC, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgment. Springer, New York. 2018 (Chapter 4 in this book)

Koch BJ, Febria CM, Cooke RM, Hosen JD, Baker ME, Colson AR, Filoso S et al (2015) Suburban watershed nitrogen retention: estimating the effectiveness of stormwater management structures. Elementa 3(July):63. doi:10.12952/journal.elementa.000063

Montibeller G, von Winterfeldt D (2018) Individual and group biases in value and uncertainty judgments. In: Dias LC, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgment. Springer, New York. 2018 (Chapter 15 in this book)

Tuomisto JT, Wilson A, Cooke RM, Tainio M, Evans JS (2005) Mortality in Kuwait due to PM from oil fires after the Gulf War: combining expert elicitation assessments. Epidemiology 16(5):S74–S75

Wilson AM, Eschenroeder AQ, Evans JS (2005) Final human health risk assessment: mortality risks from oil fire particulate matter exposure. In: Harvard school of public health report to the Kuwait public authority for assessment of compensation for damages from the Iraqi aggression. Harvard School of Public Health, Boston

Wittmann ME, Cooke RM, Rothlisberger JD, Lodge DM (2014) Using structured expert judgment to assess invasive species prevention: Asian Carp and the Mississippi—Great Lakes hydrologic connection. Environ Sci Technol 48(4):2150–2156. doi:10.1021/es4043098

Wittmann ME, Cooke RM, Rothlisberger JD, Rutherford ES, Zhang H, Mason DM, Lodge DM (2015) Use of structured expert judgment to forecast invasions by bighead and silver carp in Lake Erie. Conserv Biol 29(1):187–197. doi:10.1111/cobi.12369

Chapter 3
Validation in the Classical Model

Roger M. Cooke

Abstract Validation is the hallmark of science. For expert judgment to contribute to science-based uncertainty quantification, it must become amenable to empirical validation. Using data in which experts quantify uncertainty on variables from their fields whose true values are known post hoc, this chapter explains how validation is performed in the Classical Model for structured expert judgment and reviews results for different combination methods.

3.1 Introduction: Why Validate?

Expert Judgment (EJ) encompasses a wide variety of techniques ranging from a single undocumented opinion, through preference surveys, to formal elicitation with external validation. In the nuclear safety area, Rasmussen et al. (1975) formalized EJ by documenting all steps in the expert elicitation process for scientific review. This made visible wide spreads in expert assessments and teed up questions regarding the validation and synthesis of expert judgments. The nuclear safety community later took onboard expert judgment techniques driven by external validation (Cooke 2012a, b; Oppenheimer et al. 2016). Most recently, the National Academy of Science report on the Social Cost of Carbon spotlights the role of performance measurement "performance-weighted average of distributions usually outperforms the simple average, where performance is again measured again by calibration and informativeness" (NAS 2017, p. 339).

External validation is the hallmark of science, and is the main driver of the *Classical Model for Structured Expert Judgment* (SEJ). It has been deployed

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-319-65052-4_3) contains supplementary material, which is available to authorized users.

R.M. Cooke (✉)
Resources for the Future, 1616 P Street NW, Washington, DC, 20036, USA

Department of Mathematics, Delft University of Technology, Mekelweg 2, 2628 CD, Delft, Netherlands
e-mail: cooke@rff.org

L.C. Dias et al. (eds.), *Elicitation*, International Series in Operations Research & Management Science 261, https://doi.org/10.1007/978-3-319-65052-4_3

extensively in areas ranging from nuclear safety, investment banking, volcanology, public health, ecology, and aeronautics/aerospace. Applications are overviewed in Cooke and Goossens (2008), Eggstaff et al. (2014) and in Sect. 3.4. A Wikipedia page gives a good introduction (https://en.wikipedia.org/wiki/Structured_expert_judgment:_the_classical_model).

The classical model validates probabilistic forecasting performance in terms of statistical accuracy (sometimes called calibration) and information (Cooke et al. 1988; Cooke 1991). Since "calibration" causes confusion among engineers and is only loosely defined in decision theoretic literature, the term "statistical accuracy" is now used in its stead. *Statistical accuracy* is measured as the p-value at which one would falsely reject the hypotheses that an expert's probability assessments were statistically accurate. *Informativeness* is measured as Shannon relative information with respect to a user supplied background measure. The *combined score* is the product of the former two. Shannon relative information is used because it is scale invariant, tail insensitive, slow, and familiar. Parenthetically, information measures with physical dimensions, such as the standard deviation, or the width of prediction intervals are scale dependent: a change of units (meters to kilometers) would affect some variables but not others. The combined score satisfies a long run proper scoring rule constraint, and involves choosing an optimal statistical accuracy threshold beneath which experts are unweighted, see Chapter 2: "Elicitation in the Classical Model" in this book (Quigley et al. 2017), and the online appendix to this chapter (Cooke 2017).

Experts' combined scores are used to form a weighted average of experts' distributions, sometimes termed a weighted linear pool. Other pooling models have been proposed, without the benefit of benchmarking with expert judgment data. Section 3.2 discusses pooling and benchmarks pooling methods against empirical data.

Cooke and Goossens (2008) published the results of 45 professionally contracted SEJ studies and made this data available to the research community. Using this data, Clemen (2008) introduced the issue of out-of-sample validation. Variables of interest are typically unobservable on relevant time scales (a few exceptions are found in Cooke and Goossens 2008), and out-of-sample validation usually comes down to cross validation. Cross validation involves splitting the calibration variables, whose values are known post-hoc, into a training set and a test set. The models are initialized on the training set, then scored for performance on the test set. The studies published in 2008 contain many from the dawn of SEJ with wildly different designs. The number of experts ranged from four to seventy-seven, the number of calibration variables from five to fifty-five. Eggstaff et al. (2014) exhaustively cross validated all these studies. SEJ continues to expand, thanks in no small part to (Aspinall 2010). The more recent studies are better resourced, better executed and better documented than the very early studies. Thirty-three independent professionally contracted expert judgment studies have been performed between 2006 and 2014 in which panels of four to twenty-one experts assessed between seven and seventeen calibration variables from their fields. These studies have been recently cross validated (Colson and Cooke 2017). Cross-validation research is summarized in Sect. 3.3. Methods for identifying calibration variables

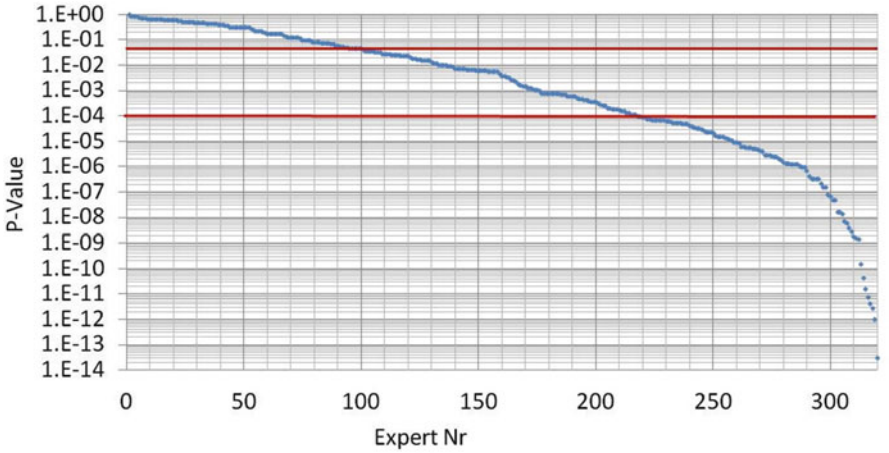


Fig. 3.1 Expert statistical accuracy scores in post 2006 data

are discussed in Chapter 2: "Elicitation in the Classical Model" in this book (Quigley et al. 2017).

Two sets of studies were not used in cross validation research. An ongoing expert elicitation program at the Montserrat Volcano Observatory (Aspinall et al. 2002; Wadge and Aspinall 2014) and a very large scale study by the World Health Organization (Aspinall et al. 2015; Hald et al. 2015; Hoffmann et al. 2016) both produced a wealth of data on expert performance. Since both sets of studies involve heavily overlapping expert panels, they do not lend themselves to cross validation analysis where the panels are considered independent.

Expert judgment data provides compelling answers to the question 'why validate?'. Figure 3.1 shows the statistical accuracy scores of the 320 experts in the 2006–2014 data, arranged from best to worse. 227 of the 320 experts have a statistical accuracy score less than 0.05, which is the traditional rejection threshold for simple hypothesis testing. Half of the experts score below 0.005, and roughly one third fall into the abysmal range below 0.0001. These numbers challenge the assumption that the predicate "expert" is a sufficient guarantee of quality, with regard to uncertainty quantification. There is however, a bright side: 93 of the 320 experts would not be rejected, as statistical hypotheses, at the 5% level. 25 of the 33 studies have at least one, and usually two or more experts whose statistical accuracy is acceptable. Those who eschew validation seem committed to the '*random expert hypothesis*' according to which one expert is as good as another and differences in expert performance are random fluctuations around a mean. On this hypothesis, combining experts on the basis of performance is just groping in the noise.

A recent study for the US Geological Service involving 32 experts and 18 calibration variables (this study is not included in the 2006–2014 data, publication in preparation) provides a nice test for the 'random expert' hypothesis. Figure 3.2 shows the statistical accuracy scores of the top five experts in the real data and in randomized data whereby 32 "Random experts" randomly pick their assessments for

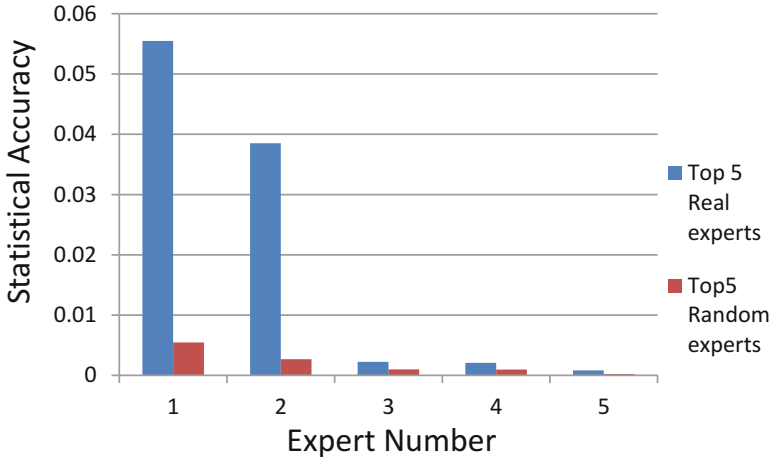


Fig. 3.2 Real and random expert statistical accuracy scores compared for US Geological Service data

each of the 18 variables, without replacement, from the original expert assessments for that variable. Were the random expert hypothesis true, then there would be no reason to expect that the best real experts are different from the best random experts. In fact, the best real experts are better than random. (The worst real experts are also worse than random: the standard deviation of the real experts is ten times that of the random experts). Of course this is but one data set. The cross validation work presented in Sect. 3.3 may be seen as a comprehensive test of the random expert hypothesis.

Simply identifying the best experts and relying on them would be a big improvement over unvalidated expert judgment. Indeed, Philip Tetlock's Good Judgment Project, the reputed winner of IARPA's 5 year forecasting tournament, radically down-selected a small set of "superforecasters" from a pool of more than 3000,[1] based on their performance.

Once we have taken that step, looking for an optimal combination of expert judgments based on their performance is inevitable.

Mathematicians and statisticians are inclined to see expert combination as a mathematics problem, as if the axioms of probability will tell us how it must be done. A thorough study of foundations teaches that expert combination is more akin to an optimization problem in engineering. A bicycle obeys Newton's laws but does not follow from them. It is designed to optimize performance under constraints. The classical model views expert judgment combination as a tool for enabling

[1]Full documentation is not available at this writing and the information here is based on http://www.npr.org/sections/parallels/2014/04/02/297839429/-so-you-think-youre-smarter-than-a-cia-agent accessed 1/12/2017 and Ungar et al. (2012).

rational consensus by optimizing performance measures under mathematical and decision theoretic constraints. Details are found in Wittmann et al. (2014) and Cooke (2015).

The following section compares two popular mathematical aggregation techniques with performance based combinations. Section 3.3 reviews out-of-sample validation. Section 3.4 overviews applications to date. An on-line appendix gives updated proofs of the scoring rule properties of weights in the classical model.

3.2 Mathematical Pooling: Harmonic, Geometric and Arithmetic Means

The classical model for SEJ employs weighted arithmetic averaging, often termed weighted linear pooling. Other pooling models have been proposed and a review of their performance is a useful introduction to the classical model.

Performance based weighting of expert judgments takes effort on the part of the analyst in developing suitable calibration questions, and also on the part of the expert in answering them. Simply averaging experts' probability distributions is much easier and has been widely used. This is termed Equal Weighting (*EW*). Geometric averaging, or Geometric Weighting (*GW*) has been advocated as being "independence preserving" (Laddaga 1977) and "externally Bayesian" (Genest and Zidek 1986). Geometric averaging tends to concentrate mass in regions where the experts agree. This tendency is even more pronounced with harmonic averaging or Harmonic Weighting (*HW*).

Harmonic weighting, under the moniker "averaging quantiles" has found recent adherents. Lichtendahl et al. (2013) suggest that averaging experts' quantiles might give a better decision maker than an equal weight, or "averaging probabilities" combination of their distribution functions. They note that *HW* is "sharper" than *EW*. Flandoli et al. (2011) also used this technique in their analysis of the Classical Model (CM). Gillingham et al. (2015) is a recent example. Averaging quantiles is easier to compute than averaging distributions, and is frequently employed by unwary practitioners.

This section shows that averaging quantiles is equivalent with harmonically averaging densities. The performances of *HW*, *EW* and Performance Weighted (*PW*) combinations are then compared on the thirty-three professional expert judgment studies since 2006. Since *EW* and *HW* bracket geometric averaging, and since none of these techniques utilizes expert performance, this comparison does not consider geometric averaging. Some of this material first appeared in Bamber et al. (2016).

3.2.1 Analysis

Let F and G be Cumulative Distribution Functions (CDFs) from experts 1 and 2, with densities f, g. Let HW, hw denote respectively the CDF and density of the result of averaging the quantiles of F, G. Then

$$HW^{-1}(r) = \frac{F^{-1}(r) + G^{-1}(r)}{2}, \quad 0 \le r \le 1 \tag{3.1}$$

A good intuitive interpretation (Andrea Bevilacqua, personal communication) notes that *HW* takes the average of the experts' median values, i.e. $r = 0.5$, and a confidence interval whose width is the average of the experts' confidence intervals. The position of the median within the confidence interval depends on the distributions.

To gain further insight into Eq. (3.1), take derivatives of both sides:

$$\frac{1}{hw\left(HW^{-1}(r)\right)} = \frac{\frac{1}{f\left(F^{-1}(r)\right)} + \frac{1}{g\left(G^{-1}(r)\right)}}{2} \tag{3.2}$$

$$hw\left(HW^{-1}(r)\right) = \frac{2}{\frac{1}{f\left(F^{-1}(r)\right)} + \frac{1}{g\left(G^{-1}(r)\right)}} \tag{3.3}$$

Equation (3.3) says that *hw* is the harmonic mean of f and g, evaluated at points corresponding to the *r-th* quantile of each distribution. The harmonic mean of n numbers strongly favors the smallest of these numbers: the harmonic mean of 0.01 and 0.99 is 0.0198. The higher concentration of the *HW* combination would be very valuable IF statistical accuracy were also achieved. Evaluating the statistical accuracy for HW requires real experts assessing real variables from their fields for which true values are known post hoc. None of the proponents of *HW* have verified its performance on real expert data.

3.2.2 Performance on Real Expert Data

Using the thirty-three 2006–2014 professional expert judgment studies, it is possible to compare *HW, EW* and performance weighting (*PW*). In performing this comparison, the global weights combination was used and experts who assessed less than the full set of seed variables were excluded. This causes the *PW* and *EW* solutions used here to differ slightly from the solutions published elsewhere. The integrity of the present comparison is not affected; it was done to facilitate checks by third parties.

The performance of *HW*, *EW* and *PW* are compared with regard to statistical accuracy, informativeness and the combined score (the product of the former two).

Table 3.1 gives the results. *HW* is the best in four of the thirty-three cases, its informativeness is slightly higher than that of *PW*, and substantially higher than *EW*. The statistical accuracy of *HW* is substantially below that of EW and *PW*. In eighteen cases (55%) the hypothesis that *HW* is statistically accurate would be rejected at the 5 percent level. In nine cases rejection would be at the 0.001 level.

Table 3.1 Performance of PW, EW and HW

	PW			EW			HW				
	P-value PW	inf PW	comb	P-value EW	inf EW	comb	P-value HW	inf HW	comb	#seeds	#exprts
Arkansas	0.499	0.337	0.168	0.386	0.198	0.076	5.55E-02	0.640	3.55E-02	10	4
Arsenic D-R	0.036	2.739	0.098	0.061	1.095	0.067	7.99E-04	1.324	1.06E-03	10	9
ATCEP Error	0.683	0.227	0.155	0.124	0.247	0.031	5.99E-04	1.066	6.38E-04	10	5
Biol agents	0.678	0.610	0.414	0.413	0.244	0.101	3.60E-02	0.884	3.18E-02	12	12
CDCROI	0.720	2.305	1.660	0.233	1.230	0.286	7.56E-01	1.565	1.18E+00	10	20
CoveringKids	0.720	0.431	0.310	0.628	0.274	0.172	9.03E-01	0.595	5.38E-01	10	5
CREATE	0.394	0.276	0.109	0.061	0.207	0.013	2.77E-04	0.52	1.44E-04	10	7
CWD	0.493	1.215	0.598	0.474	0.930	0.441	7.07E-01	1.494	1.06E-00	10	14
Daniela	0.554	0.634	0.351	0.533	0.168	0.089	1.82E-01	0.520	9.48E-02	7	4
dcpn_fistula	0.119	1.309	0.156	0.059	0.622	0.037	8.78E-08	1.125	9.88E-08	10	8
eBBP	0.833	1.406	1.172	0.358	0.316	0.113	8.04E-02	0.954	7.67E-02	15	14
EffusiveErupt	0.664	1.123	0.745	0.286	0.796	0.228	2.65E-02	1.505	3.99E-02	8	14
Erie Carps	0.661	0.856	0.566	0.182	0.281	0.051	3.87E-01	0.754	2.92E-01	15	10
FCEP Error	0.664	0.574	0.381	0.222	0.099	0.022	1.75E-05	0.771	1.35E-05	8	5
Florida	0.756	1.133	0.857	0.756	0.455	0.344	6.98E-02	0.880	6.15E-02	10	7
GL-NIS	0.928	0.209	0.194	0.044	0.307	0.014	5.53E-02	0.842	4.66E-02	13	9
Gerstenberge	0.9302	1.095	1.018	0.6439	0.4815	0.31	8.10E-02	0.966	7.82E-02	14	12

(continued)

Table 3.1 (continued)

	PW			EW			HW				
	P-value PW	inf PW	comb	P-value EW	inf EW	comb	P-value HW	inf HW	comb	#seeds	#exprts
Goudheart	0.707	0.959	0.678	0.550	0.277	0.153	6.83E-01	0.888	6.07E-01	10	6
Hemophilia	0.312	0.494	0.154	0.254	0.202	0.051	3.12E-01	0.779	2.43E-01	8	18
IceSheet2012	0.399	1.552	0.620	0.492	0.517	0.254	7.96E-02	1.201	9.56E-02	11	10
Illinois	0.337	0.647	0.218	0.620	0.264	0.163	2.37E-03	0.793	1.88E-03	10	5
Liander	0.228	0.524	0.120	0.228	0.484	0.111	2.81E-03	1.198	3.36E-03	10	11
Nebraska	0.033	1.447	0.048	0.368	0.695	0.256	2.40E-05	1.192	2.86E-05	10	4
Obesity	0.440	0.507	0.223	0.070	0.243	0.017	6.68E-04	0.745	4.98E-04	10	4
PHAC T4	0.178	0.351	0.062	0.298	0.211	0.063	1.64E-02	0.640	0.01048	12	10
San Diego	0.155	0.758	0.117	0.147	1.012	0.148	3.02E-03	1.583	3.32E-02	10	8
Sheep Scab	0.643	1.310	0.843	0.661	0.780	0.516	1.15E-02	1.411	1.63E-02	15	14
SPEED	0.676	0.777	0.525	0.517	0.751	0.389	2.97E-02	1.165	3.46E-02	16	14
TdC	0.989	1.256	1.242	0.166	0.364	0.060	1.24E-02	1.079	1.34E-02	17	18
Tobacco	0.688	1.062	0.730	0.200	0.451	0.090	2.11E-01	0.708	1.49E-01	10	7
Topaz	0.411	1.455	0.598	0.629	0.922	0.580	8.66E-05	1.528	1.32E-04	16	21
umd_nremoval	0.706	1.988	1.404	0.068	0.804	0.054	2.40E-03	1.219	2.93E-03	11	9
Washington	0.200	0.724	0.145	0.155	0.529	0.082	4.21E-01	0.862	3.63E-01	10	5
nr < 0.05	2			1			18				
nr best			26			3			4		
Ave Inf		1.042			0.531			1.077			

"#seeds" denotes the number of calibration variables used in each study, "#experts" denotes the number of experts who assessed all calibration variables in each study

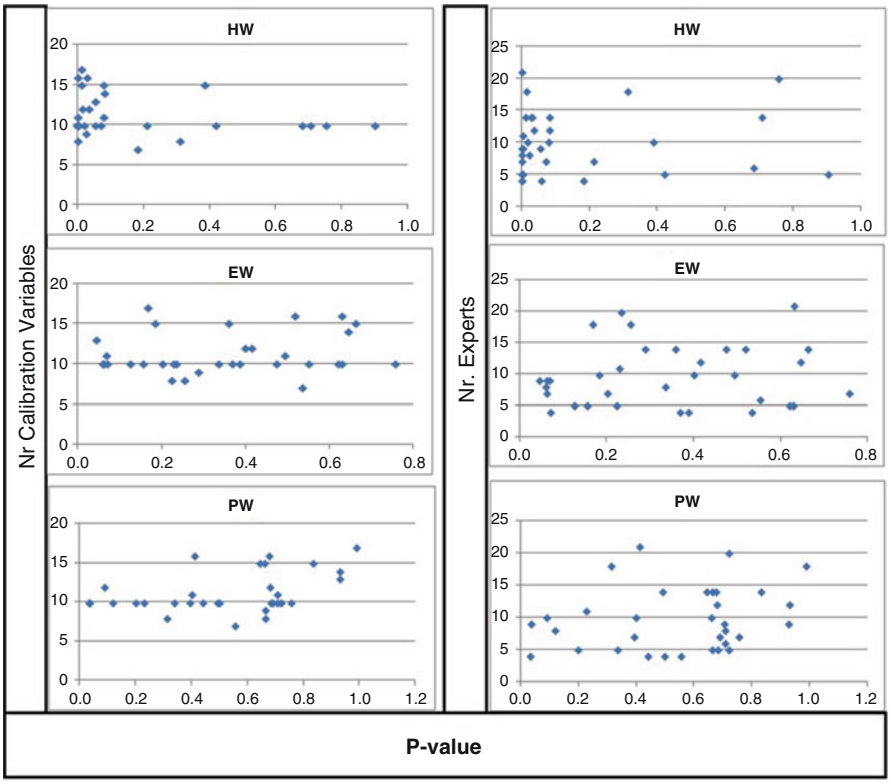


Fig. 3.3 Number of calibration variables and number of experts against P-values for HW, AvP and PW

These data provide evidence on how performance is affected by the number of experts and number of calibration variables. Focusing on statistical accuracy, Fig. 3.3 graphs the number of calibration variables and number of experts against the statistical accuracy scores, for *HW, EW*, and *PW*. *HW* degrades as the number of calibration variables increases, whereas *EW* is unaffected and *PW* actually improves. Indeed, increasing the number of calibration variables increases PW's ability to resolve expert performance. The statistical power of the measure of statistical accuracy increases with the number of calibration variables and this would tend to suppress statistical accuracy scores of all experts and combinations alike. However, no such tendency is observed for *EW* or *PW*. The number of experts does not have a marked effect on any of the combinations.

3.3 Review of Expert Judgment Cross Validation Research

Out of Sample Validation for expert judgment dates from Clemen's (2008) proposal of a Remove-One-At-a-Time (ROAT) method. Calibration variables were removed one at a time and predicted by the model initialized on the remaining calibration variables. The predictions, though originating from different decision makers, were pooled and compared with the equal weight (EW) decision maker (DM). On the fourteen studies selected for his exercise, Clemen found that performance weights (PW) outperformed EW on nine, which was not statistically significant. Cooke (2008, 2012a, b) noted that this ROAT is biased against PW since *each* calibration variable is predicted by a separate DM in which experts who assessed that particular item badly are up-weighted.

3.3.1 ROAT Bias

To understand the ROAT bias, suppose two experts state the probability of heads. Let $P_1(Heads) = 0.8$ and $P_2(Heads) = 0.2$ be the probability of heads for experts *1* and *2*. Suppose that the decision maker's probability is a weighted combination of the experts' probabilities, $P_{dm} = wP_1 + (1-w)P_2$, where the weight of each expert, given observed data, is proportional to the likelihood of each expert's distribution, given the data. Such likelihood weights are not proper scoring rules, and do not account for informativeness; nonetheless there is a strong analogy with the classical model, as the driving term in that model is the likelihood of the hypothesis that an expert is statistically accurate. After observing n Heads and n Tails, the experts' likelihood ratio is

$$\frac{0.8^n \times 0.2^n}{0.2^n \times 0.8^n} = 1 \tag{3.4}$$

so that the weights are each 1/2. If we remove *one* Tail, the likelihood ratio becomes $0.8/0.2 = 4$. We re-initialize our model and predict the Tail which was removed: we find that the predicted probability of Tails is $1-P_{dm}(Heads) = 1-[(4/5) \times 0.8 + (1/5) \times 0.2.] = 1-0.68 = 0.32$. Removing one Tail, strongly tilts the model toward expert 1 with $P(Heads) = 0.8$ and our prediction probability for heads is *0.68*. At the same time we evaluate this model on the Tail which we removed, hence the likelihood for this model on this observation is *0.32*. The same holds, mutatis mutandis, when we remove a Head. If we do this for each of ten coin tosses, the likelihood for our ROAT model is one one-hundredth of the true likelihood ($(0.32/0.5)^{10} = 0.01$).

It is commonly observed that removing one calibration variable can influence an individual expert's statistical likelihood by a factor of three or more, a feature explained by the fact that statistical accuracy is a very fast function. To illustrate, Fig. 3.4 shows the variation in weights of five experts in the EU-USNRC atmospheric dispersion study (Harper et al. 1995) as each of the twenty-three calibration variables (a large number) is removed one-at-a-time.

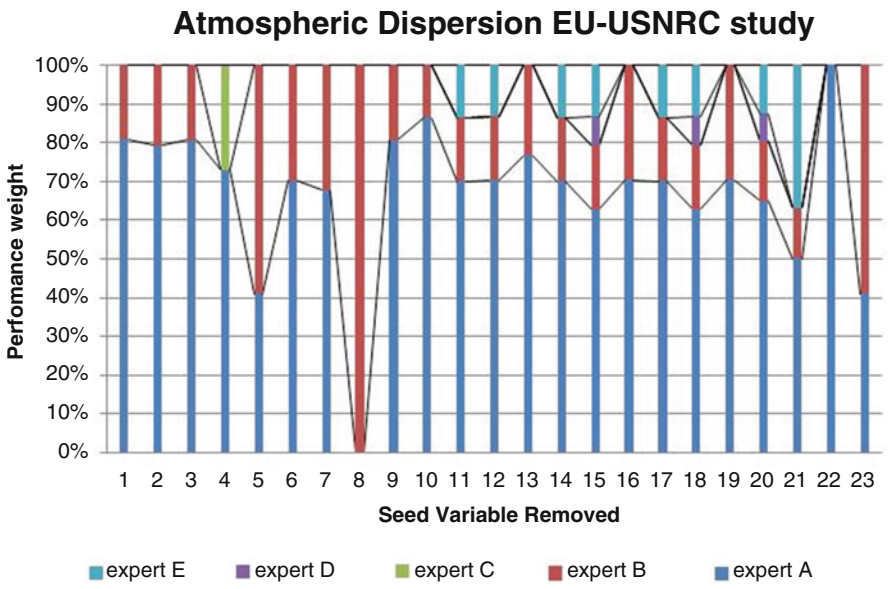


Fig. 3.4 Variation of expert weights under one-at-a-time calibration variable exclusion

Variations on the ROAT approach have been performed by other researchers. Lin and Cheng (2008) examined twenty-eight of the forty-five studies and found PW significantly out performing EW, although PW's out-of-sample performance was degraded. Lin and Cheng (2009) used ROAT on forty studies finding no significant difference between PW and EW. These publications do not report that their code has been vetted against EXCALIBUR, the standard software for implementing the Classical Model (Cooke and Solomatine 1992), and there are very large differences between the values reported in Lin and Cheng (2008) and those reported in Cooke and Goossens (2008), as shown in Table 3.2.

Table 3.2 provides a strong argument for communicating with the authors of the data set before publishing results. Such communication did not happen in this case and the numbers of Lin and Chen bear little resemblance to those of Cooke and Goossens. In six of the twenty-eight studies the names and numbers of experts and calibration variables are so divergent as to prevent determining which study is meant.

The values in Cooke and Goossens (2008) are published values some of which were computed with archaic MS DOS code. That code had a crude method for estimating the tail of a chi square distribution, leading to poor resolution below 1E–4. For large numbers of calibration variables (eg as in study 24) this problem could be acute. It was addressed by reducing the statistical power to a default value of 10. This might explain part of the discrepancy in study 24. For other studies, no explanation suggests itself.

Table 3.2 Comparison of combined scores for performance based and equal weight of Lin and Cheng (2008) and Cooke and Goossens (2008)

					Lin and Cheng (2008), Table 1 "within sample"		Cooke and Goossens (2008) Table 1	
	Lin and Cheng study name	# expert	#calibration vbls	Remarks	PWComb	EWComb	PWComb	EWComb
1	Acrylonitrile	7	10		0.47	0.44	0.764	0.423
2	Option trading	5	34	??				
3	Dike ring	17	47		0.42	0.03	0.2456	0.03768
4	Flanges	10	8		0.6	0.2	0.905	0.4274
5	Crane risk	8	12	11eff	0.93	0.28	1.148	0.345
6	Groundwater	7	10		0.95	0.05	2.106	0.158
7	Space debris	7	26	18eff	6.00E-06	0.13	0.25	0.14
8	Composite materials	6	12		0.55	0.21	0.39	0.111
9	Radiation in food	7	6	??				
10	Dry deposition	8	14		0.48	0.003	0.697	0.001
11	Atmospheric dispersion	8	23		0.38	0.18	0.9785	0.129
12	Early health effects	9	15		0.06	0.01	0.0496	0.01153
13	Radiation dosimetry	5	38					
14	Soil transfer	4	31		1.00E-06	1.00E-07	1.00E-04	9.70E-05
15	Wet deposition	7	19		0.11	0.002	0.113	0.00073
16	Gas pipelines	16	14	??				
17	MONTSE 1	9	8	??				
18	MOTHER5	5	34	??				

19	Montserrat	10	8	??				
20	Movable barriers	8	14		0.06	0.13	0.535	0.125
21	Real estate	5	31		0.7	0.001	0.6296	0.0009
22	River dredging	6	8		0.54	0.18	0.447	0.185
23	Sulpher trioxide	4	7		2.53	0.3	0.547	0.294
24	Building temperature	6	48	10eff	0.002	2.00E-10	0.2005	0.00354
25	Atmospheric dispersion (TNO)	7	36		0.09	0.002	0.604	0.24
26	Radioactive deposition (Delft)	4	24	22eff	0.3	0.22	0.741	0.415
27	Atmospheric disprsion (Delft)	11	36		0.14	0.06	0.562	0.508
28	Water pollution	11	9	11/10eff	0.62	0.4	0.6563	0.4847

The effective number of seed variables is the smallest number assessed by some expert, and is shown in column 5. The statistical accuracy is powered to the effective number of seed variables by the software used by Cooke and Goossens. In some cases (e.g. study 24) scores are powered down because of numerical limitations of archaic code

Several researchers have shared their code with the authors and exact agreement with EXCALIBUR was achieved. Exact agreement was achieved with the out-of-sample code of Eggstaff et al. (2014). The out-of-sample code of Flandoli et al. (2011) has been reviewed and found to optimize incorrectly and to conflate uniform and loguniform background measures. Two of the four cases reported in (Flandoli et al. 2011) had fifteen and sixteen calibration variables, enabling direct comparison with results from the Eggstaff code. The other cases are too large. Flandoli et al. draw 500 random samples from training sets of fixed size and compute the scores on the complementary test set. Table 3.3 compares the results with the complete sample using the Eggstaff code.

It is often suggested that cross validation should use different performance measures than those underlying the Classical Model. Lin and Huang (2012) used ROAT with the Brier score (related to the quadratic scoring rule) in a regression based study of the effects of aggregation method, dependence, number of experts and calibration variables and overconfidence on the Brier score. They follow in the footsteps of Winkler (1969), who first proposed strictly proper scoring rules for individual variables to score experts. A score is assigned to each experts' probability assessment for each calibration variable based on each realization and the scores are summed over the set of calibration variables.

This idea is strongly discouraged in Cooke (1991). A simple example shows why: Suppose an expert assess the probability of Heads for a coin of unknown composition as 1/2. On each toss with the coin, the score is the same for Heads and Tails. If these individual scores are added, then the sum score after 100 tosses is also independent of the actual sequence of outcomes; fifty Heads and fifty Tails gets the same score as 100 Heads. Table 3.4 compares the quadratic score (positively sensed, on $[-1, 1]$) averaged over 1000 predictions of rain of two experts.

Table 3.3 Results of Flandoli et al. (2011) based on 500 samples compared with the vetted code of Eggstaff based on the complete sample where SA is Statistical Accuracy, Inf is Information Score and Comb is the product of both

		PW			EW		
		SA	Inf	Comb	SA	Inf	Comb
Pbearl 7 training, 8 test	Eggstaff	0.149	0.617	0.072	0.271	0.167	0.046
	Flandoli Table 8	0.229	0.407	0.093	0.273	0.167	0.046
Vesuvius 8 training 8 test	Eggstaff	0.277	1.176	0.231	0.520	0.756	0.380
	Flandoli Table 4	0.449	0.896	0.377	0.519	0.720	0.365

Table 3.4 Two experts assessing next day probability of rain on 1000 days, quadratic score positively sensed on $[-1, 1]$

Probability of rain next day		5%	15%	25%	35%	45%	55%	65%	75%	85%	95%	Totals
Expert 1	Assessed	100	100	100	100	100	100	100	100	100	100	1000
	Realized	5	15	25	35	45	55	65	75	85	95	500
Expert 2	Assessed	100	100	100	100	100	100	100	100	100	100	1000
	Realized	0	0	0	0	0	100	100	100	100	100	500

Quadratic score expert 1 = 0.665; Quadratic score expert 2 = 0.835

Both experts are equally informative in the sense that they both attribute 5% probability to one hundred next days, etc. Expert 1 is statistically perfectly accurate, expert 2 is massively inaccurate, yet expert 2 scores better than expert 1. The reason is that such rules decompose as the sum of a "calibration" and "resolution" terms (De Groot and Fienberg 1983, or online appendix). Resolution measures the expert's ability to separate the variables into statistically distinct subsets, regardless whether the distributions assigned to the subsets correspond to the expert's assessments. High resolution overwhelms bad statistical accuracy in the above example.

Other researchers have undertaken cross validation without ROAT. Cooke (2008) looked at half-half splits in thirteen studies with at least fourteen calibration variables. Flandoli et al. (2011) examined five datasets, choosing 30 percent of the number of calibration variables as the size of the test set, provided this number was at least eight, otherwise the test set was eight. They recoded the classical model in R, but did not implement item weights or the log uniform background measure. They randomly drew 500 partitions into training and test sets of the fixed sizes.

The most extensive study of this kind is Eggstaff et al. (2014), which initializes the global weights model on all subsets of calibration variables (except the empty set and the full set) and in each case predicts the complementary subset, again using only global weights. Using primarily the pre 2006 data sets, studies with large numbers of calibration variables were split into separate studies to suppress combinatoric explosion. This resulted in 62 studies for cross validation. We note that the studies are not independent, as the split studies had the same panels of experts. Combined scores for *PW* and *EW* were aggregated per study, and their ratios are shown in Fig. 3.5.

An in-depth analysis of the 2006–2014 breaks the 'performance dividend' of performance based weighting into components and shows the dependence on the number of calibration variables. The number of calibration variables is more uniform (26 of the 33 studies have between 10 and 15), allowing to aggregate training sets based percentage of the calibration set. Aggregating over all training sets whose size is a fixed percentage of all calibration variables, scoring performance on the complementary test sets, and aggregating over all studies, the aggregate statistical accuracy (*Sa*) and informativeness (*Inf*) scores and the combined scores of PW and EW can be plotted as function of percentage training set size. Figure 3.6 left reveals an out-of-sample penalty in statistical accuracy, as the *PW* predicts variables outside the set used to initialize the model. Whereas *EWSa* grows with decreasing test set size (increasing training set size) simply because of loss of statistical power, the same is not true for *PWSa.* For small training sets, *PW* is unable to resolve experts' statistical accuracy, and performance initially lags *EWSa.* The gap starts closing as the training set is 80% of the calibration variables. At this point, the expert weights resemble the weights based on the full set of calibration variables. For Informativeness (Fig. 3.6 right) a different picture emerges. The information advantage of *PW* kicks in for small training sets. Figure 3.7 shows that the effect of the information boost overwhelms the statistical accuracy, and the combined score of *PW* consistently dominates that of *EW.*

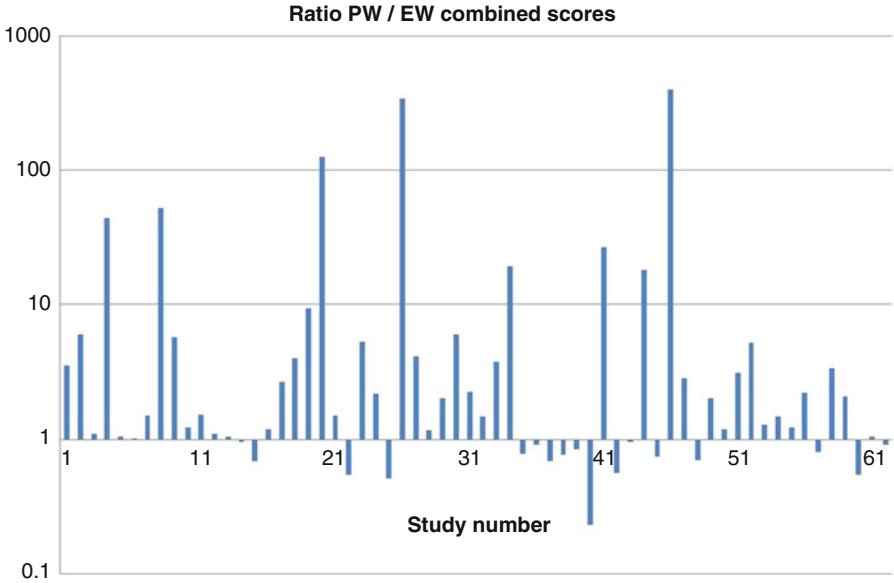


Fig. 3.5 Ratios of combined scores of PW/EW, aggregated per study over all splits into training and test sets of calibration variables. Scores for PW resp. EW were averaged for each training set size, and the averages were geometrically averaged of all sizes

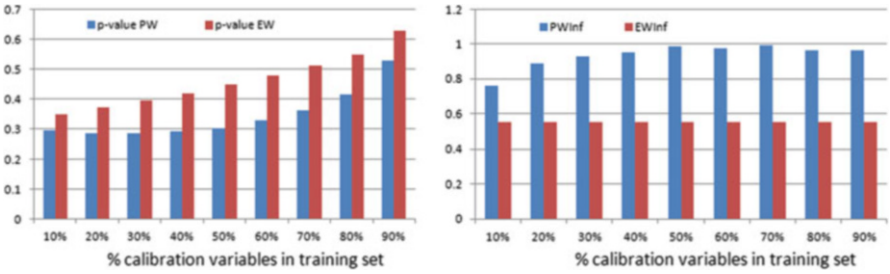


Fig. 3.6 Average over all studies per training set size percentage of the average statistical Accuracy (SA) for PW and EW (*left*) and informativeness (Inf) (*right*)

Before this data was well understood it was opined that a small number of calibration variables might be sufficient. A thorough analysis, taking account of the volatility in weights for small calibration sets (Cooke et al. 2014), re-affirms the advice to use 10 calibration variables. Pooling all vetted cross validation data to date, the hypothesis that PW is no better than EW is rejected at the 2.5E-5 level (Colson and Cooke 2017).

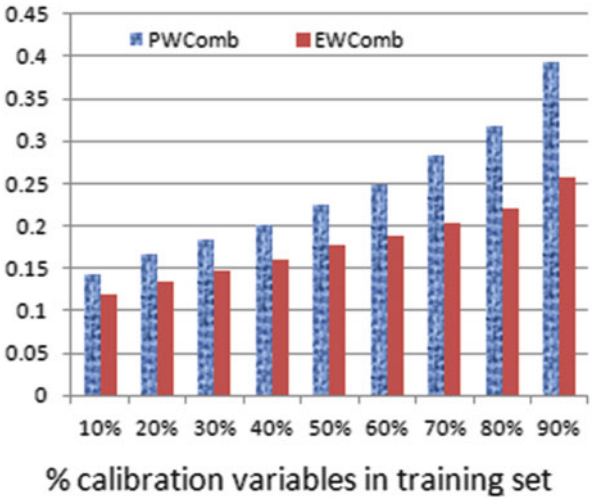


Fig. 3.7 Average over all studies per training set size percentage of the combined score for PW and EW

3.4 Post 2006 Data Sets and Applications Documentation

Expert judgment materials, including data from expert judgment studies, are available at (http://rogermcooke.net/). The studies can be read by the expert judgment software EXCALIBUR free downloadable at http://www.lighttwist.net/wp/. Summary information is presented in the following table. The list includes four studies completed after 2014 beneath the bold line. Reference number refers to the publication number in the Applications Documentation downloadable from (http://rogermcooke.net/rogermcooke_files/Supplementary%20Material%20for%20Cross%20Validation.pdf).

3.5 Conclusion

Social decision making under uncertainty remains the playground of poor ideas where private interests cherry-pick sources, pander fake news and gerrymander proof burdens to promote their agendas. Facing decisions on longer timescales, greater uncertainty and heightened impact, this playground is a luxury society can no longer afford. Developing science based methods of uncertainty quantification for data poor contexts is a priority requiring a host of tools. Recruiting and training experts, communicating to stakeholders, fostering uncertainty awareness in educational programs and the general public—these are among the areas where multidisciplinary approaches are needed. Underpinning all such efforts is validation. Without a clear idea what constitutes good uncertainty assessment and how this can be measured, broader efforts in recruitment, training, communication, education and social acceptance cannot progress beyond the parochial state in which they presently

Contracting party	Performed by	Study name	Subject	Reference nr
Robert Wood Johnson Foundation	Center for Disease Dynamics, Economics & Policy	Arkansas	Grant effectiveness, child health insurance enrollment	9.1
		Florida		
		Illinois		
		Nebraska		
		Washington		
		CoveringKids		
		Tobacco	Grant effectiveness tobacco control	9.2
		Obesity	Grant effectiveness, childhood obesity	
Disease Control Priority Project, 3rd Edition		Fistula	Effectiveness of obstetric fistula repair	3.29
		San_Diego	Effectiveness of surgical procedures	
Center for Disease Control and Prevention		CDC_ROI	Return on investment for CDC warnings	9.3
U Wisconsin, CREATE	V. Bier	Create	Terrorism	
NOAA, EPA	U Notre Dame	Erie_Carp	Establishment of Asian Carp in Lake Erie	3.25, 3.26, 3.27
		GL_NIS	Costs of invasive species in Great Lakes	3.9, 3.22,
EPA, U Maryland	B. Koch. U Maryland	UMD_NREMOVAL	Nitrogen removal in Chesapeake Bay	3.28, 3.31

UMC Utrecht	K. Fischer/M.P. Janssen, UMC Utrecht	Hemophilia	Optimal treatment of patients with severe hemophilia	3.30
National Institute for Public Health and the	TU Delft	ATCEP	Air traffic Controllers Human Error	
		FCEP	Flight Crew Human Error	7.5
Brand Preventie		Daniela	Fire prevention and control	4.11
Liander		Liander	Underground cast iron gas-lines	
U Cambridge	W. Aspinall	Arsenic	Air quality levels for arsenic	–
HPE	Biol_Agent	Human dose-response curves for bioterror agents	–	
U Ottawa	CWD	Infection transmission risks: Chronic Wasting Disease from deer to humans	3.19; 3.20; 3.21	
PrioNet	eBPP	XMRV blood/tissue infection transmission risks	3.18; 3.24	
UK Government	Eff Erupt	Icelandic fissure eruptions: source characterization	6.12	
U Bristol/BAS/Ice2sea.eu	IceSheets	Contribution to sea level rise from Ice Sheets melting due to global warming	7.1	
uOttawa	PHAC_T4	Additional CWD factors	3.21	

(continued)

Contracting party	Performed by	Study name	Subject	Reference nr
U Bristol/BRISK		Sheep	Risk management policy for sheep scab control	–
INGV		SPEED	Volcano hazards (Vesuvius & Campi Flegrei, Italy)	6.18, 6.19
NERC/ESRC/BGS		TdC	Volcano hazards (Tristan da Cunha)	6.17
NUMO/Obayashi Corp		TOPAZ	Tectonic hazards for radwaste siting in Japan	6.16
Natural Hazards Research Platform/GNS Science	M. Gerstenberger	Gerstenberger	Canterbury Seismic Hazard Model	7.2,7.3, 7.4
Embry-Riddle	Goodheart	Goodheart	Airport safety	
World Health Organization	RFF/USDA	Cooke Hoffmann Aspinall	Global burden of foodborne disease, 134 panels, 74 experts,	3.29, 3.33, 3.34
Ariel Re, AIG Life	Ismail & Ried	Insurance	Insurance risk	8.4
Resources for the Future	Colson & Cooke	BF&IQ	Breastfeeding and IQ	3.34
USGS	Aspinall	USGS	Volcanos	in prep

find themselves. The mathematical tools for validation presented here are certainly not the last word, but hopefully signal directions along which social decision making under uncertainty can advance.

References

Aspinall WP, Loughlin SC, Michael FV, Miller AD, Norton GE, Rowley KC, Sparks RSJ, Young SR (2002) The montserrat volcano observatory: its evolution, organisation, role and activities. In: Druitt TH, Kokelaar BP (eds) The eruption of Soufrière Hills Volcano, Montserrat, from 1995 to 1999. Geological Society, London, pp 71–92

Aspinall W (2010) A route to more tractable expert advice. Nature 463:294–295

Aspinall WP, Cooke RM, Havelaar AH, Hoffmann S, Hald T (2015) Evaluation of a performance-based expert elicitation: WHO global attribution of foodborne diseases. PLoS One 11(3):e0149817. doi:10.1371/journal.pone.0149817. eCollection 2016

Bamber JL, Aspinall WJ, Cooke RM (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. Clim Chang 137(3–4):321–328. doi:10.1007/s10584-016-1672-7

Clemen RT (2008) Comment on Cooke's classical method. Reliab Eng Syst Saf 93(5):760–765

Colson A, Cooke RM (2017) Cross validation for the classical model of structured expert judgment. Reliab Eng Syst Saf 173:109–120

Cooke RM (2017) Strictly proper scoring rules as weights. In: Dias LC, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgment. Springer, New York. 2018 (on-line Appendix of this book)

Cooke RM (1991) Experts in uncertainty, opinion and subjective probability in science. Oxford University Press, Oxford, p 321

Cooke RM, Solomatine D (1992) EXCALIBUR—integrated system for processing expert judgments, user's manual version 3.0. Delft University of Technology and SoLogic Delft, Delft

Cooke RM (2008) Response to Comments, Special issue on expert judgment. Reliab Eng Syst Saf 93:775–777. Available online 12 March 2007. Volume 93, Issue 5, May 2008

Cooke RM, Goossens LHJ (2008) TU Delft expert judgment data base, special issue on expert judgment. Reliab Eng Syst Saf 93:657–674. Available online 12 March 2007, Issue 5, May 2008

Cooke RM (2012a) Pitfalls of ROAT cross validation comment on effects of overconfidence and dependence on aggregated probability judgments. J Model Manag 7(1):20–22. issn: 1746-5664

Cooke RM (2012b) Uncertainty analysis comes to integrated assessment models for climate change . . . and conversely. Clim Chang 117(3):467–479. doi:10.1007/s10584-012-0634-y

Cooke RM, Wittmann ME, Lodge DM, Rothlisberger JD, Rutherford ES, Zhang H, Mason DM (2014) Out-of-sample validation for structured expert judgment of asian carp establishment in Lake Erie. Integr Environ Assess Manag 10(4):522–528. doi:10.1002/ieam.1559

Cooke RM (2015) Messaging climate change uncertainty with supplementary online material. Nat Clim Chang 5:8–10. doi:10.1038/nclimate2466

Cooke RM, Mendel M, Thijs W (1988) Calibration and information in expert resolution. Automatica 24(1):87–94

De Groot M, Fienberg S (1983) Effective scoring rules for probabilities forecasts. Manag Sci 29(4):447–454

Eggstaff JW, Mazzuchi TA, Sarkani S (2014) The effect of the number of seed variables on the performance of Cooke's classical model. Reliab Eng Syst Saf 121:72–82. doi:10.1016/j.ress.2013.07.015

Flandoli F, Giorgi E, Aspinall WP, Neri A (2011) Comparison of a expert elicitation model with the Classical Model, equal weights and single experts, using a cross-validation technique. Reliab Eng Syst Saf 96:1292–1310. doi:10.1016/j.ress.2011.05.012

Genest C, Zidek J (1986) Combining probability distributions: a critique and an annotated bibliography. Stat Sci 1(1):114–148

Gillingham K, Nordhaus WD, Anthoff D, Blanford G, Bosetti V, Christensen P, McJeon H, Reilly J, Sztorc P (2015) Modeling uncertainty in climate change: a multi-model comparison. In Working paper 21,637 NBER working paper series http://www.nber.org/papers/w21637. National Bureau of Economic Research

Hald T, Aspinall W, Devleesschauwer B, Cooke RM, Corrigan T, Havelaar AH, Gibb H, Torgerson P, Kirk M, Angulo F, Lake R, Speybroeck N, 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 11:e0145839. doi:10.1371/journal.pone.0145839

Harper FT, Hora SC, Young ML, Miller LA, Lui CH, McKay MD, Helton JC, Goossens LHJ, Cooke RM, Pasler-Sauer J, Kraan B, Jones JA (1995) Probabilistic accident consequence uncertainty study: Dispersion and deposition uncertainty assessment. Prepared for U.S. Nuclear Regulatory Commission and Commission of European Communities NUREG/CR-6244, EUR 15855 EN, SAND94–1453, Washington/USA, and Brussels-Luxembourg, November 1994, published January 1995. Volume I: main report, volume II: appendices A and B, volume III: appendices C, D, E, F, G, H

Hoffmann S, Aspinall W, Cooke RM, Cawthorne A, Corrigan T, Havelaar AH, Gibb H, Torgerson P, Kirk M, Angulo FJ, Lake R, Speyboeck N, Devleesschauwer B, Hald T, World Health Organization, Foodborne Epidemiology Reference Group, Source Attribution Task Force (2016) Research synthesis methods in an age of globalized risks: lessons from the global burden of foodborne disease expert elicitation. Risk Anal 36(2):191–202

Laddaga R (1977) Lehrer and the consensus proposal. Synthese 36:473–477

Lichtendahl KC Jr, Grushka-Cockayne Y, Winkler RL (2013) Is it better to average probabilities or quantiles? Manag Sci 59(7):1594–1611. issn: 0025-1909. doi:10.1287/mnsc.1120.1667

Lin S-W, Cheng C-H (2008) Can Cooke's model sift out better experts and produce well-calibrated aggregated probabilities? Department of Business Administration, Yuan Ze University, Chung-Li, Taiwan Proceedings of the 2008 IEEE, IEEM

Lin S-W, Cheng C-H (2009) The reliability of aggregated probability judgments obtained through Cooke's classical model. J Model Manag 4(2):149–161

Lin S-W, Huang S-W (2012) Effects of overconfidence and dependence on aggregated probability judgments. J Model Manag 7(1):6–22

NAS (2017) Valuing climate damages: updating estimation of the social cost of carbon dioxide. The National Academies Press, Washington, DC. doi:10.17226/24651

Oppenheimer M, Little CM, Cooke RM (2016) Expert judgment and uncertainty quantification for climate change. Nat Clim Chang 6:445–451. doi:10.1038/NCLIMATE2959

Quigley J, Colson A, Aspinall W, Cooke RM (2017) Elicitation in the classical method. In: Dias LC, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgment. Springer, New York. 2018 (Chapter 2 in this book)

Rasmussen NC, et al (1975) Reactor safety study. An assessment of accident risks in U. S. commercial nuclear power plants. WASH-1400 (NUREG-75/014). Rockville, MD, USA: Federal Government of the United States, U.S. Nuclear Regulatory Commission

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

Wadge G, Aspinall WP (2014) A review of volcanic hazard and risk assessments at the Soufrière Hills Volcano, Montserrat from 1997 to 2011, Ch. 24. In: Wadge G, Robertson REA, Voight B (eds) The eruption of Soufrière Hills Volcano, Montserrat, from 2000 to 2010: Geological Society Memoirs, vol 39. Geological Society, London, pp 439–456

Winkler RL (1969) Scoring rules and the evaluation of probability assessors. J Am Stat Assoc 64:1073–1078

Wittmann ME, Cooke RM, Rothlisberger JD, Rutherford ES, Zhang H, Mason D, Lodge DM (2014) Structured expert judgment to forecast species invasions: bighead and silver carp in Lake Erie. Conserv Biol 29(1):187–197. doi:10.1111/cobi.12369

Chapter 4
SHELF: The Sheffield Elicitation Framework

John Paul Gosling

Abstract The Sheffield elicitation framework is an expert knowledge elicitation framework that has been devised over a number of years and many substantial expert knowledge elicitation exercises to give a transparent and reliable way of collecting expert opinions. The framework is based on the principles of behavioural aggregation where a facilitator-guided group interact and share information to arrive at a consensus. It was originally designed for helping to elicit judgements about single uncertain variables, but, in recent years, the framework and the associated software implementations have been extended to accommodate judgements about more complex multidimensional variables and geographically-dispersed experts. In this chapter, we discuss the aims and foundations of the framework, its extensions and its notable applications.

4.1 Introduction

Meticulous preparation is required alongside consideration of potential psychological pitfalls to ensure representative judgements are captured from experts in any elicitation exercise whose results are to be used in a decision making process. The process of capturing expert judgements involves the investment of many hours of effort on the part of people involved. As such, it is important that the results are transparent and defensible: any potential user will need to understand the basis on which the judgements have been made and trust the process. Key principles for successful elicitation exercises include well-structured questions, unambiguous definitions of quantities, transparency in the process and opportunities for experts to share their expertise and reasoning (Garthwaite et al. 2005; Morgan and Henrion 1990; O'Hagan et al. 2006).

The Sheffield elicitation framework (henceforth SHELF) is an expert knowledge elicitation (EKE) protocol that provides a transparent and rigorous approach to capturing judgements from multiple experts. The synthesis of the experts'

J.P. Gosling (✉)
School of Mathematics, University of Leeds, Leeds LS2 9JT, UK
e-mail: j.p.gosling@leeds.ac.uk

L.C. Dias et al. (eds.), *Elicitation*, International Series in Operations Research & Management Science 261, https://doi.org/10.1007/978-3-319-65052-4_4

judgements is achieved through facilitated group discussion aiming to arrive at a consensus distribution using behavioural aggregation. SHELF provides a framework for capturing information about an elicitation exercise including the experts' backgrounds and potential conflicts of interest and any reasoning or key sources of information that underpin the experts' judgements. This is done through a comprehensive set of questions that are designed to cover everything a user of the elicitation exercise results needs to know before applying them to their particular decision problem. On the more technical side, SHELF was originally set up to cover a variety of univariate elicitation techniques including the roulette, bisection and range methods and uses least squares fitting to model the experts' probability distributions (which are described fully in the next section). The method has been subsequently extended to cover judgements about multidimensional parameters (including vectors of proportions).

An EKE exercise carried out using SHELF is designed to be performed by a group of experts guided by a facilitator. The individual experts are asked to make their own quantitative judgements after the quantity of interest has been discussed and then a group consensus distribution is suggested to the facilitator using linear opinion pooling with equal weights. Both the individual fitted distributions and the potential consensus distribution can be displayed to the group and discussions are encouraged as to whether the consensus distribution is valid. This final step is a behavioural approach to aggregation that can use mathematical aggregation as a point of departure. Throughout, the experts have opportunities to contribute to discussions and to revise their judgements, and a skilled facilitator is needed to tackle the difficulties of managing group dynamics.

The application of SHELF is supported by explanatory documentation, forms to capture the various stages of the process and a set of R functions (Core Team 2016; Oakley and O'Hagan 2014). The documentation gives the justification for each of the steps in the protocol along with suggestions for facilitators to help avoid potential biases in the experts' judgements. The forms are a tool for capturing information from details of the experts' and the aims of the EKE exercise through to the ultimate consensus distribution. The R functions allow the facilitator to fit distributions and demonstrate the consequences of judgements to the experts during the elicitation sessions.

The development of SHELF stems from years of experience in designing and performing EKE exercises over many different application areas. A particularly important project in the development was the Department of Health funded project, "Bayesian Elicitation of Expert Probabilities", which brought together a number of senior researchers on the topic of expert judgements and culminated in a book reviewing the topic (O'Hagan et al. 2006). This project was also supported by ongoing research in the Centre for Bayesian Statistics in Health Economics at the University of Sheffield where methods were being developed to help support decision making in health-related EKE. The final catalyst for the first version of SHELF stemmed from a project investigating microbial risk assessments where several model parameters needed to be specified using expert judgement and software was needed to help the facilitation of the EKE (Kennedy et al. 2009).

Since its inception, SHELF has been widely applied in health economics and medicine, which is unsurprising given its background. There have also been notable applications in business planning, natural hazards and environmental sciences amongst others. These are highlighted in Sect. 4.3.

There has been some use of SHELF at a government and regulator level. The Aqua Book is produced by the UK's Treasury and provides guidance on producing quality analysis for government (Treasury 2015). In the Aqua Book, the Sheffield method is highlighted as a formal method for eliciting expert knowledge. The UK's Defence Science and Technology Laboratory describe many features of SHELF for performing probabilistic elicitation of subjective data (Defence Science and Technology Laboratory 2015). The European Food Safety Authority has also recommended the use of the SHELF as one of its methods for conducting structured EKE (European Food Safety Authority 2014).

In the next section, we outline the steps in SHELF with reference to the challenges in quantifying opinions that the framework aims to solve. In Sect. 4.3, applications of SHELF are highlighted across areas such as health-related research and environmental science. Extensions to the framework that have been implemented are described in Sect. 4.4, and, in Sect. 4.5, we discuss other potential extensions to the framework alongside the benefits and challenges of applying SHELF.

4.2 The Elicitation Framework

There are several key participants in an EKE exercise. The problem owner is the person who wishes to quantify the current knowledge about the quantity of interest. The experts are the people who the problem owner believes are likely to have useful knowledge about the quantity of interest. The facilitator of the EKE exercise interacts with the experts to obtain the desired information about the quantities of interest on behalf of the problem owner. The facilitator will typically have knowledge of statistics and probabilistic reasoning and be skilled in managing meetings. In practice, many EKE exercises will benefit from having multiple facilitators: in the past, we have found it useful to have someone running the meeting and another recording the discussions and performing any necessary calculations.

The most important part of SHELF is the set of forms that guide the facilitator of an elicitation exercise through the necessary steps whilst enabling them to record what has been presented to the experts and what their responses have been. In order for a decision maker to make use of the results of an elicitation exercise, they must have faith in the results. This faith can be gained by having transparency in the process that comes from producing complete records of an exercise and recording the evidence base behind the experts' reasoning. There is the additional bonus of having such records that, if the exercise needs to be revisited, then the

comprehensive records are available to set up a new elicitation session and to find out what evidence the original judgements were based upon.

Of course, an elicitation exercise is not just about asking an expert a number of questions about the quantities of interest and fitting distributions. There are several stages the facilitator of the elicitation exercise should go through (Garthwaite et al. 2005):

- problem set-up and training of experts,
- eliciting beliefs about the quantities of interest,
- fitting of an appropriate distribution,
- feedback of implications of fitted distribution,
- revision of judgements to reach a consensus distribution.

The final three stages should be repeated until the experts are happy that the fitted distribution reflects their beliefs about the quantity of interest. The ultimate aim of expert elicitation is to finish with a probability distribution that the experts are satisfied captures their beliefs. Figure 4.1 is a flowchart outlining the steps behind these stages when employing SHELF for a single quantity of interest.

In the flowchart, we can see that there are many stages to complete before the experts are asked to make any direct judgements about the quantities of interest and, once they have made judgements, there are opportunities for them to discuss and revise their judgements as the group moves towards a consensus. Formally, SHELF provides forms with detailed instructions for capturing stages (3)–(8); however, guidance has been provided within the documentation on stages (1) and (2) alongside some briefing documents that can help with the expert selection phase.

Stages (3)–(8) take place during a facilitated workshop. In our experience, the facilitator will benefit from having the following available during the workshop (European Food Safety Authority 2014):

- a computer linked to a projector to step through the SHELF documents and provide feedback to the experts,
- a flip chart or white board to allow key information about the quantities of interest can be displayed to the experts throughout,
- name cards for each person in the room to aid in the capturing of judgements and group cohesion,
- writing materials for each expert.

For the latter item, it is beneficial to have preprinted forms for the expert to use to capture their quantitative judgements. An examples of this will be given in Sect. 4.2.5.

4.2.1 Exercise Specification

Prior to commissioning an EKE, the problem owner must have identified a need for the quantification of uncertainty about some quantities of interest due to incomplete or inconsistent information. It is important that the problem owner specifies the

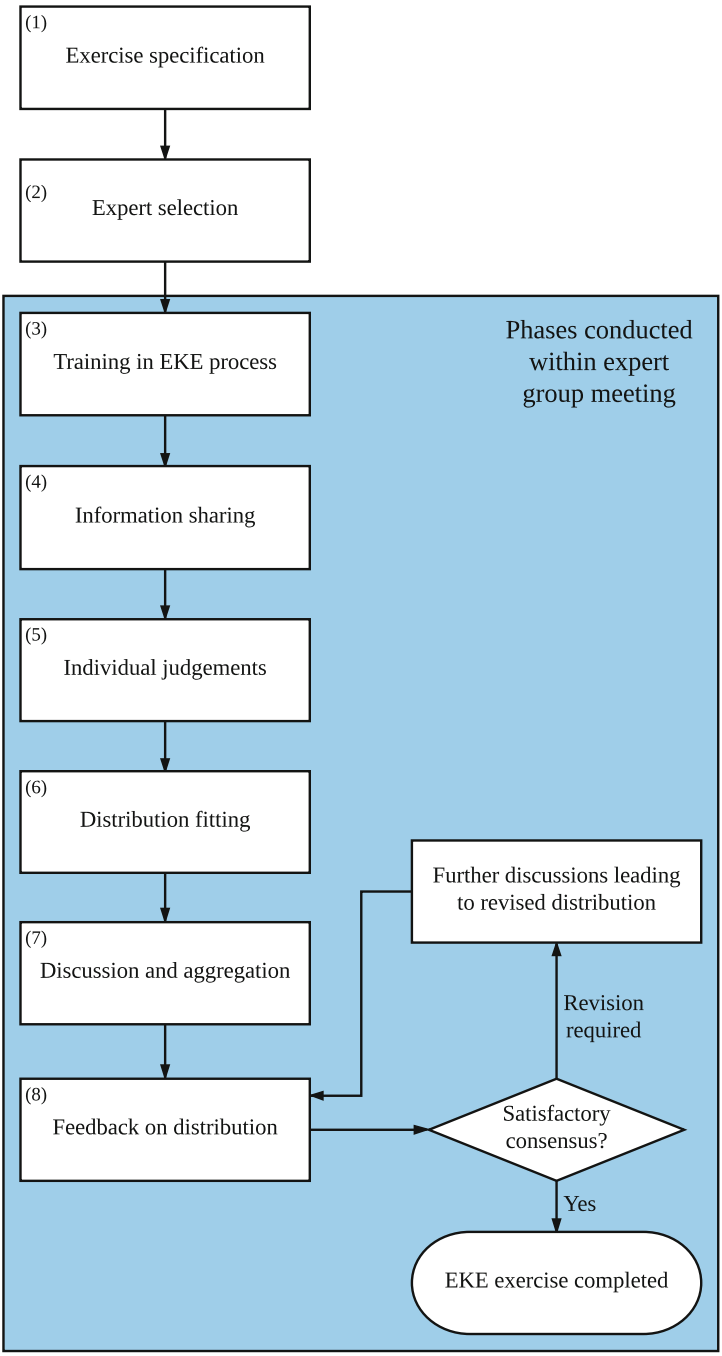


Fig. 4.1 The flow of an elicitation exercise under the SHELF protocol

quantities of interest unambiguously because ambiguity can render an EKE useless. Unclear definitions with potentially contentious wording can lead to unnecessary arguments between participating experts and may lead to misinterpretation from the ultimate users of the exercise results. It can also be beneficial for the problem owner to collect the relevant information and evidence on the quantities of interest so that it can be shared with the experts before the EKE (see Sect. 4.2.2).

It is recommended that the problem owner considers the potential impact of not assessing the uncertainty on their decision making (Morgan and Henrion 1990; O'Hagan et al. 2006). If the quantities have little bearing on the potential decisions, then the great effort needed for a successful EKE may be better spent elsewhere. Further guidance on the exercise specification and the role of the problem owners in this is discussed in European Food Safety Authority (2014).

4.2.2 Expert Selection

It is important to elicit beliefs from a group of experts, rather than a single expert, in order to synthesise the range of knowledge and opinions of the relevant expert community. Kadane (1986) recommends that prior distributions are representative of the community of experts; this may lead to any number of experts' beliefs being combined to form one prior distribution. Of course, the experts must be able to communicate information about their knowledge, but that does not necessarily mean that they will need to be able to make probabilistic statements at the outset. Within SHELF, there is ample opportunity to train and aid the experts with making such judgements (see Sect. 4.2.3).

SHELF can be used with a single expert up to any number of experts (in theory). However, experience in conducting EKEs under SHELF has led to the belief that having five-to-ten experts seems to be practicable (European Food Safety Authority 2014): the group has the potential to cover many perspectives and their are enough people for the experts to feel comfortable making judgements without feeling as if they are being interrogated, but there are not so many that the group is difficult to facilitate. In cases where there are many quantities to elicit information on, the problem owner may allow experts to enter the discussions temporarily to act in an advisory role. This has been suggested in European Food Safety Authority (2014), but the facilitator must manage such a situation carefully because the advisor might want to be more involved in the process or they may disrupt the group. In situations where there may be several quantities to elicit judgements on, there may only be one expert available on each of the topics. During the elicitation exercise, we typically see the relevant individual dominating discussions surrounding their expertise and often the others are happy to defer to their judgements.

Another consideration when selecting experts is length of time that an EKE will take. It is recommended that a SHELF exercise takes between one and two days. Experience of applying the method for multiple quantities tells us that, as the expert group become familiar with the process, this can drop to just 2–3 h (Gosling

et al. 2012) (dependent on the shared relevant information between the quantities of interest and the level of expert training required).

Recruitment is limited by the availability of experts. An example recruitment strategy of experts arose in the context of quantifying patient survival (Girling et al. 2007). The experts were recruited because they were in the same place at the same time: the 51st annual conference of the American Society for Artificial Organs. Of course, conferences can provide a good opportunity to get experts together, but there is a danger that the selected experts might be a biased sample because they might all be from a particular academic field and not cover all backgrounds of interest. When deciding the make up of the group, the problem owner must also consider the effects of perceived seniority on group dynamics. Different people respond differently in group situations to the presence of people they consider to be more senior (in terms of experience or institutional hierarchy). Part of the facilitator's task is to make every expert realise that their opinions and judgements are valued (see Sect. 4.2.4); this can be aided by avoiding expert groups that have widely varying levels of seniority. A much more comprehensive discussion of these recruitment issues can be found in Chap. 16 of this book (see Bolger 2018).

As part of the recruitment process when using SHELF, a briefing document is sent to the identified experts outlining the purpose of the EKE and stating the use of SHELF. The suggested text from Oakley and O'Hagan (2014) is as follows:

> "The purpose of the elicitation meeting is to obtain probability distributions to represent your uncertainty about various quantities of interest. The elicitation will be conducted following the Sheffield Elicitation Framework (SHELF), based on elicitation practice recommended in O'Hagan et al. (2006). You will be given training in the process of elicitation at the start of the meeting, which will include a practice exercise to familiarise you with the procedure.
>
> "It is important to note that you will not be asked to provide single estimates of any of these quantities. The elicitation process will instead involve considerations such as what a plausible range of values would be for each unknown quantity, and whether, in your opinion, some values are more likely than others. You may have considerable uncertainty about some of these quantities (though less than that of a lay person). This will not be of concern during the elicitation itself, as the outputs from the elicitation will reflect large uncertainty when it is present.
>
> "Due to the subjective nature of elicited probability distributions, it is important to make the elicitation process as transparent as possible. A written record will be kept of the meeting, which will include details of experts present at the meeting, a summary of each expert's relevant expertise, and any declarations of interest."

The final sentence states that a list of the experts will be recorded for transparency's sake, but the experts are offered partial anonymity in that the individual comments and judgements that will appear in the SHELF reports will not be attributed to any individual. Full anonymity can also be offered if appropriate, but, when not absolutely necessary, it is more useful to maintain the partial anonymity to help the group have ownership for the their judgements. In practice, this means that a full list of experts and their affiliations will be added to the EKE exercise record, but no judgement or line of reasoning will be attributed to any single expert. Whatever the method is used for recruitment, there must be transparency in the process: it is important that subsequent users of the elicitation exercise know whose experience the judgements are based on and why.

Alongside the text, the experts are provided with details of the background to the problem, clear definitions of the quantities of interest and any evidence that has been identified by the problem owner as particularly relevant. The experts are asked to complete a form given this information asking for the following:

- declarations of interests in the outcomes of the EKE,
- expertise of the respondent relevant to the quantities of interest,
- additional evidence that is relevant to the quantities of interest.

Although experts will often be stakeholders in the ultimate decision problem, the declarations of interests enable the expert group to recognise the potential vested interests within the group so that they can discuss the quantities openly. They also help the facilitator to be aware of possible tensions. It should be noted that the experts may be employees who will benefit from success in the enterprise to which the elicitation contributes or they may be invited specifically to represent a stakeholder group or point of view.

Although training is part of the formal SHELF exercise (see Sect. 4.2.3), it is often beneficial to give the experts some background material (or even online training) on expressing uncertainty through probabilities. It should be noted that some experts may already be comfortable with probabilistic reasoning and some experts may need more training in order to understand this use of probability. Of course, sending a briefing document explaining the objectives of an EKE exercise and getting a group of experts into a room do not guarantee that the experts will feel part of the process or that the group will function in a way conducive to providing useful information. At the expert meeting, SHELF begins with a statement by the facilitator reiterating the briefing document and assuring the experts that all of their opinions and judgements on the topic are valuable to the problem owner including a statement:

> "Participants are aware that this elicitation will be conducted using the Sheffield Elicitation Framework, and that this document, including attachments, will form a record of the session."

4.2.3 Training in EKE Process

Having expertise in a particular field does not guarantee an ability to make probabilistic judgements about quantities of interest in that particular field. In EKE exercises, we often have experts who have little or no knowledge of probability or statistics. It is therefore important that the facilitator guides the expert through the process based on a toy problem. There are two approaches to this: the first is to give an almanac-type problem where the subject is reasonably well known and the second is to tailor the problem so that it is similar in nature to the ultimate quantities of interest. It is common for researchers to use road distances between cities, timings of train journeys or heights of mountains in the first approach (European Food Safety Authority 2014; O'Hagan 1998). It can be useful when using such quantities if the

true value is not so simple to look up for the experts (although this is difficult given access to the internet). The exercise may be more beneficial if there will be a range of experiences and uncertainties across the experts. Asking about the population of a city can work with regards to this because experts local to that city may expect to have more knowledge about the population.

Training using quantities that are similar to the ultimate quantities of interest may give the experts an experience that is closer to what is coming later in the elicitation exercise, but it also may cause other problems. It can be difficult to find similar enough quantities to the ones of ultimate interest such that the values are not known by the experts. The major problem is that having a relevant quantity in the training phase may mean that the experts become too involved in that they spend too much time discussing the quantity.

In the training stage, the facilitator should step the experts through stages (5)–(9) of the protocol as shown in Fig. 4.1. The training phase should be as close to the real elicitation as possible in the use of the SHELF forms and the sharing of the experts' individual fitted distributions (see Sect. 4.2.5). The facilitator should take care to explain the probabilistic judgements that are required and point out any incoherences in the experts' judgements. This is also an opportunity for the facilitator to help work on the well-known issue of experts being overconfident in their judgements (Kadane and Wolfson 1998). First, the facilitators can identify and question experts who have relatively narrow uncertainty ranges, and, secondly, the facilitators can reveal the true value of the training quantity of interest to help shift the experts' opinions.

In some expert elicitation exercises, there are experts who do not want to engage fully in the process and may feel unable to give quantitative judgements (Morgan and Henrion 1990). The training stage may help the facilitators to identify these experts; however, the fabricated nature of the training exercise may not be a true reflection of how an expert will operate when considering the real quantities of interest.

4.2.4 Information Sharing

Before quantitative judgements are made about the quantities of interest, information about the experts and their expertise is needed. This serves three purposes: first, the problem owner needs to know who is in the group and what expertise is covered so they can have faith in the results from the EKE exercise; secondly, this stage helps the experts to focus on the problem in hand and to remember relevant information sources; and, thirdly, it helps to improve the confidence that the experts have in each other.

The formal SHELF questioning begins with
"Have you got any interests that are related to the variable under consideration?", which is asking for a declaration of interests. Although these will have been captured to some extent if a pre-elicitation questionnaire has been administered, recognising

their own potential vested interests before the group and those of other participants helps the experts to report their beliefs openly and in an informed way. It is also important for the decision maker to be aware of possible biases.

The second question in this stage is:

"What is your expertise in relation to the quantity under consideration?"

The purpose of this question is self explanatory. As a decision maker, we would want to know about the experts' experience in areas related to the quantities. At this stage, there should also be the opportunity for the experts in the group to suggest what expertise might be missing in the room. It is useful for the problem owner and future users of the results to know what the perceived weaknesses of the group were, and it may help direct further study or future EKE exercises.

After the initial sharing of information about expertise and the training exercise, the experts are then asked to refocus on the quantities of interest. The first question on the SHELF forms for this stage is:

"What facts are important when making judgements about the variable under consideration?"

This leads to a list of influencing factors that the experts need to keep in mind when making the judgement. We have found it useful to record these on a flip chart or white board so that all the experts have access to these key points whilst making their judgements. The facilitators can also use this question to check for ambiguities in the definition of the variable (for example, has the scenario that is been conditioned on been defined clearly enough?).

Once the list has been produced, the experts are asked to relate these to evidence:

"What quantitative or qualitative evidence have you seen relating to the variable under consideration?"

This could be a list of key publications and reports, which could build upon the shared materials before the elicitation meeting. The evidence could also be experiences that the experts have had in their careers and research. Apart from recording the evidence base that the experts were using, there is an additional benefit similar to the benefit of the earlier question on expertise: the experts are once again reminded of their expertise and value. During these discussions, the experts may also have comments on the dependency between the quantities of interest. The SHELF forms have a section entitled "Structuring" with instructions:

"Record any choices made to structure the quantities of interest in terms of others that may be easier to elicit."

Structuring or elaboration is an important tool when eliciting information about several quantities. It is generally better to structure the quantities of interest in terms of other quantities that the experts judge to be independent of each other and in terms of quantities the experts feel more comfortable making judgements about O'Hagan (1988).

Although these discussions are important to have transparency and a functioning, engaged group, the facilitator must act to curtail peripheral discussions due to the inevitable time constraints.

4.2.5 Individual Judgements

The SHELF documentation supports the recording of many types of quantitative expert judgements including tertiles, quartiles and direct assessments of the cumulative distribution. In this stage, the experts will discuss the evidence related to the quantities, but they are asked to keep their quantitative judgements to themselves. Here we list the questions that are suggested in the documentation for both the quartile and roulette methods (which have been the most popular methods within the context of SHELF, see Sect. 4.3).

The quartile method (which is referred to as the "Sheffield method" in European Food Safety Authority 2014) is based upon the bisection method of Raiffa (1968). Before asking for quantitative judgements, if there any many quantities under consideration, it is worth reiterating the precise definition of the quantity of interest and the key evidence (ideally, these will still be displayed on a flip chart or white board). The facilitator starts by asking for bounds for the quantity of interest; these may be physical bounds or provide a range for which it is extremely unlikely that the true value will be outside. As part of this, the facilitator should prompt the experts to think about circumstances that can rise to values that are past the lower and upper bounds to test if the experts are missing plausible values. The quantitative questioning begins with the extreme ends to help with overcoming two well known biases in making probabilistic judgements: anchoring and overconfidence (Kadane and Wolfson 1998). We do not start with a best guess for the quantity of interest because experts have a tendency to anchor on this value and adjust their uncertainty judgements away from this value. Also, thinking about extremes can make values away from the best guess feel more plausible to the experts and this can counteract the natural tendency to be overconfident in predictions.

The next question aims to elicit the experts' medians:

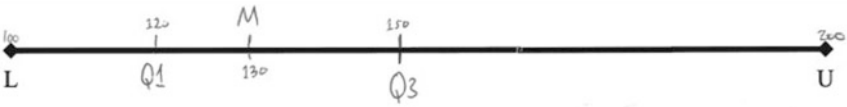
"Can you specify a value such that it is equally likely that the true value lies below or above it?"

After recording their own judgements, the expert are then asked about the remaining quartiles:

Lower quartile: "Suppose the true value is definitely below the median you have specified. Can you specify a value such that it is equally likely that the true value lies below or above it?"

Upper quartile: "Suppose the true value is definitely above the median you have specified. Can you specify a value such that it is equally likely that the true value lies below or above it?"

Here the experts are bisecting their initial ranges. We have found that providing the form in Fig. 4.2 can be helpful to experts when making these judgements because they can visualise the line that they are bisecting. Here, the facilitator should be mindful of another cause of poor judgement: the range-frequency compromise where people tend to want to share probability reasonably evenly across the range. The symptoms of this are easy to spot in the roulette method (described later), and,



Lower bound (L)	100	Upper bound (U)	200
	Median (Q2)	130	
Lower quartile (Q1)	120	Upper quartile (Q3)	150

Fig. 4.2 An example sheet for capturing each expert's judgements

when eliciting quartiles, the effect of this is that experts will tend to specify their lower and upper quartiles in the middle of the ranges under consideration.

The second most popular method within SHELF is the roulette method where the experts are asked to build histographic representations of densities that reflect their beliefs about the quantities of interest. In practice, we have found that this method can work well when experts have a solid background in statistics or if they have been exposed to probability density functions that encode beliefs through other elicitation exercises. This method shares that same start as the quartile method with the experts being asked to provide a range for the quantity of interest, but, using this method, the experts must agree to share the plausible range to make the judgement entry easier when using SHELF's R functions. Once the range is specified, it is split into ten equal length "bins" where the experts will be able to allocate "chips". Using ten here means that it is simple to do the mathematics and that the experts will not be making judgements to an unrealistic level of accuracy. Of course, if there are reasons why more or fewer are needed, then the number of bins can be changed; however, it is not useful to deviate from equal spacing because the interpretation of the chips will need to change. Experts should ideally have preprinted sheets with ten bins marked out, and they will have space to add in the bin boundaries. Experts will also have been given a number of chips each to place. It is recommended that the experts are given physical chips rather than asking them to mark the sheet with pen, because this engages the experts and allows a visual representation of their uncertainty (O'Hagan et al. 2006; Oakley and O'Hagan 2014).

In order to place chips in the bins, experts are asked to consider the relative probabilities of the true value for the quantity of interest falling within each bin (like placing bets on a roulette table). They should also be shown an example of chip allocation based upon a small number of chips so they understand how the chips translate into probability statements. For example,

	Bin.lower	Bin.upper	Expert1.chips	Expert2.chips	Expert3.chips	Expert4.chips
1	100	110	0	0	0	0
2	110	120	0	0	0	0
3	120	130	0	0	0	0
4	130	140	0	0	0	0
5	140	150	0	0	0	0
6	150	160	0	0	0	0
7	160	170	0	0	0	0
8	170	180	0	0	0	0
9	180	190	0	0	0	0
10	190	200	0	0	0	0
11						

Fig. 4.3 Entering individual judgements with the roulette method

"we have five chips in bin A, three chips in bin B and two in bin C; this translates to a 50% probability for the true value to be in bin A."

The facilitator may advise the experts that a realistic expression of uncertainty should involve concentrating chips in relatively few bins, but not too few. Also, the overall specified range is dictated by physical bounds, the fact that it is considered implausible for the true value to be outside these bounds suggests that the probability of being outside is so small that even a single chip would give too much weight to those regions. The experts should be offered no more than 30 chips (with no expectation that they will use them all) and the experts can adjust their deployment of chips until they are satisfied with the distribution.

As mentioned before, this stage should be completed individually; therefore, some consideration has to be made regarding the seating arrangements of the experts and the size of the chips to help prevent copying in the allocations. We have found it beneficial, when preparing for the fitting stage and subsequent aggregation, to photograph each expert's chip allocations. This way the experts do not need to shout out their judgements and the facilitator has a hard copy of the judgements for entering the values into the computer (see Fig. 4.3) and for the records of the EKE exercise.

Within the SHELF documents, there are also supporting forms and guidance for the tertile method (which is analogous to the quartile method, but the questions lead to tertiles rather than quartiles) and hybrid methods that use direct probability judgements alongside the one of the aforementioned three methods.

After the experts have made the individual judgements, the facilitator collects in the information without anyone declaring the values that they have specified. The facilitator will then record the judgements just assigning them to suitable aliases (for example, "expert 1", "expert 2" and so on).

4.2.6 Distribution Fitting

The judgements are not enough (at least in the continuous variable case) to specify a probability distribution fully. Extra assumptions about distributional form are needed to arrive at a fitted distribution. Here, the facilitator's experience and knowledge of statistics is crucial. We have found that it is worthwhile being flexible here and giving the experts chance to comment on the distributional choices if they feel able to (Gosling et al. 2012, 2013).

In order to fit a distribution to the judgements, it is common to employ a least-squares fitting procedure (O'Hagan 1998). In such a procedure, the elicited judgements are compared against the corresponding theoretical quantities from a fully-specified probability distribution. We select the parameters of that distribution by finding the parameters that minimise the squared difference between the elicited and the theoretical quantities. For instance, if we have elicited the median, Q_2 say, and the lower and upper quartiles, Q_1 and Q_3, we wish to find the distribution with cumulative distribution function $F(.)$ that minimises

$$F(Q_0)^2 + [F(Q_4) - 1]^2 + \sum_{i=1}^{4} \{[F(Q_i) - F(Q_{i-1})] - 1/4\}^2 , \tag{4.1}$$

where Q_0 is the specified lower bound and Q_4 is the upper bound. The measure in (4.1) is easily extended to more judgements of different types. Because we have a finite number of judgements, there are infinitely many distributions that will minimise this measure so, in practice, the facilitator must use their judgement as to which distribution (or family of distributions) would best accommodate the expert's judgements. Minimising this measure does not necessarily result in an appropriate distribution being fitted. For example, if the experts' judgements indicate that the distribution is likely to be heavily skewed, we will not be able to find a set of parameters for a normal distribution such that that distribution is an adequate representation of the experts' beliefs. Often, we have information on the likely shape of the distribution prior to the judgements being made (for instance, we may know that the distribution is bounded between 0 and 1); we can use this information to choose appropriate distributions on which to attempt the fit. For a categorical variable, this type of fitting is unnecessary because the full distribution will have been defined by the expert in the previous step.

The R functions provided with SHELF allow the facilitator to enter and store the judgements for each of the experts and automate the fitting process. Figure 4.4 shows a screen shot from using those functions to enter information from six experts using the quartile method. The minimisation routine is fast enough for the fitting to be completed in real-time. The functions give the facilitator the option of choosing the distribution family or allowing the program to select the best fitting distribution family automatically. Version 2 of SHELF has the following distributions built in for univariate quantities:

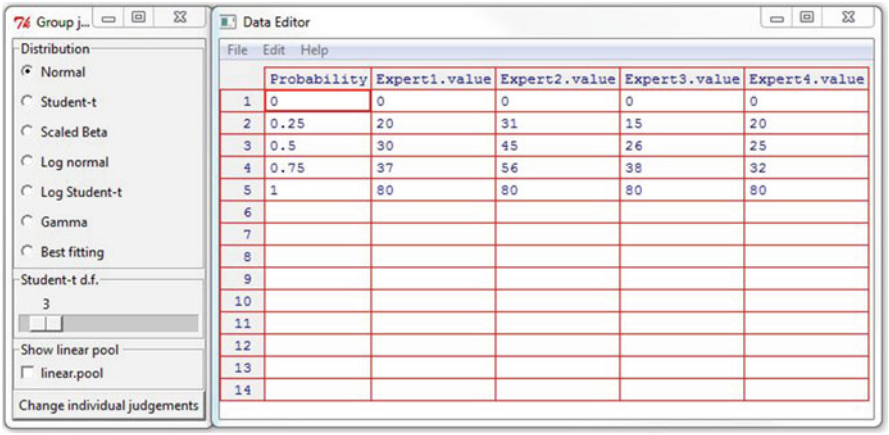


Fig. 4.4 Entering individual judgements and choosing distributional families

- normal,
- Student's-*t*,
- scaled beta,
- log-normal,
- log-Student's-*t*,
- gamma.

Due to potential identifiability issues given the low number of judgements, the facilitator is required to choose the number of degrees of freedom when using the Student's-*t* or log-Student's-*t* distributions.

4.2.7 Aggregation of Distributions

The individual distributions should be shown to the experts, but, at this stage, revisions are not invited unless an expert is insistent that the fitted density badly distorts their beliefs. However, the facilitator may take the opportunity to work with the experts to correct any incoherences that may have occurred in their judgements (as suggested in Brown and Lindley 1982).

As a guide for the facilitators, the linear opinion pool of Stone et al. (1961) can be calculated as part of the SHELF R functions. The linear opinion pool offers a mathematical framework to form a prior distribution for the whole group. If we have $f_i(\theta)$ to represent the density fitted to the ith expert's judgements about θ, then the linear opinion pool can be used to calculate an aggregated density $f_A(\theta)$ for N experts:

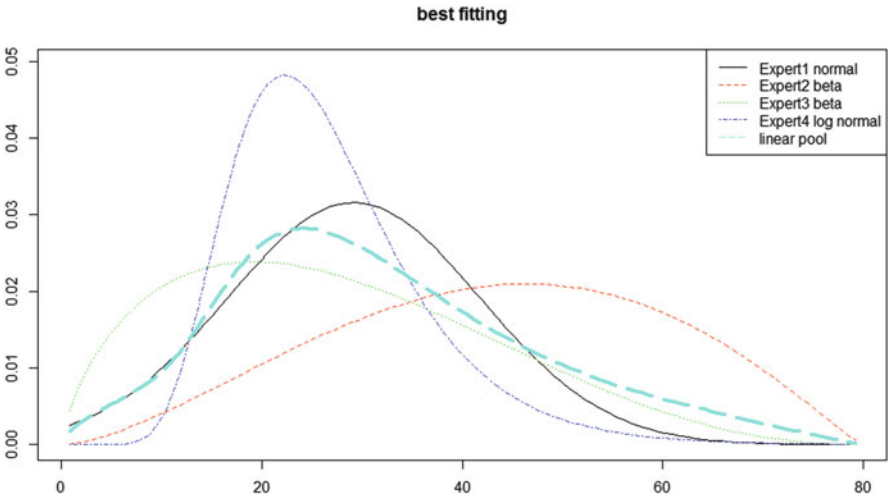


Fig. 4.5 Example distributional fits within SHELF with linear pool shown

$$f_A(\theta) = \frac{1}{N}\sum_{i=1}^{N} f_i(\theta). \tag{4.2}$$

The combining of beliefs means that each individual's opinion must be pooled together in a way in which everyone is satisfied. French (1985) investigated the problems of putting opinion-pooling techniques into practice and the problem of distinguishing between the experts and the decision makers. French commented that this is an extremely difficult task and a simplistic, democratic technique for constructing consensus probabilities is the only viable method. In our experience, we found that asking the experts to discuss the quantity and relevant evidence and then arrive at judgements that were group consensuses has been effective at capturing a group's opinion. In contrast, mathematical aggregation of the individual expert's beliefs could lead to distributions that no one person agreed with.

Some practitioners of the SHELF method choose to share the pooled distribution with the group (see Fig. 4.5), but we have found that this can distract the experts and may be used as a fall-back by the experts who may prefer a seemingly objective way of combining the judgements.

Producing the pooled distribution can help the facilitator to identify experts that have relative extreme views in terms of the location of their judgements and the amount of uncertainty. This information is useful in the next stage where feedback on what has been collected so far (and fitted) is presented to the experts. If feasible, the facilitator should compute the median and quartiles of an equally-weighted average of the density functions. Ideally, these should not be revealed to the experts immediately, but may be used at the facilitator's discretion in the next stage.

4.2.8 Feedback on Distributions

In the feedback stage, the experts judge the assumption that the fitted distribution is an adequate representation of their beliefs as a group. This is similar to the satisficing prior distribution of Winkler (1967): the experts will be content to adopt that prior distribution at that moment in time as being representative of the group. As the expert cannot differentiate between well-fitted distributions, there are an infinite number of distributions that the expert would accept as their own distribution. As the experts are often not experts in interpreting plots of density functions, there is little point to just showing a graph and asking if that conforms with what they had in mind. Even if the experts could do this, we could not expect the experts to differentiate between several distributions that have similar characteristics. Often, the experts will find it useful to be given statements or summaries about the fitted distributions that are in a similar format to the original questions. For instance, we might produce a credible interval based on the fitted distributions. Appropriate questions at this stage include:
"Your responses suggest that a value of X (maybe using the 99th percentile here) is highly unlikely; do you agree with this?"
and
"According to your judgements, there is a 1 in 5 chance of the variable falling outside the range (u,v)?"

A picture like that of Fig. 4.5 is shown to the experts (with or without the pooled density) and the facilitator prompts a discussion of the different distributions. Questions about why one distribution is far from another in terms of location and spread can help to identify incoherences in an expert's judgements and reasons for differing opinions. Because there is potential for wildly contrasting views and that this is the first time that each individual is aware of the others' opinions on the quantity of interest, this discussion can take a substantial amount of time. The facilitator should be careful to only cut short the discussions if the experts are no longer exchanging information and arguments, but are just repeating opinions. The facilitator needs to manage the discussion so that divergent views are properly considered, and to ensure that strong personalities or groups with overlapping experiences do not dominate inappropriately. A skilled facilitator will be aware of psychological literature warning of group dictators and the polarisation that can occur in extended group discussions (Myers and Lamm 1975; O'Hagan et al. 2006).

When using the quartile method, the aim of the facilitator is to find values for the three quartiles that the experts can agree on as being representative of the group's views. The agreed median will inevitably be some sort of compromise. Before discussion, there are two components of uncertainty in the group—uncertainty that each expert has and is expressed in that expert's quartiles, as well as variability between the experts' judgements. The agreed quartiles should reflect the group's overall uncertainty that remains after the discussion thus capturing lack of knowledge and variability across the experts in the room.

It is possible that our reported distribution for the experts' density may not agree with what the experts really thinks: the distributions we have chosen might be inappropriate and/or the experts may have given us probability judgements that do not really match their beliefs. In this case, the experts could give us different or more information to help update the fitted distribution.

This is part of the "feedback loop" where stages (5)–(8) are repeated until the experts are satisfied that their beliefs have been captured adequately. The SHELF forms have space to record the process of iteratively fitting, feeding back and revising the group judgements. In an EKE exercise, we must be careful about the fact that an expert might just get bored with the process and accept anything after a few iterations of the feedback loop.

Throughout this process, the value of a skilled facilitator can be seen. Because of the interaction with the experts and the gauging of reaction to fitted distributions, a poor facilitator could influence the outcome of the EKE exercise more than if a more straightforward mathematical approach was taken to the aggregation. However, we have found in numerous elicitation exercises that it is these discussions that help the experts understand how to make judgements about their uncertainty and more detailed information about each individual's reasoning will be captured that could be beneficial to the problem owner and other subsequent users of the results. Often, conflicts that are apparent in the experts' individual fitted densities are due to misunderstandings in making coherent quantitative judgements and differences in perceptions of the relevant evidence.

4.2.9 Completing the Exercise

Stages (4)–(8) must be repeated for each quantity of interest in the exercise. If some of the quantities being considered are closely related to quantities already covered, there may be much overlap in the information sharing stage and there may be an instinct for the experts to repeat their previous judgements. When this situation arises, the facilitator may wish to order the quantities being considered to help prevent judgement reuse. However, in long elicitation sessions, fatigue can set in with all concerned due to the intensity of the process and experts can begin to repeat judgments even if quantities are not related. Therefore, when employing SHELF for multiple quantities (assuming it is important to elicit information about all the quantities), adequate time should be left for breaks and the facilitator should be realistic about what can be achieved in a single session.

As already stated, it is important to be as transparent as possible when using SHELF due to the inherently subjective nature of EKE. Therefore, whenever the EKE has been used, we must report all the information about the process alongside the elicited consensus distribution. Here is a list of information that could be included in the supporting documentation for an EKE exercise if the SHELF protocol has been followed:

- List of experts along with their expertise and any declarations of interest;
- List of agreed and unambiguous descriptions of the quantities of interest;
- Experts' answers to qualitative questions as described in Sect. 4.2.4;
- Experts' answers to qualitative and quantitative questions as described in Sect. 4.2.5 (including any disagreements between the experts);
- Details of the distributional fitting procedure;
- Experts' answers to feedback questions as described in Sect. 4.2.8;
- Details of any revisions and the experts' reactions to them;
- The final fitted consensus distributions for each of the quantities of interest.

Throughout the documentation, once the list of experts has been given, only aliases should be used as described at the end of Sect. 4.2.5. Of course, there are situations where the experts must remain completely anonymous, and this will have to be stated in the supporting information.

An important part of each of the SHELF forms is the recording of the timings for each stage. For any users of the results, this is indicative of the amount of effort spent on each quantity of interest. For the facilitator, such information is invaluable when planning future SHELF sessions.

4.3 Notable Applications of the Framework

SHELF has been widely available since 2008 and, as such, has been applied in a number of EKE exercises. As mentioned earlier, the early development of SHELF was done with healthcare in mind (see Sect. 4.3.1), but it has also been taken up in the environmental sciences (see Sect. 4.3.2) and in business planning amongst other applications (see Sect. 4.3.3). In this section, we briefly outline some of the method's applications and comment on the use of the results.

4.3.1 Healthcare and Medicine

SHELF was considered as a tool for clinical trial planning in Kinnersley and Day (2013) where its utility was demonstrated in a case study involving a hypothetical trial in ankylosing spondylitis and in Ren and Oakley (2014) where assurance calculations based upon expert judgements. The experts were asked to use the roulette method to specify their individual probability densities. In this exercise, the experts found that the roulette method was appropriate (despite some difficulties in specifying tail probabilities); however, it should be noted that the experts in this case were statisticians working on clinical trials and had significant exposure to expressing uncertainty using probability density functions. In a recent exercise to inform decisions around the treatment of persistent cervical cancer, a similar scheme was used where experts were asked to declare their background and

relevant expertise prior to making judgements (Meads et al. 2013). In that exercise, the roulette method was used to elicit individual judgements with a paper based exercise with no group interaction for the formal elicitation part and mathematical aggregation was done without the experts' intervention. Related to this, the reporting guidelines for expert judgements in model-based evaluations in health economics as offered in Iglesias et al. (2016) aligns closely with what is reported at the end of a SHELF exercise.

The SHELF method was implemented as part of characterisation of a human toxicological safety assessment in Gosling et al. (2013). The method was used to gather knowledge from risk assessors and toxicologists on the potency of certain chemicals and the relationships between various animal and non-animal potency tests. The final results of the EKE were used to populate a Bayes linear updating scheme aimed at quantifying the experts' beliefs about true human potency, and, as such, the facilitators needed to elicit information on correlations between quantities of interest that were arrived at by using SHELF to elicit beliefs about differences between the various quantities. In the information sharing and structuring phase of SHELF (see Sect. 4.2.4), the experts identified conditional independencies to make this task more manageable.

SHELF has been implemented for the elicitation of expert beliefs for veterinary treatments Higgins et al. (2012). In this study, the cure rates of cows under different treatments in intra-mammary dry cow therapy was considered. Practitioners were paid at a rate of £100 per hour for their time to encourage participation . Also, the 24 experts were selected at random from a pool of 77 practitioners from 13 veterinary practices using stratified sampling to get a range of view across the veterinary practices. Due to the large number of experts, they were split into five groups and the consensus distributions for each cluster were reported. The facilitators used the quartile method and fitted beta distributions (as the variables were bounded between zero and one).

In Gosling et al. (2012), SHELF was used to quantify government experts' opinions on the costs and rate of exotic disease outbreaks in UK livestock. In contrast to Higgins et al. (2012), the experts were selected from an internal pool of employees based upon their expertise. The expert elicitation sessions covered 40 variables of interest and elicitation workshops were ran over four non-consecutive days. Here having a skilled facilitator was crucial to the exercises success; however, the long days and sheer number of variables being considered meant that the experts repeated earlier judgements and the task would have been more difficult if the quantities were not so closely related.

4.3.2 Environmental Sciences

SHELF has also been applied in the environmental sciences partially due to the interests of the creators and partially due to the rise of uncertainty analyses and Bayesian methods in that discipline.

SHELF was employed in error modelling for geological boundaries in Lark et al. (2015). In this study, five geologists were asked to consider errors in boundaries under six different hypothetical scenarios. These scenarios were designed prior to the EKE sessions, but the facilitators allowed the experts to have input in altering the scenarios to make them more realistic and to avoid future ambiguities. The training exercise for the experts focussed on the distribution of the ages of delegates to the 2013 European Geosciences Union congress. Beliefs about the age of a randomly selected individual involves both natural variability and uncertainty, which was akin to the ultimate quantity of interest because there is a population of errors for which the experts are unsure of the distribution.

Uncertainty analyses of complex computer models require the specification of distributions for the input parameters, which are often specified through EKE (Oakley and O'Hagan 2002). When modelling the atmospheric carbon flux, EKE procedures similar to SHELF were employed in Kennedy et al. (2008) around the time SHELF was being developed. A formal use of SHELF for the purpose of specifying input distributions can be found in Lee et al. (2013) where 28 model input parameter were considered by aerosol modellers.

4.3.3 Other Applications

There are several other documented uses of SHELF for substantial applications in other fields.

The UK's Centre for Workforce Intelligence employed SHELF to inform future workforce planning for the health and social care system in England (Centre for Workforce Intelligence 2015). Before employing SHELF, they reviewed the protocol alongside traditional Delphi techniques that have been employed in horizon scanning for similar initiatives (Linstone and Turoff 1975). They concluded that the number of quantities that could be considered in a SHELF exercise was far less than what is possible in traditional Delphi exercises and a greater level of effort was required by the experts. These considerations led to SHELF only being employed for quantities that are of most importance to the model of workforce planning. Also, in these studies, gaps in the method were highlighted with respect to capturing correlations between quantities.

An adaptation of SHELF was used to model expert opinions of failure times for water pipe networks (Scholten et al. 2013). The standard SHELF procedure was followed until the point when the experts had to make their individual quantitative judgements. The individual judgements were made over three rounds:

1. Each expert added paper clips to their own time line to build up a picture of when failures would occur;
2. The quartiles were then elicited from each expert with reference to the results of the first round;
3. Each expert had the opportunity to make qualitative statements about the shape of the density that they felt would represent their beliefs.

Instead of combining the individual distributions using behavioural aggregation, linear pooling (see Stone et al. 1961) and using an hierarchical model of the expert opinions (akin to Lindley et al. 1979).

A study of expert opinion on medium-term effects of the economy and climate on the energy sector (Usher and Strachan 2013) employed a SHELF-like scheme when eliciting judgements from individual experts using the quartile method. That particular study stopped short of using behavioural aggregation to combine the experts' distributions because the focus was on the range of opinions and they just reported the individual judgements alongside the linear opinion pool. The experts were training in this EKE exercise using the almanac question: "*What is the length of the Moscow underground network in kilometers?*"; this was appropriate here because the quantities of interest were subject to epistemic uncertainty alone.

4.4 Extensions of the Framework

Many extensions and adaptations to SHELF have have been proposed and implemented. There have also been efforts to make the accompanying R software more flexible and user friendly. The R package SHELF has been created to give updated versions of the supporting R functions and to host additional functionality (for example, this includes a function for viewing experts' judgements as a histogram, which can highlight incoherence in the judgements; see Fig. 4.6). The latest version (1.2.0, released 17th August 2016) has a browser-based interface and includes options for eliciting multivariate distributions (see Sect. 4.4.1). There has also been interest in developing the protocol to handle expert groups that cannot physically meet (see Sect. 4.4.2) and to extend the experts' judgements to attempt to capture more qualitative information about the uncertainties that cannot be quantified (Gosling et al. 2012, 2013). In this section, we review the extensions to the original method and some approaches that have deviated from the original framework.

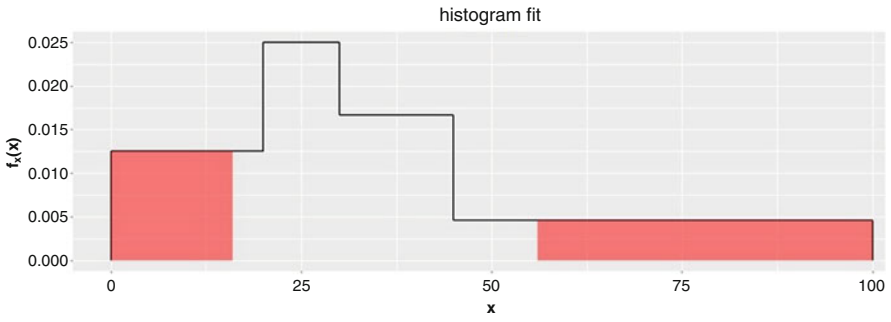


Fig. 4.6 Visualising judgements in the SHELF R package

4.4.1 Elicitation for Multivariate Quantities

The first extension for multivariate quantities considers fitting distributions for sum-constrained vectors. A method was developed in Zapata-Vázquez et al. (2014) for fitting Dirichlet distributions to judgements on such vectors that avoids direct questioning on correlations or dependencies across vector elements. The individual judgement stage of SHELF proceeds in this case by separately eliciting judgements on each of the elements of the vector. The distribution fitting finds marginal beta distributions for the set of judgements for each element and the means of those distributions are taken alongside to specify a Dirichlet distribution. However, to complete the specification, an extra parameter is needed that controls the spread in the distribution. This can be calculated through optimisation where the fitted Dirichlet distribution's marginal standard deviations are selected to best match the standard deviations from the fitted beta distributions. The facilitator is free to make other choices here (like matching the most diffuse fitted beta distribution), but the idea is to get a number of indicative distributions before starting the feedback and aggregation stages of the process. The feedback options in the SHELF package are limited to the marginal distributions, but it may be helpful to expose the experts to some ternary plot representations of the fitted distributions as well as conditional probability statements. The difficulty here is that the Dirichlet distribution is inflexible in terms of correlation structure. The original paper (Zapata-Vázquez et al. 2014) also suggested a more flexible form of the Dirichlet distribution based on Dickey (1983), but that is not yet implemented in the SHELF R package.

For more flexibility in modelling expert's uncertainty for two quantities of interest, an extension of SHELF utilising copula-based representations of probability distributions has been added to the SHELF R package following the approach of Clemen and Reilly (1999). The experts are asked to make judgements about the two quantities of interest separately so that marginal distributions can be fitted (essentially, following the univariate SHELF scheme as described in Sect. 4.2). Following this, they are asked to make judgements about the concordance probability. The concordance probability here is of the form

$$\Pr\{[(X > x) \cap (Y > y)] \cup [(X < x) \cap (Y < y)]\},$$

where X and Y are the two quantities of interest and x and y are median values taken from the experts' fitted marginal distributions (or original judgements if appropriate). The R function allows the experts to visualise the consequences of their specified probability in the form of a scatterplot as shown in Fig. 4.7. When fitting univariate distributions, the facilitator must choose which distributions might be appropriate to capture the experts' beliefs adequately. Similarly, there is a choice for the copula form. Currently, the R function only offers the bivariate normal copula for fitting bivariate distributions.

Given the growth of Bayesian approaches in geostatistics, Truong et al. (2013) extended SHELF to accommodate expert judgements about variograms that specify

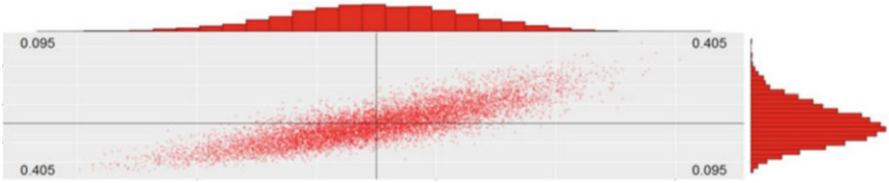


Fig. 4.7 Tool for eliciting concordance probabilities

correlations over spatial fields. Like in the copula approach, marginal distributions are elicited using the standard approach for values of the variable at multiple sites. Under assumptions of stationarity over the spatial field, judgements about differences in the variable at different spatial locations can be used to capture information on the variogram for the entire spatial field. Again, assumptions were made about the underlying spatial process when fitting to the judgements (Gaussian process with Matérn correlation function).

It is clear that there is still some effort to needed to extend SHELF for general multivariate problems. Simple conversion to separate univariate EKE exercises using products or differences for quantities on a similar scale (like in Gosling et al. 2013) will prove to be too inefficient given the number of pairs of variables that need to be considered and the fatigue that can set in during long elicitation exercises. Of course, there are efficiencies to be gained in the structuring phase by identifying (conditional) independencies, but the facilitator may be end up making the experts accept more assumptions about their distributions when it is difficult to make such judgements.

4.4.2 Distributed Experts

It is not always possible to get the experts together at the same time. Internet connectivity allows for online applications to be developed that can aid the remote capture of expert judgements. A SHELF exercise requires skilled facilitation and difficulties can arise if the group are not guided through the process by such an individual (see French 2007; Morgan and Henrion 1990). If the process can be facilitated remotely through telephone or video conferencing, it can be challenging for the facilitator to track group dynamics and respond appropriately. However, due to time restrictions and travel costs, it can often be infeasible for the experts to be in the same room. The R program behind the SHELF implementation has been modified for remote use in the online MATCH uncertainty elicitation tool (Morris et al. 2014). MATCH is implemented as a browser based elicitation tool (see Fig. 4.8) and covers the functionality of the SHELF R functions for fitting univariate distributions via five elicitation methods. MATCH covers stages (5)–(6) of SHELF as described in Sect. 4.2. The facilitators may choose to conduct the first and final

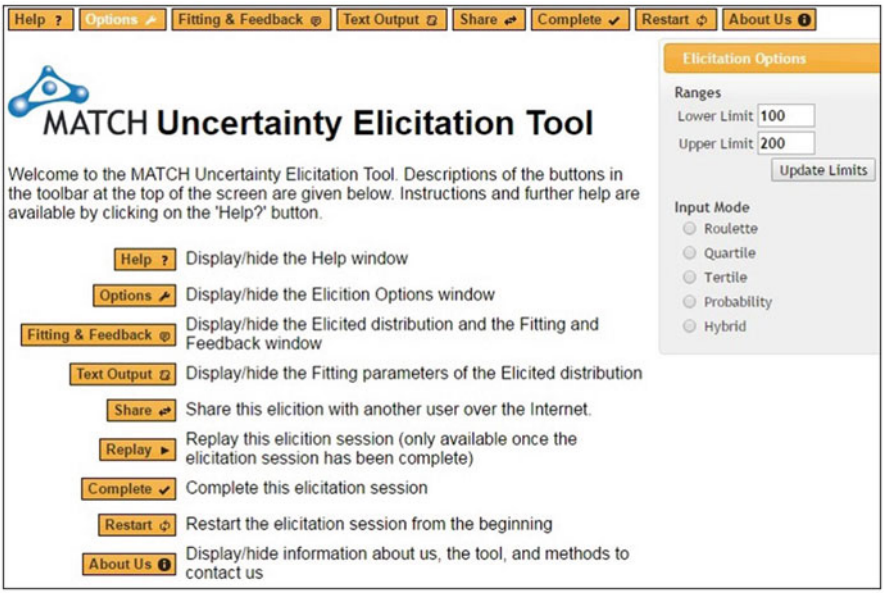


Fig. 4.8 The home page of the MATCH uncertainty elicitation tool

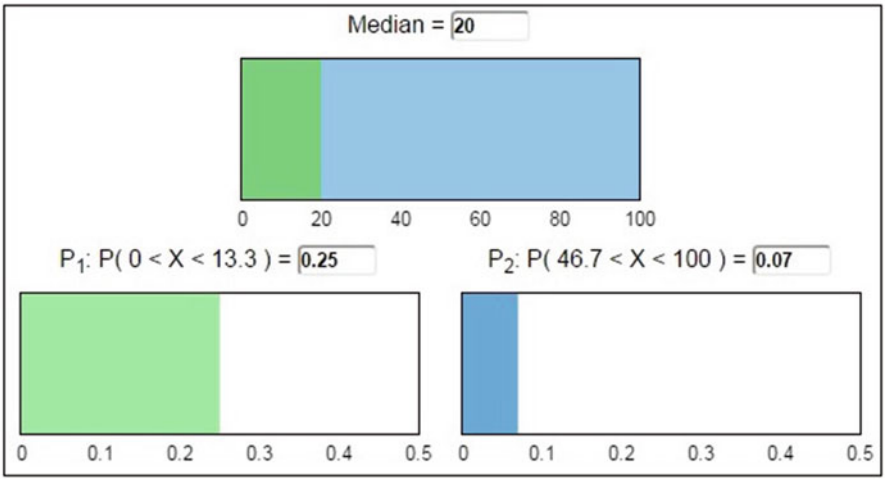


Fig. 4.9 Entering judgements on median and probabilities in MATCH

parts of SHELF using a teleconference and use MATCH as part of remote one-to-one interviews to get the individual distributions.

A key feature of the MATCH implementation is that the browser window can be shared across multiple users so that the facilitator can guide the individual though the SHELF steps. In Figs. 4.9 and 4.10, screen shots are given for the hybrid method

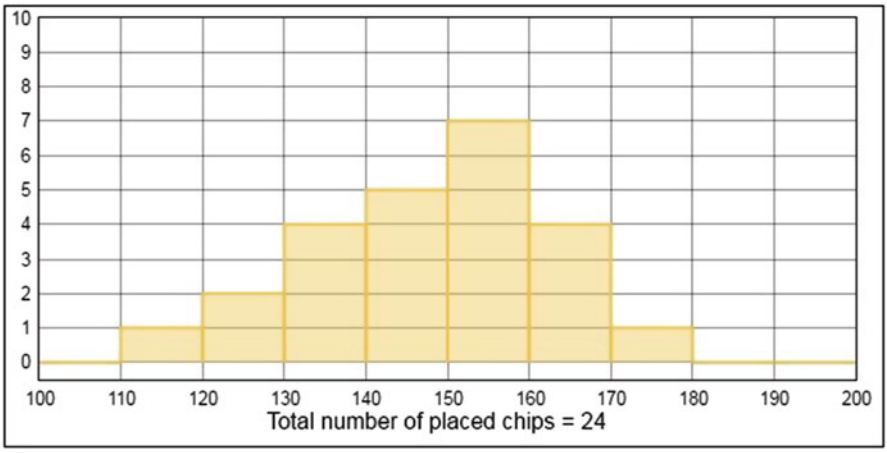


Fig. 4.10 Entering judgements via the roulette method in MATCH

and roulette method respectively. In the hybrid method, the experts are asked for their medians and two probabilities and they are presented with visualisations of those judgements that might help some experts. For people with limited statistical training, it can be challenging to think about medians and probabilities or to build up a probability density function so having a facilitator (albeit remotely watching) to guide can be crucial. This approach follows the recommendation of the joint use of facilitation and remote software given in Anson et al. (1995).

In MATCH, once the judgements have been made, the experts can get instantaneous feedback through a plot of fitted distributions alongside percentiles (as shown in Fig. 4.11). This part of the software uses the fitting procedure described in Sect. 4.2.6 and enables the users to select which percentiles they want to consider in the feedback phase.

A similar protocol to SHELF has also been implemented as part of the EC-funded UncertWeb project (Bastin et al. 2011). In this implementation, named 'The Elicitator', online forms are provided to capture all the briefing documents for an elicitation exercise, to invite experts to participate in the elicitation exercise and to keep track of the progress of all the experts during the exercise (see Fig. 4.12).

Like in the MATCH implementation, the experts are exposed to real-time feedback to their judgements through plots of the fitted densities (see Fig. 4.13) and selected feedback questions. The aggregation stage can be done automatically using a Vincentization procedure (see Thomas and Ross 1980) or the results of the individual distribution fitting can form the basis of discussions in a shorter teleconference or face-to-face meeting.

The Elicitator also has an option to collect expert judgements on categorical variables through a roulette-type method. Here the experts are given an inexhaustible supply of beans to place in various pots that represent the possible values of

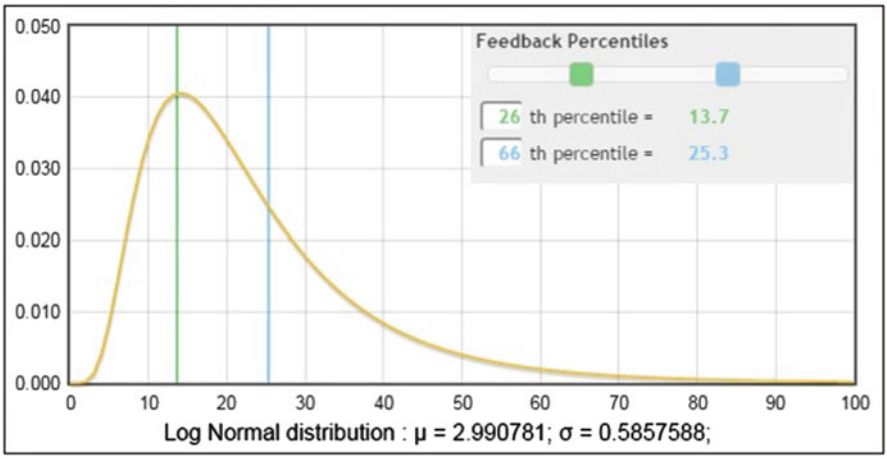


Fig. 4.11 Fitted density and feedback percentiles in MATCH

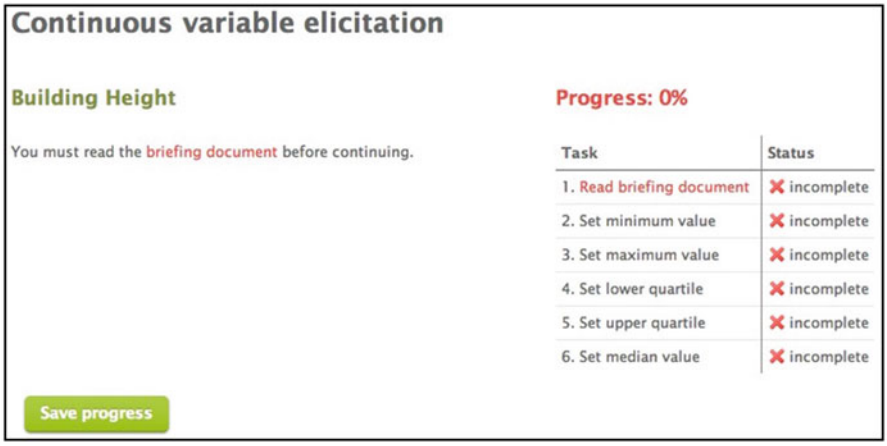


Fig. 4.12 Tracking the progress through part of a SHELF exercise with The Elicitator

the variable (see Fig. 4.14). Again, in the feedback phase the experts are asked questions about the probabilities of falling in one of multiple categories to check for coherence.

Remote expert elicitation exercises have been attempted through several rounds of emails using SHELF as a basis. Such innovations could be especially valuable to facilitators looking to apply a behavioural aggregation method with the costs (but without the benefits) of meeting. SHELF has been adapted to an Microsoft Excel spreadsheet-based tool that has been used via email for several different applications and has been piloted for capturing disease prevalence (Sperber et al. 2013) and for food safety assessments (which is in fact a modified Delphi technique) (European

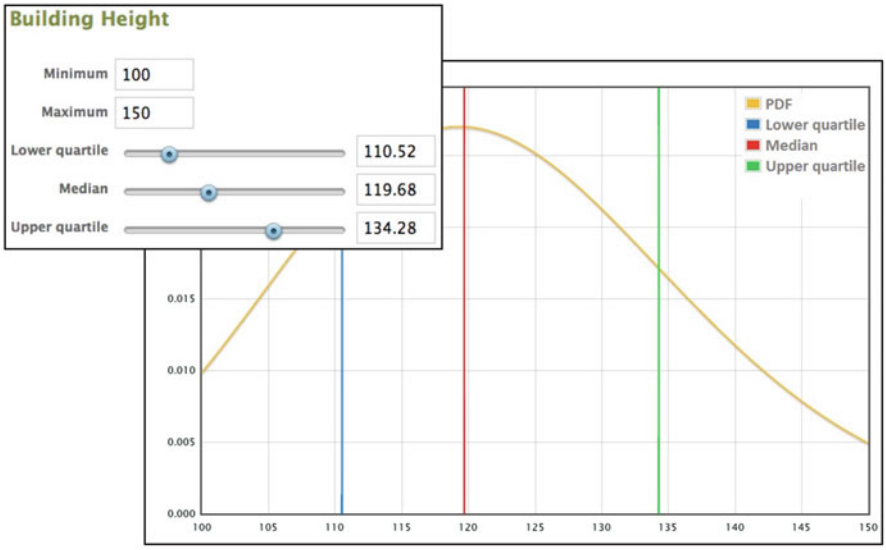


Fig. 4.13 Specifying bounds and the three quartiles with real-time feedback in The Elicitator

Food Safety Authority 2014). The flexibility of modern spreadsheet software means that the necessary information can be easily recorded and instantaneous feedback can be provided in bespoke ways with little knowledge of graphical user interface development. Also, health experts are likely to be familiar with spreadsheets and they (and the facilitators) may even use them for some of their probabilistic modelling. However, it is important that the experts buy-in to an elicitation exercise and a long spreadsheet-based questionnaire may not be ideal.

The Delphi method is well-known technique for capturing expert opinion (Dalkey 1969; Rowe and Wright 2001). A Delphi exercise involves involves several rounds of questionnaires where each expert has access to the opinions of the other participants. In order to make the Delphi technique more relevant to probabilistic assessments, it has been suggested that the SHELF method could be embedded within a Delphi-like process (European Food Safety Authority 2014). Here the idea is to use the SHELF process to capture all the relevant information about the elicitation exercise whilst avoiding the need to meet. The experts are asked to make their judgements separately and then have several rounds of revisions based on the other experts' judgements and the fitted consensus distribution. Like SHELF, the relative simplicity of technique makes it relatively easy to produce software, but given the remote nature of these exercises and lack of facilitator interaction, great care must go into software design.

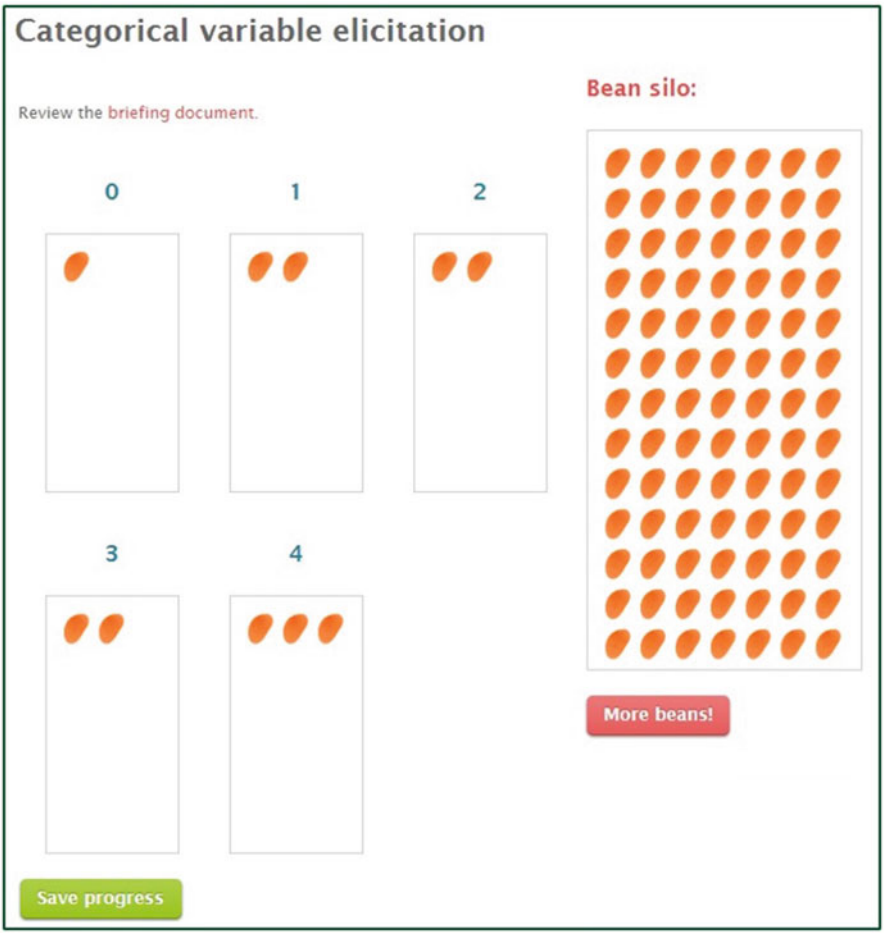


Fig. 4.14 Capturing individual judgements on categorical variables in The Elicitator

4.5 Discussion

The core principles behind the SHELF method are:

- group information sharing,
- capturing individual judgements whilst helping to avoid incoherence,
- facilitating discussions to reach consensus,
- documented procedure open to external scrutiny.

SHELF is designed to create an environment within which the experts can make useful judgements whilst minimising the effects of the usual heuristics and biases. A SHELF exercise should also lead to a set of documents that are accessible to users of

the elicited distributions so they understand the provenance of the results. However, no matter the amount of care put into the formulation of an elicitation protocol, there is no guarantee that it has been applied correctly. Each part of SHELF and the questions stem from years of experience and research into elicitation methods. An unskilled facilitator could undo all of the good intentions by mismanaging the group or deviate from the ordering of the process.

The final consensus distribution may be thought of as the distribution a rational independent observer may arrive at after taking into account the discussions and judgements made at the meeting. It is clear that this is wholly dependent on the experts in the room. Therefore, great care must be taken when setting up a SHELF exercise to find experts who will have useful opinions for subsequent users of the results. Further guidance on expert selection is given in Ayyub (2001) and Morgan and Henrion (1990). In scientific studies, we would like to be able to reproduce results of an experiment. An EKE exercise is an experiment of sorts where the facilitator is attempting to measure the experts' beliefs about some quantities of interest. It cannot be guaranteed that the same group of experts will give the same judgements at a different time (even if their baseline knowledge of the variables has not changed). SHELF enables us to capture what the experts were thinking at the time of the judgements and provides a framework within which we can easily highlight differences between two elicitation sessions.

Of course, there is potential for great disparity between the experts and conflicts may occur in both the discussion and in the fitted distributions. Another innovation to help in this situation was suggested in the veterinary treatment work of Higgins et al. (2012) to help understand the differences in expert judgements. As part of the feedback phase (see Sect. 4.2.8), the differences between experts were highlighted by using multidimensional scaling where each expert was represented in two dimensions based upon their vector of judgements about the quartiles. This is a useful tool for the facilitator to demonstrate differences in opinions and clusters of experts with similar judgements (and is also easy to implement in R in real time). Such a visualisation could be used to help guide discussion in the feedback and revision stages of SHELF especially when conflicts occur. If these disparities cannot be resolved, there is scope within SHELF to accommodate the various viewpoints, report distributions for each conflicting view point and record the reasons for the differences (which will likely be differing interpretations of the same evidence).

As already discussed in Sect. 4.4, many efforts are currently being undertaken to improve and extend the SHELF R functions that support the framework's implementation. On the theoretical side, more research is needed on handling multivariate quantities within the framework and taking account of uncertainties that the experts feel unable to quantify. Given the choices that the facilitator makes when fitting a distribution, the reported distribution could also take into account the uncertainty in distribution fitting (as infinitely many distributions would be satisfactory) using the approaches of Gosling et al. (2007) and Oakley and O'Hagan (2007).

References

Anson R, Bostrom R, Wynne B (1995) An experiment assessing group support system and facilitator effects on meeting outcomes. Manag Sci 41:189–208

Ayyub BM (2001) Elicitation of expert opinions for uncertainty and risks. CRC Press, Boca Raton

Bastin L, Williams M, Gosling JP, Truong P, Cornford D, Heuvelink G, Achard F (2011) Web based expert elicitation of uncertainties in environmental model inputs. In: Abstracts of the European Geosciences Union General Assembly 2011

Bolger F (2018) The selection of experts for (probabilistic) expert knowledge elicitation. In: Dias LC, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgment. Springer, New York

Brown RV, Lindley DV (1982) Improving judgment by reconciling incoherence. Theory Decis 14:113–32

Centre for Workforce Intelligence (2015) Elicitation methods: applying elicitation methods to robust workforce planning. http://www.cfwi.org.uk/publications/elicitation-methods-applying-elicitation-methods-to-robust-workforce-planning. Cited 15 Sept 2016

Clemen RT, Reilly T (1999) Correlations and copulas for decision and risk analysis. Manag Sci 45:208–24

Dalkey N (1969) An experimental study of group opinion: the Delphi method. Futures 1:408–426

Defence Science and Technology Laboratory (2015) The probabilistic elicitation of subjective data. https://www.gov.uk/government/publications/the-probabilistic-elicitation-of-subjective-data. Cited 15 Sept 2016

Dickey JM (1983) Multiple hypergeometric functions: probabilistic interpretations and statistical uses. J Am Stat Assoc 78:628–37

European Food Safety Authority (2014) Guidance on expert knowledge elicitation in food and feed safety risk assessment. EFSA J 12(6):3734

French S (1985) Group consensus probability distributions: a critical survey. In: Bernardo JM et al. (eds) Bayesian statistics 2. Oxford University Press, Oxford, pp 183–202

French S (2007) Web-enabled strategic GDSS, e-democracy and Arrow's theorem: a Bayesian perspective. Decis Support Syst 43:1476–1484

Garthwaite PH, Kadane JB, O'Hagan A (2005) Statistical methods for eliciting probability distributions. J Am Stat Assoc 100:680–701

Girling AJ, Freeman G, Gordon JP, Poole-Wilson P, Scott DA, Lilford RJ (2007) Modeling payback from research into the efficacy of left-ventricular assist devices as destination therapy. Int J Technol Assess Health Care 23:269–77

Gosling JP, Oakley J, O'Hagan A (2007) Nonparametric elicitation for heavy-tailed prior distributions. Bayesian Anal 2:693–718

Gosling JP, Hart A, Mouat D, Sabirovic M, Scanlon S, Simmons A (2012) Quantifying experts' uncertainty about the future cost of exotic diseases. Risk Anal 32:881–893

Gosling JP, Hart A, Owen H, Davies M, Li J, MacKay C (2013) A Bayes linear approach to weight-of-evidence risk assessment for skin allergy. Bayesian Anal 8:169–186

Higgins H, Dryden I, Green M (2012) A Bayesian elicitation of veterinary beliefs regarding systemic dry cow therapy: variation and importance for clinical trial design. Prev Vet Med 106:87–96

HM Treasury (2015) The Aqua Book: guidance on producing quality analysis for government. https://www.gov.uk/government/publications/the-aqua-book-guidance-on-producing-quality-analysis-for-government. Cited 15 Sept 2016

Iglesias C, Thompson A, Rogowski W, Payne K (2016) Reporting guidelines for the use of expert judgement in model-based economic evaluations. PharmacoEconomics 34(11):1161–1172

Kadane JB (1986) Progress toward a more ethical method for clinical trials. J Med Philos 11:85–404

Kadane JB, Wolfson L (1998). Experiences in elicitation. Statistician 47:3–19

Kennedy M, Anderson C, O'Hagan A, Lomas M, Woodward F, Gosling JP, Heinemeyer A (2008) Quantifying uncertainty in the biospheric carbon flux for England and Wales. J R Stat Soc Ser A 11:109–135

Kennedy M, Clough H, Turner J (2009) Case studies in Bayesian microbial risk assessments. Environ Health 8:S19

Kinnersley N, Day S (2013) Structured approach to the elicitation of expert beliefs for a Bayesian-designed clinical trial: a case study. Pharm Stat 12:104–113

Lark R, Lawley R, Barron A, Aldiss D, Ambrose K, Cooper A, Lee J, Waters C (2015) Uncertainty in mapped geological boundaries held by a national geological survey: eliciting the geologists' tacit error model. Solid Earth 6:727–745

Lee L, Pringle K, Reddington C, Mann G, Stier P, Spracklen D, Pierce J, Carslaw K (2013) The magnitude and causes of uncertainty in global model simulations of cloud condensation nuclei. Atmos Chem Phys 13:8879–8914

Lindley DV, Tversky A, Brown RV (1979) On the reconciliation of probability assessments. J R Stat Soc Ser A 1:146–80

Linstone HA, Turoff M (1975) The Delphi method: techniques and applications. Addison-Wesley, Boston

Meads C, Auguste P, Davenport C, Małysiak S, Sundar S, Kowalska M, Zapalska A, Guest P, Thangaratinam S, Martin-Hirsch P et al (2013) Positron emission tomography/computerised tomography imaging in detecting and managing recurrent cervical cancer: systematic review of evidence, elicitation of subjective probabilities and economic modelling. Health Technol Assess 17:1–323

Morgan M, Henrion M (1990) Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis. Cambridge University Press, New York

Morris D, Oakley J, Crowe J (2014) A web-based tool for eliciting probability distributions from experts. Environ Model Softw 52:1–4

Myers DG, Lamm H (1975) The polarizing effect of group discussion. Am Sci 63:297–303

Oakley J, O'Hagan A (2002) Bayesian inference for the uncertainty distribution of computer model outputs. Biometrika 89:769–84

Oakley JE, O'Hagan A (2007) Uncertainty in prior elicitations: a nonparametric approach. Biometrika 94:427–41

Oakley J, O'Hagan A (2014) SHELF: the Sheffield elicitation framework (version 2.0). http://www.tonyohagan.co.uk/shelf/. Accessed 7 Sept 2016

O'Hagan A (1988) Probability: methods and measurement. Chapman and Hall, London

O'Hagan A (1998) Eliciting expert beliefs in substantial practical applications. Statistician 47:21–35

O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, Oakley JE, Rakow T (2006) Uncertain judgements: eliciting experts' probabilities. Wiley, Chichester

Raiffa H (1968). Decision analysis: introductory lectures on choices under uncertainty. Addison-Wesley, Reading

R Core Team (2016) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna https://www.R-project.org/. Accessed 7 Sept 2016

Ren S, Oakley J (2014) Assurance calculations for planning clinical trials with time-to-event outcomes. Stat Med 33:31–45

Rowe G, Wright G (2001) Expert opinions in forecasting: the role of the Delphi technique. In: Armstrong JS (ed) Principles of forecasting. Springer, New York, p 125–144

Scholten L, Scheidegger A, Reichert P, Maurer M (2013) Combining expert knowledge and local data for improved service life modeling of water supply networks. Environ Model Softw 42:1–16

Sperber D, Mortimer D, Lorgelly P, Berlowitz D (2013) An expert on every street corner? Methods for eliciting distributions in geographically dispersed opinion pools. Value Health 16:434–437

Stone M (1961) The opinion pool. Ann Math Stat 32:1339–1342

Thomas EA, Ross BH (1980) On appropriate procedures for combining probability distributions within the same family. J Math Psychol 21:136–52

Truong P, Heuvelink G, Gosling JP (2013) Web-based tool for expert elicitation of the variogram. Comput Geosci 51:390–399

Usher W, Strachan N (2013) An expert elicitation of climate, energy and economic uncertainties. Energy policy 61:811–821

Winkler RL (1967) The quantification of judgment: Some methodological suggestions. J Am Stat Assoc 62:1105–1120

Zapata-Vázquez R, O'Hagan A, Soares Bastos L (2014) Eliciting expert judgements about a set of proportions. J Appl Stat 41:1919–1933

Chapter 5
IDEA for Uncertainty Quantification

Anca M. Hanea, Mark Burgman, and Victoria Hemming

Abstract It is generally agreed that an elicitation protocol for quantifying uncertainty will always benefit from the involvement of more than one domain expert. The two key mechanisms by which judgements may be pooled across experts are through striving for consensus, via behavioural aggregation, where experts share and discuss information, and via mathematical methods, where judgements are combined using a mechanistic rule. Mixed approaches combine elements of both deliberative (behavioural) and mechanical (mathematical) styles of aggregation.

This chapter outlines a mixed-aggregation protocol called IDEA. It synthesises specific elements from several of the *classical* structured expert judgement approaches. IDEA encourages experts to Investigate, Discuss, and Estimate, and concludes with a mathematical Aggregation of judgements.

5.1 Introduction

Several elicitation protocols developed over the last decades have been deployed successfully in political science, infrastructure planning, volcanology, etc. (e.g. Aspinall 2010; Aspinall and Cooke 2013; Bolger et al. 2014; Cooke and Goossens 2008; O'Hagan et al. 2006). The protocols detailed in Chaps. 2 and 3 of this book (see Quigley et al. 2018 and Gosling 2018 respectively) are two of the most notable examples of structured protocols that follow thoroughly documented methodological rules. They differ in several aspects, including the way interaction between experts is handled, and the way in which experts' judgements are pooled.

The Classical (Cooke's) Model detailed in Chap. 2 of this book (Quigley et al. 2018) uses mathematical aggregation. In mathematical aggregation approaches, interaction between experts is generally limited to training and briefing (e.g. Valverde 2001; Cooke 1991), since it is believed that more interaction may induce

A.M. Hanea (✉) • V. Hemming
CEBRA, University of Melbourne, Parkville, VIC, Australia
e-mail: ahanea@unimelb.edu.au; hemmingv@student.unimelb.edu.au

M. Burgman
Centre for Environmental Policy, Imperial College London, London, UK
e-mail: m.burgman@imperial.ac.uk

L.C. Dias et al. (eds.), *Elicitation*, International Series in Operations Research & Management Science 261, https://doi.org/10.1007/978-3-319-65052-4_5

dependence between elicited judgements (e.g. O'Hagan et al. 2006), adversely affecting them. Chapter 9 of this book (Wilson and Farrow 2018) discusses the aggregation of correlated judgements in detail; here we touch on this subject very briefly.

The main advantage of mathematical aggregation is that it makes aggregation explicit and auditable. The choice of the aggregation rule is nevertheless difficult. Different rules possess different properties and it is not possible to have all desirable properties in one rule (Clemen and Winkler 1999). The Classical Model uses an unequally weighted linear pool, distinguished by the use of calibration variables to derive performance based weights. Techniques for testing and evaluating experts' performances necessarily play an important role in exploring the performance of experts. Commonly used metrics are designed to be objective. However, different metrics focus on (and measure) different attributes of performance.

Another class of methods of aggregating experts' judgements is referred to as *behavioural aggregation*, and involves striving for consensus via deliberation (O'Hagan et al. 2006). The Sheffield protocol, detailed in Chapter 3 of the book (Gosling 2018), is an example. When experts disagree, the advocates of behavioural aggregation recommend a discussion between the experts with divergent opinions, resulting in a "self-weighting" through consensus.[1] But this comes at the cost of verifiability and reproducibility. Moreover, such interaction is prone to group dynamic biases including overconfidence, polarisation of judgements and groupthink (Kerr and Tindale 2011).

Mixed approaches combine behavioural and mathematical aggregation techniques. The most common mixed approach is the Delphi protocol (Rowe and Wright 2001), in which experts receive feedback over successive question rounds through a facilitator, in the form of other group members' judgements. Experts remain anonymous and do not interact with one another directly. As originally conceived, the Delphi method strives to reach consensus after a relatively small number of rounds (Dalkey 1969), though in modern usages achieving consensus is not necessarily the primary aim (e.g. von der Gracht 2012). While research supports a general conclusion that Delphi methods can improve accuracy over successive rounds, this is by no means guaranteed. Critical reviews suggest that even though individual judgements may converge (von der Gracht 2012), this convergence does not necessarily lead to greater accuracy (e.g. Murphy et al. 1998; Bolger et al. 2011). Moreover, the Delphi method is widely used for the elicitation of point estimates rather than probability distributions.

The IDEA protocol described in this chapter synthesizes specific elements from all the approaches described above. In doing so, it aims to minimize the

[1]However, where a group consensus judgement cannot be reached, individual expert distributions can be elicited and combined using a mathematical aggregation technique. Or alternatively, where consensus is not the aim, the resulting spread of expert viewpoints following discussion can be maintained and presented to decision-makers (Morgan 2015).

disadvantages of existing approaches and optimise their advantages. The majority of elements that characterise IDEA are not new; its novel contribution is in the structured approach to the combination of these elements.

The reminder of this chapter is organised as follows: Section 5.2 introduces the IDEA protocol, Sect. 5.3 discusses the analysis of expert data collected using IDEA and Sect. 5.4 offers guidance for facilitators to use IDEA to elicit and quantify uncertainty.

5.2 The IDEA Protocol

The acronym *IDEA* arises from the combination of the key features of the protocol that distinguish it from other structured elicitation procedures: it encourages experts to *Investigate* and estimate individual first round responses, *Discuss*, *Estimate* second round responses, following which judgements are combined using mathematical *Aggregation* (Hanea et al. 2016).

An outline of the basic approach is as follows. First, experts provide private, individual estimates in response to the questions posed to them. They receive feedback in the form of the judgements of the other experts. With the assistance of a facilitator, the experts discuss their initial estimates with the others, sharing information, clarifying terms, and establishing a shared understanding of the problem. This discussion stage may take place remotely (e.g. Wintle et al. 2012; McBride et al. 2012; Hanea et al. 2016) or face-to-face (e.g. Burgman et al. 2011). During the discussion stage, ideally the anonymity of the individual estimates is maintained to counter possible unwanted dominance and halo effects. Experts are asked to revise their judgements in light of this discussion and make a second, private and anonymous estimate. These second round estimates are finally combined mathematically (see Fig. 5.1).

The motivation behind the use of the IDEA protocol is that while interaction between experts can be detrimental during the initial development of arguments and

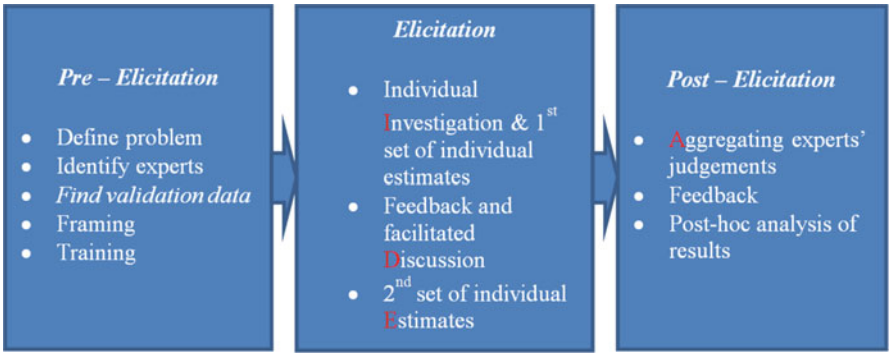


Fig. 5.1 The IDEA protocol

responses, its use during the evaluation stage can be beneficial: allowing experts to better clarify reasoning and assumptions, and to benefit from the gains arising from well-functioning behavioural groups. The controlled interaction and feedback allow for exchange of information independent of its source, thereby removing some of the more negative aspects of behavioural groups. Using a (final) mathematical aggregation lessens the pressure for experts to reach consensus. In making their estimates for each question, experts answer using either a 4-step format for eliciting information about quantities, or a 3-step format for eliciting probabilities of binary variables (Burgman 2016). These formats draw on empirical findings from cognitive psychology and they have been shown to mitigate overconfidence (Speirs-Bridge et al. 2010; Soll and Klayman 2004).

5.2.1 Eliciting Probabilities

When eliciting probabilities of binary variables (or event' occurrences), IDEA uses three questions, termed a 3-step format, one for a *best estimate* and the other two for an interval that captures uncertainty around it. The bounds are asked for before the best estimate, to get experts to think about the extreme conditions. The first two questions are prefaced with statements that urge them to think about evidence that points in one direction, and then the other, as shown in Fig. 5.2.

Other approaches, including Cooke's protocol, ask the experts to assign events to probability bins $b_i = (p_i, 1 - p_i)$, where p_i corresponds to the probability of occurrence. Bins can have the following form: $b_1 = (0.1, 0.9)$, $b_2 = (0.2, 0.8)$, $b_3 = (0.3, 0.7)$, etc. if the continuous probability of occurrence scale is discretized into ten intervals. An expert assigns an event to the b_2 bin if their best estimate (about the probability of occurrence) is anywhere between 0.1 and 0.2. So, in a way, these approaches only ask for *best estimates*, acknowledging the imprecision in the experts' judgements by allowing a fixed interval around them (equal to the respective bin's length).

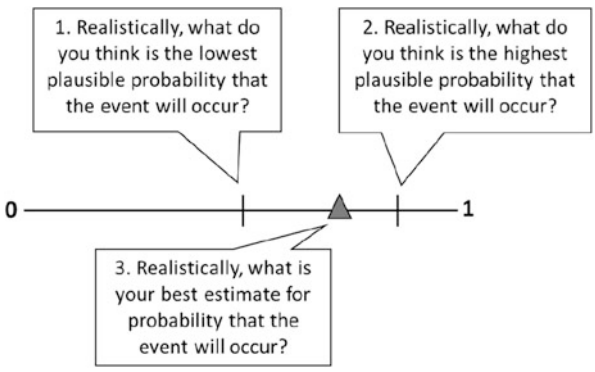


Fig. 5.2 The 3-step format

The probabilities of binary variables can sometimes be interpreted in terms of relative frequencies. It is then legitimate to ask experts to quantify their degree of belief using a subjective distribution. In this case the upper and lower bounds asked for in the 3-step format may be thought of as quantiles of this subjective probability distribution. However, when the relative frequency interpretation is not appropriate the 3-step format may be criticised for lacking operational definitions for the upper and lower bounds. We emphasize that in such cases the bounds are elicited to improve thinking about the best estimates. They are not used in a probabilistic framework.

In both situations, if questions resolve within the time frame of the study, and using the experts' best estimates only, experts' performances can be assessed in terms of accuracy and calibration. For calibration measures, the best estimates are placed in probability bins. For example, using the notation above, best estimates between 0.2 and 0.3 are assigned to bin b_3. This construction allows the evaluation of calibration measures used in other protocols, e.g. Cooke's protocol. Sections 5.3.2 and 5.3.3 discuss a comparison of such measures evaluated using a dataset detailed later in this chapter.

5.2.2 Eliciting Quantiles of Probability Distributions

When IDEA is used to elicit continuous quantities (continuous random variables) this procedure uses four questions to elicit the values of variables (corresponding to different quantiles), termed a 4-step format. This approach draws on research from psychology on the effects of question formats, mitigating much of the overconfidence typically observed in expert estimates (e.g. Soll and Klayman 2004; Speirs-Bridge et al. 2010).

In the 4-step format (like the 3-step format above) bounds are elicited before asking for the best estimate, to encourage experts think about extreme values, and to prevent them from anchoring on their best estimate. The first three questions are used to elicit three values of the variable, corresponding to three different quantiles, and the fourth question is used to identify the probabilities corresponding to the upper and lower quantiles specified by the experts (Fig. 5.3).

The best estimate corresponds to the median.[2] The lower and upper bounds correspond to upper and lower quantiles (denoted q_l and q_u), such that their difference corresponds to the specified confidence level. If, for example, an expert provides a 50% confidence level, q_l and q_u will be taken to be the first and the third quartiles. When experts provide different confidence levels, their estimates are

[2]The best estimate may be also interpreted as the mode of the distribution. Methods for building a distribution that complies with the mode and two specified quantiles are proposed in Salomon (2013). However the interpretation of the best estimate and its use in constructing a distribution should be clearly specified prior to the elicitation.

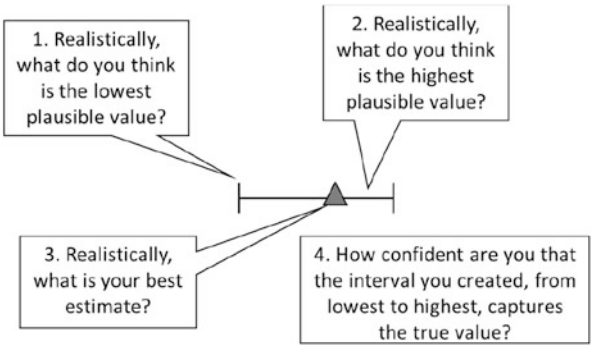


Fig. 5.3 The 4-step format

rescaled to a consistent confidence level (e.g. 90% confidence) such that experts' distributions can be further compared and aggregated. Several methods may be used to rescale to a fixed pair of quantiles, ranging from a simple linear extrapolation to fitting a parametric distribution to the elicited quantiles and extracting the required quantiles from the fit. The sensitivity of an aggregated distribution (calculated for example as a weighted combination of individual rescaled expert distributions) to the choice of the rescaling method is assumed low (as supported by anecdotal evidence). However this topic requires additional research.

A slightly different version of this procedure, where the elicited quantiles are fixed, corresponds to the way questions are asked in the Sheffield method and in Cooke's protocol. Once rescaled to these fixed quantiles, the answers obtained using the 4-step format can be mathematically aggregated using the mathematical apparatus of Cooke's protocol.

5.3 Data Analysis

The IDEA protocol was refined and tested as part of a forecasting "tournament" that started in 2011 as an initiative of the US Intelligence Advanced Research Projects Activity(IARPA).[3] Five university-based research teams were involved in predicting hundreds of geopolitical, economic and military events, with the goal of finding the key characteristics of efficient protocols for eliciting and aggregating accurate probabilistic judgements. The project used real events that resolved in the near-future to test the accuracy of forecasts. Thousands of forecasters made over a million forecasts on hundreds of questions (Ungar et al. 2012; Mellers et al. 2015). The data elicited with the IDEA protocol represent the answers to a subset of the questions developed by IARPA. All questions considered correspond to Bernoulli variables of the following sort: "Will the Turkish government release imprisoned Kurdish

[3] http://www.iarpa.gov/index.php/research-programs/ace.

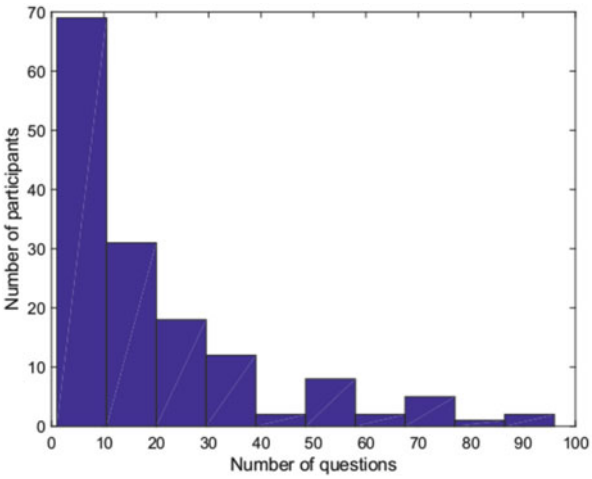


Fig. 5.4 The number of questions answered by participants over 4 years

rebel leader Abdullah Ocalan before 1 April 2013?", which were answered using the 3-step format outlined above. All questions usually resolved within 12 months, hence they were suited for empirical validation studies. The elicitation took place remotely, initially via email, and from the second year of the tournament through a dedicated website[4] which was set up for the participants to answer the questions, discuss and upload/download necessary materials.

The tournament operated on a yearly basis, over the course of 4 years. Each year, new participants joined the IDEA group, and other participants dropped out. There were 150 participants (over the 4 years) who answered at least one question (both rounds). Eight of these participants returned each year. The level of participants' expertise covered a very wide range from self-taught individuals with specialist knowledge to intelligence analyst. A total of 155 questions were answered by at least one participant. However, no participant answered more than 96 questions. Figure 5.4 shows the distribution of the number of questions answered by the participants. The participants were divided into groups and the number of groups varied across years to keep the number of participants per group fairly constant (typically ten). Starting from the third year *Super-groups* were formed composed of the best performing participants from the previous year.[5] The number of participants composing the Super-group was equal to the number of participants from any other group.

Initial training of the participants took place before the game started. Some of the participants engaged in initial face-to-face training, where they learned about how the questions would be asked, why they were asked in this manner, and

[4] http://intelgame.acera.unimelb.edu.au/.

[5] Performance was measured using the average Brier score. This measure was imposed by the forecasting tournament rules and all participating team had to use it.

most importantly, the cognitive biases and group issues that can occur during an elicitation, and ways to mitigate them. Participants who did not receive face-to-face training, received online or telephone training. Training materials/documents that outlined and explained the issues above were also uploaded to the website for access and reference. Even though probabilistic training was not offered, many probabilistic concepts were introduced through practice questions that were part of the training.

5.3.1 Measures of Performance

This section outlines some of the approaches to measuring expert performance and dependencies among experts' estimates that we have investigated for the dataset described above. Hence we restrict attention to evaluating assessments of binary variables. Experts are asked to represent their uncertainty as a subjective probability and their assessments may then be scored. Roughly speaking, a scoring rule is a numerical evaluation of the accuracy of expert assessments against actual outcomes (de Finetti 1962; Savage 1971; Winkler and Jose 2010). Despite the simplicity of this idea, there are many ways to score experts, deserving careful attention. Scoring rules are called *proper* if their expected pay-off is maximised when experts accurately express their true beliefs about the predicted event. Proper scoring rules encourage the experts to make careful and honest assessments (Winkler and Murphy 1968).

Along with evaluating individual experts' performances, we are also interested in experts' joint behaviour. Expert judgements are (in general) correlated with one another, if for no other reason, because people have access to similar information and have similar training and experiences (e.g. Booker and Meyer (1987)). This subject is discussed in Chap. 9 of this book (Wilson and Farrow 2018); here we only present the analysis of the dataset introduced above.

We are concerned with scoring as a way of rewarding those properties of expert subjective probability assessments that we value positively. We have investigated three of these properties: accuracy, calibration and informativeness.

5.3.1.1 Accuracy

Accuracy measures how close an expert's best estimate is to the truth. One tool to measure accuracy is the Brier score (Brier (1950)), a proper scoring rule. The Brier score for events is twice the squared difference between an estimated probability (an expert's best estimate) and the actual outcome; hence it takes values between 0 and 2. Consider question/event i with two possible outcomes j. The Brier score of expert k assessing event i is calculated as follows:

$$BrierScore_i^k = \sum_{j=1}^{2}(p_{ij}^k - x_{ij})^2,$$

where p_{ij}^k is expert k's probability for event i, output j, and x_{ij} is 1 if output j occurs and 0 otherwise. The above formula measures the accuracy of one estimate made by one expert for one question. Lower values are better and can be achieved if an expert assigns large probabilities to events that occur, or small probabilities to events that do not occur. An experts' accuracy can be then measured over many questions (N) and averaged to represent their overall accuracy:

$$BrierScore^k = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{2}(p_{ij}^k - x_{ij})^2$$

The number of questions and their overall sample distribution play an important role in interpreting such a score. By an overall sample distribution, we mean the inherent uncertainty of the events represented by the questions. This is also called the *base rate* and it is different for each different set of questions. However, its value contributes to the value of the average Brier score, even though it has nothing to do with the expert's accuracy. This challenges the comparison of experts' scores calculated from different sets of questions, with different base rates. Nevertheless, comparisons will be more meaningful when made on the same set of questions.

5.3.1.2 Calibration

To deal with the *base rate* problem, Cooke discusses the benefits of using scores for average probabilities, rather than average scores for individual questions (variables) in Cooke (1991). He opts for calibration (which he calls *statistical accuracy*) rather than accuracy measures for evaluating experts' performance. A scoring rule is essentially a random variable and interpreting the scores' values requires knowledge about the score's distribution. An important justification for Cooke's proposal is that his (asymptotically proper) score has a known distribution, as opposed to (for example) the average Brier score, which does not. The average Brier score is a single number summary of the joint distribution of forecasts and observations. An empirical distribution of the average Brier score can be obtained for a given joint distribution of the forecasts and observations. However, this empirical distribution will differ for different joint distributions.

Before introducing Cooke's calibration score for events,[6] we need some notation. Assume the experts are asked to assign events to probability bins b_i. Let p_i be the

[6]The calibration measure for events is based on similar concepts as the ones presented in Chap. 2 of this book (Quigley et al. 2018), when the calibration score is described for evaluating assessments about continuous variables.

probability of occurrence that corresponds to bin b_i. Each expert assigns events to bins. Let n_i denote the number of events assigned (by an expert) to the bin b_i. Let s_i denote the proportion of these events that actually occur; s_i can be thought of as the empirical distribution of b_i, whose theoretical distribution is p_i. Ideally s_i and p_i should coincide. Nevertheless, in practice, they often do not. Cooke's calibration is essentially a comparison between the empirical and theoretical distributions, per bin, per expert. The discrepancy between the two is measured in terms of the relative information[7] $I(s_i, p_i)$ of s_i with respect to p_i, defined in Chap. 2 of this book (Quigley et al. 2018). The relative information of one distribution with respect to another is a non-negative measure that equals zero iff $s_i = p_i$. Increasing values of $I(s_i, p_i)$ indicate increasing discrepancy. The relative information is calculated as follows:

$$I(s_i, p_i) = s_i \ln\left(\frac{s_i}{p_i}\right) + (1 - s_i) \ln\left(\frac{1 - s_i}{1 - p_i}\right)$$

A result in Hoel (1971) shows that for n_i independent events whose probability of occurrence is p_i, $2n_iI(s_i, p_i)$ is asymptomatically Chi-squared distributed with one degree of freedom. Then, if ten bins are used and if all events are independent $\sum_{i=1}^{10} 2n_iI(s_i, p_i)$ is asymptomatically Chi-squared distributed with ten degrees of freedom. Under the (null) hypothesis that the experts estimate the theoretical distribution correctly, Cooke's calibration is defined as the probability of obtaining a result equal to or more extreme than the one observed. Hence, it corresponds to the p-value of a statistical test:

$$Cal(e) = 1 - \chi^2_{10}\left(\sum_{i=1}^{10} 2n_iI(s_i, p_i)\right),$$

where χ^2_{10} is the cumulative distribution function of a Chi-squared random variable with ten degrees of freedom.

For the Chi-square approximation to be reasonably close, the number of questions assessed by each expert should be quite large (hundreds). Since this is very rarely the case in practice, the empirical distribution of $\sum_{i=1}^{10} 2n_iI(s_i, p_i)$ (obtained via simulation) is used instead.

As for the average Brier score case, ideally expert performances should be compared on the same set of questions. When experts assess different questions, the power of the test used in measuring calibration should be adjusted to account for the different number of samples (the different number of questions) (Cooke 1991). Incorporating this adjustment into the simulated empirical distribution of the score is far from trivial. If a score has an exact distribution, rather than an asymptotic one, the power adjustment is not crucial.

[7]The relative information is usually known as the Kullback–Leibler divergence, or information divergence, or information gain, or relative entropy.

Using the same notation we could measure a different sort of calibration through the average Brier score discussed above. The average Brier score can be decomposed into two additive components called *calibration* and *refinement* (Murphy 1973). The calibration term for N questions can be calculated as follows:

$$\sum_{i=1}^{10} \frac{n_i(p_i - s_i)^2}{N}$$

Very roughly, the refinement term is an aggregation of the resolution and the inherent uncertainty of the events assessed. The resolution term rewards expert estimates that are consistent with event probabilities. Other measures of resolution based on the notion of entropy associated with a probability mass function can be formulated. Entropy is a measure of the degree to which the mass is *spread out* and can be used in several ways to describe aspects of an expert's informativeness.

5.3.1.3 Informativeness

Entropy is very often taken as a measure of lack of information in a distribution. The entropy of the distribution $(p_i, 1 - p_i)$, denoted $H(p_i)$ is calculated as follows:

$$H(p_i) = -p_i ln(p_i) - (1 - p_i) ln(1 - p_i)$$

The maximum value of $H(p_i)$ is $ln(2)$ and it is obtained when $p_i = 0.5$. Thus, the uniform distribution is the most entropic. The most informative distribution corresponds to the distributions with minimal entropy, 0. This is obtained only if $p_i = 0$ or $p_i = 1$. The entropy in the joint distribution of independent variables is the sum of entropies in the distributions of the individual variables. Two different entropy measures are defined in Cooke (1991), the average *response entropy* and the average *sample entropy*. The average response entropy in an expert's joint distribution on N events is defined as:

$$H_r = \frac{1}{N} \sum_{i=1}^{10} n_i H(p_i)$$

The response entropy measures the entropy in what the expert says. It does not depend on the actual occurrences of events. The average sample entropy, denoted H_s, is calculated as follows:

$$H_s = \frac{1}{N} \sum_{i=1}^{10} n_i H(s_i)$$

The sample entropy measures the entropy in the expert's performance, but it does not correspond to the distribution that the expert (or anyone else) believes. In contrast, response entropy corresponds to the distribution connected to the calibration hypothesis described above. If an expert is perfectly calibrated, then $H_s = H_r$. Unfortunately, $H_s = H_r$ does not imply perfect calibration.

An expert's informativeness may be also measured with respect to their choice of the probability bins. The choice (alone) of a more extreme probability bin (i.e. assigning a probability close to 0 or 1) can give yet another indication of the expert's informativeness. The average *response informativeness*, introduced in Hanea et al. (2016) is defined as follows:

$$I_r = \frac{1}{N} \sum_{i=1}^{10} n_i I(p_i, 0.5)$$

The response informativeness attains its minimum in 0, when all the variables are placed in the $(0.5, 0.5)$ bin. A higher informativeness score is preferred since it indicates that more variables were placed in more extreme bins.

All the formulations above assume that experts have placed events in probability bins. However IDEA asks experts to provide a best estimate and an uncertainty interval around their best estimate. In our analysis, the above measures are calculated by placing the best estimates into the bins and ignoring the upper and lower bounds. Nevertheless, the *interval' widths* can be considered as a measure of the experts' confidence, or lack thereof. A larger (smaller) interval may be interpreted as decreased (increased) confidence. Narrower bounds around a judgement are often interpreted as greater informativeness. Hence we can investigate the length of the uncertainty interval as a measure of confidence and the relationship between this measure and the measures of informativeness discussed above. These relationships are investigated in Hanea et al. (2016) for the dataset described above.

5.3.1.4 Correlated Expert Judgements

Correlated expert judgements have been discussed occasionally in the literature but, to our knowledge, there has been little research on evaluating the extent to which this dependence is practically relevant. Cooke (1991) postulates that such correlation is:

> usually benign, and always unavoidable.

In contrast O'Hagan et al. (2006) worries that:

> groups of similar experts will receive too much weight and minority views will be under-represented.

Chapter 9 Wilson and Farrow (2018) of the book discusses this subject from a more general perspective. In contrast, we investigate only two particular conjectures about the dependence between the participants' answers elicited using the IDEA protocol (which permits and encourages interaction between the two

elicitation rounds). We conjecture that any additional dependence between judgements introduced through the discussion is justified by the increase in information resulting from discussion and by the reduction of misunderstandings or unintended dichotomies in responses. Moreover, this discussion takes place within groups, so our second conjecture is that the dependence structures within and between the groups are similar. If/when that is true, the expert data analysis can be (statistically) strengthened by pooling the estimates from all groups.

5.3.2 The Merits of Discussion

Results on the benefits of the discussion between rounds, based on part of the 4 year dataset described earlier are presented in Hanea et al. (2016). The analysis was undertaken within groups and per year, hence the claimed benefits lack statistical power. However, the second conjecture formulated above is supported by the data analysis from Hanea et al. (2016), so we feel comfortable in pooling the expert data to form a larger dataset and hence permit more powerful statistical tests. This allows us to investigate how some of the performance scores detailed in Sect. 5.3.1 change per expert after discussion. Figure 5.5 shows pairs of four different scores (before and after discussion) corresponding to all participants who answered at least four questions. The crosses represent the average Brier scores, the diamonds represent the average confidence as measured by the length of the uncertainty intervals, and the x's represent the calibration terms of the Brier score. For all three measures low scores represent better performance. The dots represent the average response informativeness; better informativeness corresponds to larger values. The main diagonal is plotted for better visualisation. For the first three measures (Brier scores, confidence, and calibration), most of the points fall below the main diagonal, indicating better performance after discussion. For the fourth measure (informativeness), most of the points fall above the main diagonal, again indicating better performance in the second round.[8]

All the investigated measures of performance point to the value of facilitated conversations between experts in reconciling language based misunderstandings and interpretations of evidence. The relationship between these measures remains unclear in general. For this particular dataset, the authors of Hanea et al. (2017) found no, or little correlation between how accurate experts' estimates are, and how informative they are.

[8]Three quarters of the Brier scores and the average confidence scores are better in the second round, and two thirds of the calibration scores and the informativeness scores are better in the second round.

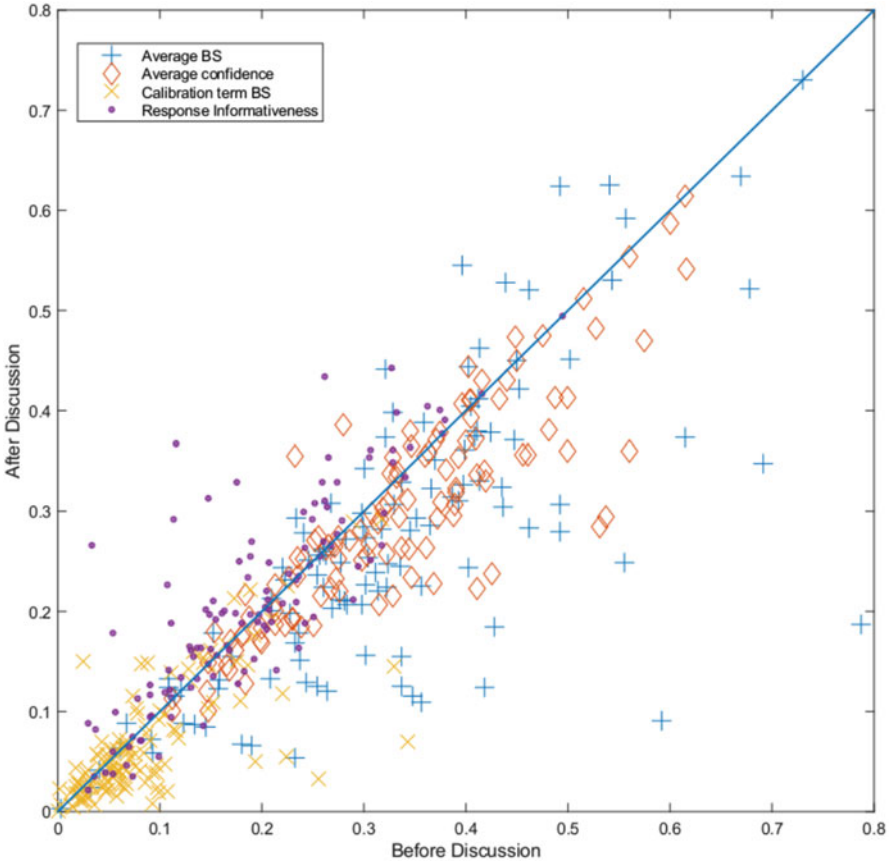


Fig. 5.5 The average Brier scores, the average confidence,the calibration term of the Brier score and the average response informativeness of all participants, before and after discussion

5.3.3 Prior Performance as a Guide to Future Performance

Each year of the tournament, we compared an equally weighted combination of all participants after the first round of opinions ("the wisdom of crowds") to the equally weighted opinions of the groups after discussion, using a within-subject design. In 1 year alone (2013–2014) we had sufficient data to calculate differential weights using Cooke's calibration score. Figure 5.6 shows the average Brier scores of the equally weighted combination of all participants' first round judgements (before discussion), compared with the equally weighted judgements of each group after discussion, together with their corresponding confidence intervals. An unequal, performance-based weighted combination of the super-group participants' judgements is shown in the same figure.

Although not statistically significant, the super-group (G1) outperformed the other groups of participants, suggesting that prior performance is a useful guide to

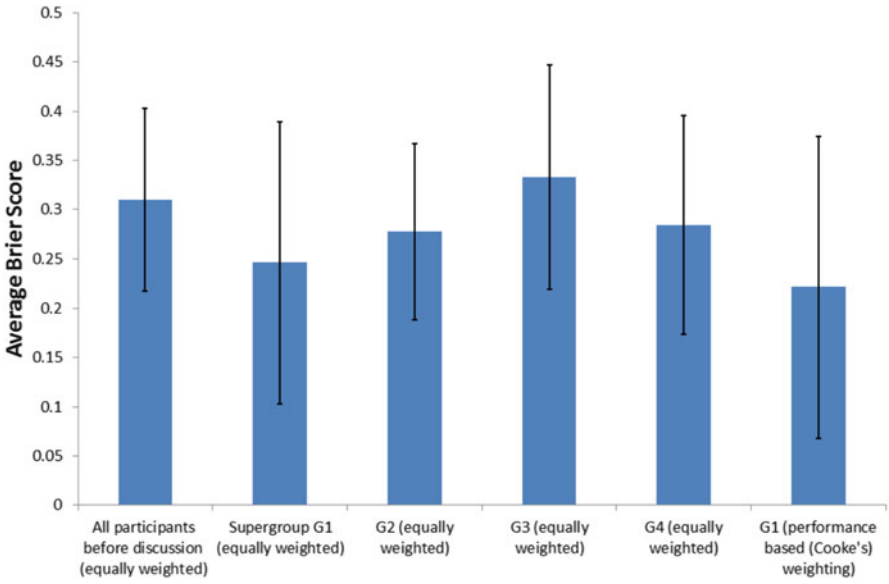


Fig. 5.6 Forcasting tournament, year 2013–2014

future performance on similar estimation tasks. The same was observed for the forth year of the tournament. This finding is in agreement with the findings of Mellers et al. (2014). Using Cooke's calibration to derive performance based weights for an unequally weighted combination generates a slight improvement in performance.

These signals illustrate one of the most important lessons of empirical studies over the last decade: an expert's performance on technical questions may be predicted to some extent by the history of their performance on similar questions previously. Taking advantage of this phenomenon, Cooke's approach to differential weighting assimilates each expert's confidence and statistical accuracy into a single weight. The result is that group performance improves. Our results demonstrate that even in the relatively difficult conditions imposed in answering binary questions on the outcomes of geopolitical events, performance based differential weights calculated using Cooke's method improve the performance of groups, even those composed of relatively reliable forecasters.

5.4 A Guide to Facilitating the IDEA Elicitation Protocol

The purpose of this section is to present a summary guide for analysts and facilitators who intend to use the IDEA protocol in an uncertainty quantification exercise. Some of the recommended steps are similar to those needed when using other protocols, however, several are specific to IDEA. Much of this section has

been adapted from Hemming et al. (2017), and we suggest referring to this paper for more comprehensive advice and examples. In this section, we assume that the problem structuring, modelling, identification of data gaps and the requirements for expert input have been decided upon.

5.4.1 Preparing for an Elicitation

Careful planning is necessary to ensure that experts are aware of time constraints, and that the deliverables of the elicitation become available in the time necessary. Below we briefly discuss a number of key elements to be taken into account prior to the elicitation.

5.4.1.1 Key Documents

Time-Line and Key Dates
A list of tasks and a schedule of key dates for each of the steps of the elicitation before commencing the process is necessary. An elicitation using the IDEA protocol can take up to 6 weeks if using remote elicitation, or as little as 3 days if using a face-to-face elicitation. Additional time is required for the development of questions, recruitment of experts, approval of human research approvals, and the analysis of data. A sample timeline can be found in the supplementary material of Hemming et al. (2017).

Human Subjects Research Ethics Approvals
These approvals may be required, particularly if results are to be published, or to be used to inform decisions. If approval is necessary this may substantially delay the project.

A Project Description
This document outlines the purpose of the project, the relevant time-frames, the required expert input, and any payments. It also includes instructions on how the collected data will be used.

A Consent Form
A consent form should accompany the project description and be sent to experts to formalize their agreement to take part in the study and for the data to be retained and used for the specified purpose.

Briefing Document
The purpose of this document is to guide experts through the IDEA elicitation protocol. It should include instructions on how to answer the questions, reiterate that experts must make an initial private and anonymous estimate, whilst they are free

to talk to people outside of the elicitation group, they cannot discuss their estimates with anyone inside the group until the discussion round. Instructions should also explain the four-step or three-step format, and how their estimates will be interpreted or scored. The document should re-iterate the time-lines for the elicitation.

5.4.1.2 The Questions

Even when the quantities to be elicited are identified, the elicitation questions should be framed such that the quantities to be elicited relate to potentially verifiable facts and have a clear operational meaning. Moreover, the questions should include details such as units, time-scales, and metrics. Vague, ambiguous or underspecified questions which could result in multiple interpretations should be avoided.

Ideally, one or two experts who will not participate in the elicitation should scrutinize the draft questions, ensuring (as far as possible) that the questions are fair and reasonable, within the domain of expertise of the participants, free from linguistic ambiguity or biases, and they can be completed within the allocated time-frame. The total number of questions that can be asked during an elicitation depends on the availability and the motivation of the experts. It also depends on the type (remote or face-to-face) and time-frame of the elicitation exercise. The authors of Hemming et al. (2017) suggest that no more than 20 questions should be asked within a single day of elicitation; many more can be asked if more time is available or through remote elicitation, but asking more questions may come at the cost of expert fatigue. Different settings will be detailed later in this section. When experts' judgements are aggregated using differential weighting schemes, calibration questions should be added to the set of questions.

5.4.1.3 The Experts

Chapter 16 of this book (Bolger 2018) is dedicated to expert selection. We only very briefly touch upon this subject. The IDEA protocol relies on recruiting a diversity of experts. To generate a diverse group of experts, we recommend employing a range of techniques including professional network searches, peer-recommendations, on-line searches, and literature reviews. The techniques employed can have inherent biases and lead to the selection of older, well regarded individuals, or people whose ideas are in line with popular belief (often older males with a tertiary education). This may lead to a homogeneous and systematically biased group. Diversity should be reflected by variation within the group in age, gender, cultural background, life experience, education or specialisation, years of experience and position on the questions at hand.

5.4.1.4 The Facilitator

A key requirement of a good facilitator regardless of the protocol they employ is that are neutral to the outcome of the elicitation, and capable of retaining objectivity. The facilitator must be competent in diplomatically handling a wide range of personalities, be able to encourage critical thinking within groups, and to pose counterfactuals.

When facilitating an elicitation using the IDEA protocol, the facilitator should be familiar with the aims and limitations of the IDEA protocol. This means they should be acutely aware of the various biases and heuristics common to expert judgement, and how elements of the IDEA protocol aim to counteract the expression of these biases. The facilitator should understand and be capable of explaining both the mathematical and the psychological theory behind the specific elicitation type and the aggregation method.

5.4.2 Implementing the IDEA Protocol

5.4.2.1 The Initial Meeting

The IDEA protocol commences with an initial meeting between the project team and the experts. The first project meeting is vitally important for establishing a rapport with the experts. A teleconference of approximately an hour is usually sufficient. During the meeting, the motivation for the project is introduced and the unavoidable frailties of expert judgement are explained. The motivation for a structured protocol is the desire to ensure the same level of scrutiny and neutrality is applied to expert judgement as is afforded to the collection of empirical data.

During this meeting the outline of the IDEA protocol, and the motivation behind its key steps are discussed. The format of the questions, the cognitive biases and group issues that can occur during an elicitation, and ways to mitigate them are explained. Probabilistic training may be included if experts do not have a minimum level of understanding of necessary probabilistic concepts. One rule is emphasised: the experts must not speak to one another prior to the discussion stage within the IDEA framework. However, they can and should speak to anyone else they like, and use any sources that may be relevant. We recommend going through one or two practice questions if time allows, as they help the experts familiarise themselves with the questions style and the overall process; otherwise practise questions can be incorporated subsequently. Finally, reiterate the time-lines and allow sufficient time for experts to ask questions. The supplementary material of Hemming et al. (2017) provides an example of how the project team might structure the teleconference.

5.4.2.2 The Elicitation

The IDEA protocol provides a flexible approach to the elicitation of experts which enables on-line and remote elicitation, or to undertake the entire elicitation through a workshop (face-to-face). The choice of method will usually be a result of budget and time constraints, however, if the option is available then it is recommended that at the very least the discussion phase should be undertaken with use of a face-to-face elicitation.

IDEA On-line
The experts should be (individually) provided with the questions (including practice questions if they were not dealt with during the initial meeting), a briefing document to guide them through the elicitation process and to reiterate key steps, and training materials. The experts then create a unique codename/number which retains their anonymity in group discussions, but allows them to easily identify their own estimates. They should be sent a reminder about 3 days before the close of the first round to get their results in by the deadline. Ideally, allow 2 weeks for experts to complete the first round estimates.

Each expert sources information and consults colleagues independently, before answering the questions. Once all answers are collected, allow time for the expert data to be cleaned. If outliers or implausible values are revealed during this process, then it is best to clarify with experts whether these are true beliefs or mistakes before analysing the data.

After all the above steps are completed, a graphical output of the data should be collated and circulated among the experts. Compile the comments, rationales, re sources and links provided by the experts together with their estimates and distribute them together with the graphical output.

The discussion phase commences once experts have received the consolidated results of the first round estimates. This can be undertaken by email, a teleconference, or a web forum. The key aims of discussion are (1) to reduce linguistic uncertainty and (2) to make sure that experts have considered counter-factual explanations, contrary evidence and alternative models. The role of the facilitator is to guide and stimulate discussion but not dominate it. For example, the facilitator should pick some contrasting results and ask questions which help to determine the source of variation.

Following the discussion, facilitators should clarify meaning and/or better define the questions. If questions are reformulated or modified in any way, the new versions should be sent back to the experts, who now need to make second, anonymous and independent estimates for each. Another week or two should be allowed for the second round estimates. It is possible to ask many more questions, when elicitations run remotely (over the web or by email). People then have enough time to spread the tasks over several days.

IDEA Face-to-Face
Face-to-face workshops are time consuming and expensive, but they usually result in better buy-in and acceptance of the outcomes than do elicitations that are exclusively

remote. The duration of the workshop depends on the resources: it can range from 1 day to 3 days. If time allows the initial meeting can be part of the workshop, prior to training the experts, and discussing the questions to be elicited. Experts provide individual, anonymous initial estimates based on their prior knowledge and any information they can gather from the web or other immediately accessible sources.

A graphical output of the data is then collated and presented to the experts. The discussion stage starts and questions are analysed in turn. Typically, some questions are more problematic than others and require longer discussion. As above, the facilitator prompts the experts to think about alternative explanations and to reconcile different linguistic interpretations of the questions. The facilitator judges when the discussion has reached a point when no more useful contributions remain to be made and the questions are sufficiently clarified. The experts then make their second, anonymous and independent estimates for each question.

Hybrid On-line and Face-to-Face IDEA

Combining remote and face-to-face elicitation steps is also possible, and several options are available. A recommended combination (in case of restricted resources) is to elicit the first round estimates remotely, and then conduct face-to-face discussions and elicitation of the second round estimates during a 1 day workshop. Other combinations are nevertheless possible. Chapter 17 of this book Barons et al. (2018) presents an application of the IDEA protocol, where a 1 day face-to-face workshop was used to elicit the questions of interest, followed by a remote IDEA protocol for eliciting calibration questions.

5.5 Discussion

Expert judgements are part of the fabric through which scientists communicate with policy makers and decision makers. In most circumstances, the data we require for decisions are unavailable or incomplete. Expert judgements are an unavoidable part of every-day decision-making in all technical domains. Structured techniques such as those outlined here (and in the rest of the book) are perhaps surprisingly a relatively new initiative. A handful of publications in the early 1990s have been followed by a flowering of ideas, methods and empirical tests in the 2000s. Despite these developments, for the most part, scientists and decision makers alike have been satisfied with informal deliberation processes and ad-hoc methods for acquiring and combining opinions. Evidence accumulated since the 1950s in cognitive psychology especially has illuminated how subjective and unstructured deliberations are prey to a host of frailties that may substantially influence scientific estimates. Most worryingly, the scientist themselves will be unaware of these biases. Thus, these methods represent a critical advance in the place of science in decision making and policy development.

Here, we have outlined the IDEA protocol for structured expert judgement that takes several of the most promising elements of these emerging techniques,

combining them in a way that takes advantage of their strengths, and avoiding their potential weaknesses. The data presented here suggest that some of the potential flaws of this new combined approach are not serious impediments to its deployment. In particular, the potential for generating unwanted correlation structures seems to be outweighed by the improvement in the quality of individual estimates, and subsequently (aggregated) in group judgements.

We have also discussed some of the practical aspects of involving small groups in the process, face-to-face and remotely. This is especially important for the adoption of protocols by organisations such as regulatory agencies and businesses. Often, there is a need to acquire the *best possible* or *best available* expert opinion. Previously, this has been achieved by organisations going to the most highly regarded individual they can find, and using their opinion uncritically. Structured techniques outperform individuals of any status consistently and by a considerable margin. Thus, by using these techniques, organisations may discharge due diligence in decision making. The methodological details provided here ensure that their deployment can be practical and time-efficient.

References

Aspinall W (2010) A route to more tractable expert advice. Nature 463:294–295

Aspinall W, Cooke R (2013) Quantifying scientific uncertainty from expert judgement elicitation. In: Rougier J, Sparks S, Hill L (eds) Risk and uncertainty assessment for natural hazards, chap 10. Cambridge University Press, Cambridge, pp 64–99

Barons M, Wright S, Smith J (2018) Eliciting probabilistic judgements for integrating decision support systems. In: Dias L, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgment, chap 17. Springer, New York

Bolger F (2018) The selection of experts for (probabilistic) expert knowledge elicitation. In: Dias L, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgment, chap 16. Springer, New York

Bolger F, Stranieri A, Wright G, Yearwood J (2011) Does the Delphi process lead to increased accuracy in group-based judgmental forecasts or does it simply induce consensus amongst judgmental forecasters? Technol Forecast Soc Chang 78(9):1671–1680

Bolger F, Hanea A, O´ Hagan A, Mosbach-Schulz O, Oakley J, Rowe G, Wenholt M (2014) Guidance on expert knowledge elicitation in food and feed safety risk assessment. EFSA J 12(6):3745

Booker JM, Meyer M (1987) Sources of correlation between experts: empirical results from two extremes. Los Alamos National Lab., NM (USA); Nuclear Regulatory Commission, Washington, DC (USA). Office of Nuclear Regulatory Research

Brier G (1950) Verification of forecasts expressed in terms of probability. Mon Weather Rev 78:1–3

Burgman M (2016) Trusting judgements: how to get the best out of experts. Cambridge University Press, Cambridge

Burgman M, Carr A, Godden L, Gregory R, McBride M, Flander L, Maguire L (2011) Redefining expertise and improving ecological judgment. Conserv Lett 4:81–87. doi:10.1111/j.1755-263X.2011.00165.x

Clemen R, Winkler R (1999) Combining probability distributions from experts in risk analysis. Risk Anal 19:187–203

Cooke R (1991) Experts in uncertainty: opinion and subjective probability in science. Environmental ethics and science policy series. Oxford University Press, Oxford

Cooke R, Goossens L (2008) TU Delft expert judgment data base. Reliab Eng Syst Saf 93(5): 657–674

Dalkey N (1969) An experimental study of group opinion: the Delphi method. Futures 1:408–426

de Finetti B (1962) Does it make sense to speak of 'good probability appraisers'? In: Good, J. (ed) The scientist speculates: an anthology of partly baked ideas. Basic Books, New York, pp 357–363

Gosling, J (2018) SHELF: the Sheffield elicitation framework. In: Dias L, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgment, chap 4. Springer, New York

Hanea A, Burgman M, McBride M, Wintle B (2017) The Value of performance weights and discussion in aggregated expert judgements. Risk Anal (re-submitted June 2017)

Hanea A, McBride M, Burgman M, Wintle B (2016) Classical meets modern in the IDEA protocol for structured expert judgement. J Risk Res doi:10.1080/13669877.2016.1215346 (Available online 9 Aug)

Hanea A, McBride M, Burgman M, Wintle B, Fidler F, Flander L, Manning B, Mascaro S (2016) Investigate discuss estimate aggregate for structured expert judgement. Int J Forecast doi:10.1080/13669877.2016.1215346 (Available online 8 June)

Hemming V, Burgman M, Hanea A, McBride M, Wintle B (2017) Preparing and implementing a structured expert elicitation using the IDEA protocol. Methods Ecol Evol, Accepted on 20.07.2017

Hoel P (1971) Introduction to mathematical statistics. Wiley, New York

Kerr N, Tindale R (2011) Group-based forecasting?: a social psychological analysis. Int J Forecast 27:14–40

McBride M, Garnett S, Szabo J, Burbidge A, Butchart S, Christidis L, Dutson G, Ford H, Loyn R, Watson DM, Burgman M (2012) Structured elicitation of expert judgments for threatened species assessment: a case study on a continental scale using email. Methods Ecol. Evol. 3: 906–920

Mellers B, Ungar L, Baron J, Ramos J, Gurcay B, Fincher K, Tetlock P (2014) Psychological strategies for winning a geopolitical forecasting tournament. Psychol Sci 25(4):1106–1115

Mellers B, Stone E, Atanasov P, Rohrbaugh N, Metz S, Ungar L, Bishop M, Horowitz M, Merkle E, Tetlock P (2015) The psychology of intelligence analysis: drivers of prediction accuracy in world politics. J Exp Psychol Appl 21:1–14

Morgan MG (2015) Our knowledge of the world is often not simple: policymakers should not duck that fact, but should deal with it. Risk Anal 35:19–20. doi:10.1111/risa.12306

Murphy A (1973) A new vector partition of the probability score. J Appl Meteorol 12(4):595–600

Murphy M, Black N, Lamping D, Mckee C, Sanderson C (1998) Consensus development methods and their use in clinical guideline development. Health Technol Assess 2(3):1–88

O'Hagan A, Buck C, Daneshkhah A, Eiser J, Garthwaite P, Jenkinson D, Oakley J, Rakow T (2006) Uncertain judgements: eliciting experts' probabilities. Wiley, London

Quigley J, Colson A, Aspinall W, Cooke R (2018) Elicitation in the classical method. In Dias L, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgment, chap 2. Springer, New York

Rowe G, Wright G (2001) Expert opinions in forecasting: the role of the Delphi technique. In: Principles of forecasting: a handbook for researchers and practitioners. Kluwer Academic Publishers, Norwell, pp. 125–144

Salomon Y (2013) Unimodal density estimation with applications in expert elicitation and decision making under uncertainty. Ph.D. thesis, Department of Mathematics and Statistics, The University of Melbourne

Savage L (1971) Elicitation of personal probabilities and expectations. J Am Stat Assoc 66: 783–801

Soll J, Klayman J (2004) Overconfidence in interval estimates. J Exp Psychol Learn Mem Cogn 30:299–314

Speirs-Bridge A, Fidler F, McBride M, Flander L, Cumming G, Burgman M (2010) Reducing overconfidence in the interval judgments of experts. Risk Anal 30:512–523

Ungar L, Mellers B, Satopaa 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. In: AAAI fall symposium series. (AAAI Technical Report FS-12-06)

Valverde L (2001) Expert judgment resolution in technically-intensive policy disputes. In: Assessment and management of environmental risks. Kluwer Academic Publishers, Norwell, pp 221–238

von der Gracht H (2012) Consensus measurement in Delphi studies: review and implications for future quality assurance. Tech Forcasting Soc Chang 79:1525–1536

Wilson K, Farrow M Combining judgements from correlated experts. In: Dias L, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgment, chap 9. Springer, New York (2018)

Winkler R, Jose V (2010) Scoring rules. In: Wiley encyclopedia of operations research and management science. Wiley, New York

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

Wintle B, Mascaro M, Fidler F, McBride M, Burgman M, Flander L, Saw G, Twardy C, Lyon A, Manning B (2012) The intelligence game: assessing Delphi groups and structured question formats. In: Proceedings of the 5th Australian security and intelligence conference

Chapter 6
Elicitation and Calibration: A Bayesian Perspective

David Hartley and Simon French

Abstract There are relatively few published perspectives on processes and procedures for organising the elicitation, aggregation and documentation of expert judgement studies. The few that exist emphasise different aggregation models, but none build a full Bayesian model to combine the judgements of multiple experts into the posterior distribution for a decision maker. Historically, Bayesian concepts have identified issues with current modelling approaches to aggregation, but have led to models that are difficult to implement. Recently Bayesian models have started to become more tractable, so it is timely to reflect on elicitation processes that enable the model to be applied. That is our purpose in this Chapter. In particular, the European Food Safety Authority have provided the most detailed and thorough prescription of the procedures and processes needed to conduct an expert judgement study. We critically review this from a Bayesian perspective, asking how it might need modifying if Bayesian models are included to analyse and aggregate the expert judgements.

6.1 Introduction

Proposals for the use of expert judgement to provide inputs to decision analysis when there is little relevant data are almost as old as decision theory itself. Early mathematical formalisms for combining judgements from several experts were either simple averaging, known as opinion polling in this context, or essentially Bayesian (French 1985, 2011). However, the Bayesian approach,[1] although very helpful for thinking about the principles behind the use of structured expert judgement (SEJ), did not prove tractable and in many ways fell into the background of the subject. In practice, opinion polling, particularly Cooke's development known

[1]Introductions to Bayesian approaches to statistics, risk and decision analyses may be found in Smith (2010).

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

L.C. Dias et al. (eds.), *Elicitation*, International Series in Operations Research & Management Science 261, https://doi.org/10.1007/978-3-319-65052-4_6

as the Classical Model (Cooke 1991, 2007), dominated among the mathematical approaches to eliciting and aggregating expert judgement data. Many practical studies also used more behavioural approaches to combining experts' assessments in which groups of experts discussed and agreed on consensus probability distributions (Garthwaite et al. 2005).

We believe that it is timely to reconsider the process of incorporating expert judgement data into risk and decision analyses for two reasons. Firstly, Bayesian methods are more tractable with the advent of more effective computational approaches, particularly MCMC (Wiper and French 1995; Clemen and Lichtendahl 2002; Lichtendahl 2005; Albert et al. 2012). Secondly, the use of Bayesian methods to think conceptually about the principles behind the use of expert judgement has been less common among those developing practical prescriptions and procedures. Our aim in this chapter is to take the latter direction, though we shall to a lesser extent note some of the more technical advances in the application of Bayesian approaches.

Although the last two decades have seen many applications, there are relatively few published perspectives on formal processes and procedures regarding the elicitation, aggregation and documentation of Expert Judgement Studies. Three key texts provide the broadest overviews: Meyer and Booker (1991), Cooke and Goossens (2000) and EFSA (2014). Meyer and Booker's text and the European Food Safety Authority's (ESFA) Guidance provide the most comprehensive discussion and recommendations on process. The Classical Method described in Chap. 2 of this book (Quigley et al. 2017), on which Cooke and Goossens' recommendations are based, has had the most practical usage, but fewer procedural details are specified. Procedural issues are also carefully managed and documented in using the Sheffield Method described in Chap. 4 of this book (Gosling 2017) of behavioural aggregation, developed by O'Hagan and Oakley[2]: see the EFSA guidance for details. None of these recommendations consider how one might use a full Bayesian model to combine the judgements of multiple experts into the posterior distribution for the decision makers from a procedural perspective. That is our purpose in this chapter.

6.2 Context

French in 1985 distinguished three contexts in which one might wish to combine expert judgements of uncertainty.

1. *The Expert Problem.* Here the group of experts are asked for advice by a decision maker who faces a real decision. She[3] formulates the problem and asks the

[2] http://www.tonyohagan.co.uk/shelf/.

[3] Following convention, we will refer to the decision maker as female.

experts for advice on the uncertainties relating to events and quantities that she has defined. In this context the emphasis is on how she should learn from the experts' statements on the uncertainties involved. Ideally, of course, she would want to learn in a rational way.

2. *The Group Decision Problem.* Here there is still the focus of a real decision; but it is the group itself that is jointly responsible and accountable for the decision. The decision makers are their own experts. They formulate the problem, defining events and quantities of interest. During discussion the decision makers may, usually will,[4] share and discuss their uncertainties and they will wish to learn from each other and to do so rationally. However, their responsibility for making the decision may bring democratic issues into play and, sadly, principles of rationality and democracy are often in conflict (Taylor 1995; French 2007, 2011). Moreover, the responsibilities and possible benefits or harm that may come to them individually from the outcome of the decision may consciously or unconsciously bias their statements.
3. *The Textbook Problem.* The group may simply be required to give their judgements for others to use in the future in as yet undefined circumstances; there is no predefined decision. For instance, when an issue or general risk is of current public concern, governments may commission reports from a panel of experts. In this case the experts have to identify events and quantities that *might* be of future interest, but without having the clear focus of a precise decision problem. The textbook problem has been discussed little in the literature, but is growing in importance (French 2012).

It is possible, indeed likely, that some combination of these three contexts may occur. For instance, a group of decision makers might be informed by a group of experts before making their decision. However, we avoid such complexities and use just these three contexts as they are sufficient to articulate some of the issues and principles that are of concern. We shall discuss, for instance, that whether it is appropriate to calibrate expert judgements depends to some extent on the context.

Within societal decision making, the Expert and Textbook problems are very common contexts. For example as an instance of an Expert Problem, a regulator may wish to act as a single rational decision maker taking advice from a panel of experts. The EFSA guidance is written very much from this perspective. As we have noted, the Textbook Problem corresponds very much to circumstances when a government sets up a panel of experts to advise on some issue of public concern. One might think that the Group Decision Problem would correspond to decision making within many areas of government, but that is seldom the case. Political processes dominate and seldom correspond to the simplicity of any decision theoretic model (French et al. 2009). Within the private sector, it is more common to find circumstances with the structure of the Group Decision Problems, possibly because members

[4]This is becoming less true when web-based group decision support systems are used and decision makers may simply enter numerical judgements of uncertainties and values into the system, while being separated in space and possibly time (French 2007).

of a commercial organisation have more closely correlated values than is found across the whole of society meaning that some of the paradoxes and inconsistencies possible in theory do not occur in practice (Bacharach 1975).

In all contexts we assume that the experts articulate their uncertainties in terms of probabilities, perhaps on a single event A, a series of events $A_1, A_2, \ldots, A_n$, or the probability distribution of an unknown quantity θ, either discrete or continuous. We also assume that these will be used in a statistical, risk or decision analysis. We would expect these to be structured according to the Bayesian paradigm (see, e.g., French and Rios Insua 2000; Bedford and Cooke 2001; Gelman et al. 2013), because such approaches are consistent with subjective theories of probability which also underpin the use of probability to encode expert judgement. However, as the EFSA guidance shows, it is quite possible to move back to non-Bayesian approaches in using the probabilities from expert judgement studies. We also note that just as the elicitation and aggregation of expert judgement requires attention to the processes and procedures of which the mathematical analysis forms a part, so there are many 'softer' issues to be considered in conducting a statistical, risk or decision analysis (Edwards et al. 2007; French et al. 2009). We would contend that all these procedures and processes need be consistent and based on compatible principles.

It is important to realise that in assuming that the experts provide probability assessments, we do not claim that they will do so 'accurately' in the sense that any probability that they give relates to reality in an unbiased way. Ideally we would wish that an articulated probability of, say, 60% corresponds to something that is observed to occur roughly 60% of the time. However, experts, as indeed everyone else, are liable to biases which manifest in many ways (Kahneman et al. 1982; Lichtenstein 1982; Gigerenzer 2002; Kahneman and Klein 2009; Kahneman 2011). For instance, experts are often overconfident, assigning probabilities of 100% and 0% to events that subsequently prove to be neither certain nor impossible. This means that we should consider the calibration of experts, just as we might consider the calibration of instruments rather than simply accepting their readings without question. We shall see that the calibration of experts has been a significant hurdle to developing practical and tractable Bayesian methods for assimilating and aggregating expert judgements into any analysis.

Calibration issues arise because of the psychology of the individual experts. They are confounded by further pressures which can bias expert judgement. These can arise from social pressures and legal responsibilities that impact on the experts. For instance, geologists, particularly in Italy, would inevitably consider their assessments in the light of the indictments of the experts who advised on the likelihood of an earthquake at Aquila (Alexander 2014). Although they were subsequently acquitted on appeal, it is far from clear that this will be last time that experts in any country are placed at risk in court because of their advice and the subsequent outcomes; and that possibility will always be in the background of their thinking.

Often the advice of 'independent experts' is sought, but they are as unlikely to be found as the elixir of life. Experts inevitably share much experience and education in common, which correlates their advice (Wilson 2016), see Chap. 9 of this book

(Wilson and Farrow 2017). Modelling dependence between experts has proved to be another significant hurdle to the implementation of practical Bayesian Methods.

6.3 The Bayesian Approach to Structured Expert Judgement

Bayesian approaches treat the experts' judgements as *data* and then develop appropriate likelihood functions to represent the information implicit in their statements. Thus applying Bayes theorem to assimilate their advice on uncertain events to give the posterior probabilities given the experts' statements:

$$P_{DM}(A|\mathbf{Q}) \propto P_{DM}(\mathbf{Q}|A) \times P_{DM}(A), \tag{6.1}$$

where:

- A is an event or series of events $A_1, A_2, \ldots, A_n$
- $\mathbf{Q}$ are the experts' statements, $\mathbf{Q} = (\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_e, \ldots, \mathbf{q}_E)$, expert e stating $\mathbf{q}_e$ for $e = 1, 2, \ldots E$
- $P_{DM}(A)$ are the decision maker's prior probability distributions for A
- $P_{DM}(\mathbf{Q}|A)$ are the decision maker's probabilities for the experts stating $\mathbf{Q}$ given A, i.e. the likelihoods,
- $P_{DM}(A|\mathbf{Q})$ are the decision maker's posterior probability distributions for A given the experts' statements $\mathbf{Q}$

or, to assimilate their advice on uncertain quantities:

$$P_{DM}(\theta|\mathbf{Q}) \propto P_{DM}(\mathbf{Q}|\theta) \times P_{DM}(\theta), \tag{6.2}$$

where:

- θ is the unknown quantity of interest to the DM,
- $P_{DM}(\theta)$ are the decision maker's prior probability distributions for θ
- $P_{DM}(\mathbf{Q}|\theta)$ are the decision maker's probabilities for the experts stating $\mathbf{Q}$ given θ, i.e. the likelihood function,
- $P_{DM}(\theta|\mathbf{Q})$ are the decision maker's posterior probability distributions for θ given the experts' statements $\mathbf{Q}$

The constant of proportionality is in principle simple to evaluate. It is found by remembering that probability distributions integrate to one. Unfortunately the integration is not easy in many cases. We shall return to this point in the next section.

A key difficulty in the Bayesian approach is the development of tractable likelihood models, $P_{DM}(\mathbf{Q}|A)$ or $P_{DM}(\mathbf{Q}|\theta)$, that capture the decision maker's understanding of:

- the ability of the experts to encode their knowledge probabilistically and their potential for overconfidence (Clemen and Lichtendahl 2002; O'Hagan et al. 2006; Hora 2007; Lin and Bier 2008);
- the correlation between expert judgements that arises from their shared knowledge and common professional backgrounds (Shanteau 1995; Mumpower and Stewart 1996; Wilson 2016), see Chap. 9 of this book (Wilson and Farrow 2017);
- the correlation between the experts' judgements and the decision maker's own judgements (French 1980);
- the effects of other biasing pressures such as 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 shared knowledge and the correlations this brings between the judgements of the experts; other approaches to aggregating expert judgements do not. As any statistician knows, ignoring dependences 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.

Bayesian modelling extends to allow for both expert judgement and empirical data to be assimilated. All that is required is that the likelihood function models the observation processes of both, which, of course, may not be a simple task. Moreover, while we are assuming for the present that the experts articulate their uncertainties in terms of probabilities, they could give means, variances or some other moments of their distributions, and again the likelihood function could model this.

The Bayesian model as stated here clearly maps onto the expert problem described above. It describes how the decision maker should update her beliefs in the light of the information she receives from the experts. In the case of the group decision problem, some approaches apply the Bayesian model separately for each group member to represent how they should learn from each other, so developing theories of *Bayesian conversations* (French 1981; Kadane 1993). Other approaches to group decision making assume that the members articulate their probabilities after sharing all their information through discussion, so that the process of learning from each other is never formalised. This is true of the Sheffield method and more generally of decision conferencing.

It is clear that the use of probability to represent uncertainty is key to our discussion and it would be wise to pause and be clear on what we mean by probability. Firstly, we note that expert judgement studies have to recognise two broad forms of uncertainty:

- *aleatory uncertainty* or *randomness*: such uncertainty relates to natural variation and randomness such as the unpredictability of the weather or variations within a species;
- *epistemological uncertainty*: such uncertainty relates to a lack of knowledge or scientific understanding.

In expressing their opinions, experts have to integrate both forms of uncertainty into a holistic expression of the total uncertainty in the event or quantity of concern. But there are more subtle issues relating to the meaning of probability, which is a surprisingly controversial topic that has been of interest to many philosophers over the centuries. Broadly there are four families of approaches (Barnett 1999; French 2013).

The *classical* view of probability simply partitions the future into n equally likely primitive events and then to get the probability of a more complex event counts those primitive events that comprise it, say q. The probability of the complex event is then taken as q/n. This works well in game of chance, for which it was devised in the seventeenth century, but quickly falls apart in less well-structured problems in which partitions of equally likely events are absent. Nonetheless, for most of us it is the form of probability that we meet earliest in our schooldays.

The *frequentist* view of probability takes probability to be the long run frequency of occurrence of some event in a series of repeated trials. It is the conception that underlies our first introduction to statistics. It is intuitive and very powerful, *when it works*. However, it does not deal with epistemological uncertainty since it is very difficult to imagine a series of trials in which one learns and unlearns knowledge repeatedly. Learning a piece of knowledge is, at least for a rational person with a good memory, a one-off, unrepeatable event. So frequentist approaches to probability do not fit with expert judgement studies.

Both classical and frequentist views of probability make probability an objective property of the system under observation. The *logical* view of probability takes a different approach. It assumes that probability is an objective property of the language in which the system is described. As knowledge accumulates some propositions in the language are observed to be true and others false. If one starts conceptually with the language and absolutely no knowledge, i.e. no proposition is known to be true or false, then one can develop a full theory of probability in which the probability of a proposition represents how likely it is to be true. Moreover, the way in which knowledge accumulates is entirely compatible with Bayesian updating: see, e.g., Jeffreys (1961). Unfortunately, it turns out to be very difficult, many believe impossible, to articulate the complete ignorance with which such theories of probability need to begin.

Which brings us to subjective views of probability (Ramsey 1926; Savage 1972; De Finetti 1974, 1975). Here probability is a property of an observer of a system, representing his or her degree of belief in something happening or of the value of a quantity. There are subjective theories which simply articulate an individual's uncertainty in this sense; and there are also approaches which combine subjective uncertainty with subjective values to create theories of rational decision based upon

subjective expected utility (French and Rios Insua 2000). It is these latter theories that underpin much of modern Bayesian risk and decision analyses and that are implicit in the Bayesian approaches to the use of expert judgement. In this chapter, we adopt an explicitly subjective approach to probability.

There is a further issue we need consider in thinking about the meaning of probability in expert judgement studies. *Whose* probabilities are we modelling: the decision maker's, the experts', someone else's, ...? Clearly in the strict Bayesian formalism above, the experts articulate their own probabilities and these are captured in the $\mathbf{Q}$. The decision maker's beliefs are modelled by the probability distributions $P_{DM}(A), P_{DM}(\mathbf{Q}|A), P_{DM}(A|\mathbf{Q})$, etc. In the expert problem this is a straightforward interpretation. In the group decision problem, the same is true if we consider each member individually. In the textbook problem, things are far from clear: there is no decision maker to own the probabilities. Moreover, if we consider societal risk and decision analysis as it needs to be practised by regulators such as EFSA, then again things become unclear. There may be a well-structured risk or decision model, but the situation may not correspond neatly to either the expert problem or the group decision problem. Either there is no explicit decision or, more likely, once scientific and socio-economic analyses are complete, decisions are taken by political processes, often somewhat nebulous political processes. There is no use of subjective expected utility to guide the final decision (French and Argyris 2016). In such cases, it is common to re-interpret the analyses as producing a model of what a rational scientist or supra-decision maker might believe in the light of the evidence, including expert advice, available. We shall return to this point in the discussion below. For a related comments, see O'Hagan and Oakley (2004). Need the experts, decision maker and analyst use the *same* interpretation? This may seem a strange comment, but from a Bayesian perspective, while the decision maker's probabilities are clearly subjective, to her the experts' probabilities are simply *data*. The decision maker's likelihood encodes how see perceives this data, i.e. their statements, relate to the uncertainties of interest. Such likelihoods could be constructed, at least in principle, however each expert interprets the probability that he or she states. Indeed, different experts might use different interpretations; what matters is that the decision maker and analysts know which interpretation they are using.

The ethics behind calibration is interesting. In the context of the expert problem, the experts' judgements are data to the decision maker and it is entirely appropriate for her or her analyst to correct/adjust their judgements for poor calibration, just as it is appropriate to correct empirical data for known biases and flaws in the observation process. It may be difficult to do so, but the ethical position is clear. However, in the case of group decision making, things are not so clear. Each member might adjust what he or she learns from the other group members' judgements for any miscalibration in their judgements. But should each member adjust their own judgements? Why, if it is what they truly believe? Surely, as decision makers sharing in responsibility and accountability for the decision, they should each vote

or whatever according their beliefs, even if others perceive these as miscalibrated? Moreover, any process which involves an individual recalibrating his or her own probability risks entering an infinite regress.

If we turn to societal decision making and the textbook problem, things become more interesting. Firstly, there is no clear individual decision maker, so the Bayesian paradigm which is essentially individualistic, does not apply transparently. One has to develop the concept of a hypothetical rational scientist or supra decision maker, who listens to all the expert judgements and forms a reasonable synthesis of these. With this interpretation, it is sensible to assume that this hypothetical being is well-calibrated, but that the real experts may be poorly calibrated. Here it seems reasonable to recalibrate the experts' judgement. But there are legal issues. If a panel of experts is charged with giving their best judgements to some body such as a regulator or the government itself, is it legitimate to adjust those judgements in subsequent analyses? Even if adjustment is legitimate, would it be acceptable to many stakeholders? Will they placidly accept some risk management strategy when the judgements that have been used to justify it have been 'blatantly tampered with'?

The Bayesian perspective does not give any easy answers to any of these questions, but it does make them explicit and open to discussion.

6.4 Survey of Bayesian Models for Structured Expert Judgement

Bayesian models have evolved considerably since their inception. The first models used conjugate prior methods, which simplify the calculations by utilising a restricted set of distributions in the Bayesian model to make the mathematics easier. Examples of such models may be found in Winkler (1981), Lindley et al. (1979) and Wiper and French (1995). These models conceptually demonstrated the power of the Bayesian approach, often producing favourable results on small datasets, however, in practice were not broadly adopted. The reasons for this relate to the restrictions the conjugate assumption put on the model, the complexity in modelling approach in comparison with intuitively simpler opinion pooling methods, and the sensitivity to inputs that was apparent in some of these approaches.

Following the conjugate prior models, there was some further investigation into other Bayesian approaches that could be fruitful. Some progress was made on Bayesian Nonparametrics (Lichtendahl 2005) and Copulas (Jouini and Clemen 1996). Working with multivariate distributions, those with more than one dimension (a phenomena prevalent throughout structured expert judgement (SEJ) studies), can increase complexity significantly. Copulas simplify the process by separating what are known as the marginal distributions, which are distributions for the individual dimensions, from the dependence structure which demonstrates how they are linked

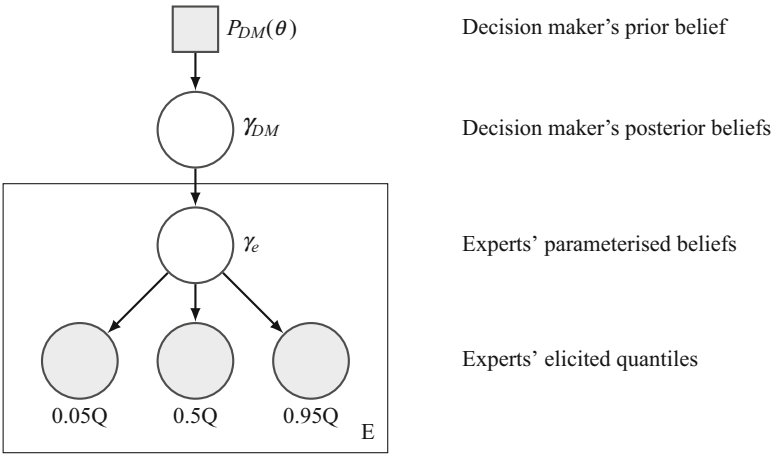


Fig. 6.1 Bayesian Network for aggregating expert judgement

together. There were some positive signs from these models, however, empirically there were questions over the method (Kallen and Cooke 2002)[5]

In more recent years there has been some resurgence in the use of Bayesian analysis for expert judgement studies; however the focus has now shifted to Markov Chain Monte Carlo (MCMC) approaches. These methods are a class of algorithms for stochastically approximating Bayesian posterior multi-dimensional distributions. Despite MCMC approaches having existed for many years, the use of these techniques in expert judgement studies is a recent development. One of their attractions to Bayesian expert judgement studies is the ability to describe the model through the use of a Bayesian Network. This seemingly simple box and arrow diagram, Fig. 6.1, can allow the analyst to easily demonstrate to the decision maker what the different variables or parameters are and how they are linked together, thereby hiding much of the modelling complexity.

In this toy example we see a simple expert judgement aggregation model for the expert problem[6] described. Here, experts have their beliefs elicited in the form

[5]It is important to note that the models themselves encode expert judgement and therefore cannot be thought of as ad-hoc. Indeed it would be reasonable to question whether it is appropriate to separate the elicitation exercise from the modelling process, however in doing this some of this encoded knowledge would be lost. Therefore we need to be very careful with the treatment of elicited data outside of the modelling paradigm, a particular challenge in the Textbook context outlined earlier.

[6]It is interesting to consider whether it would be possible to create a Bayesian Network (BN) for the group problem or the textbook problem in addition to the expert problem. Unfortunately this is significantly less trivial. For the textbook problem, by definition, the problem statement is not known at the time of elicitation and therefore it is impossible to generate a corresponding BN. It would be feasible to generate specific networks as individual problems are solved but a generic version does not exist. For the group problem, there is significant work looking at complex

of a set of quantiles, these quantiles are utilised as representatives of the experts 'true' belief which is in the form of a parameterised distribution with parameters γ_e (e.g. in the standard Gaussian $\gamma_e = (\mu_e, \sigma_e)$). Quantiles are often elicited, rather than parameters, due to the relative complexity of asking experts to think in terms of distributions. The decision maker's beliefs, represented by γ_{DM}, again parameterised, are, as in the standard Bayesian model outlined, updated based on the underlying prior she had $P_{DM}(\theta)$, combined with the probability assessments from each expert. The actual inference over this network would be produced by running an MCMC algorithm which would infer an output distribution for each of the unknown variables. Another advantage of MCMC algorithms is that they maintain uncertainty throughout and therefore it is possible to visualise the underlying distributions for each of the points of interest within the model, such as the expert's true beliefs or the decision makers full posterior distribution. With the appropriate treatment, this also allows the decision maker to understand more complex items such as the correlation between experts, or their relative calibration, without explicitly trying to elicit or model these separately.

One of the early MCMC models for SEJ was from Clemen and Lichtendahl (2002). In this paper the authors tackled a specific portion of the expert judgement problem, the issue of expert calibration. It is important to remember the change in terminology here vs. previous chapters. Bayesian models normally approach calibration by uncovering parameters by which the decision maker believes the expert's tend to over/under forecast and adjust the resultant forecasts accordingly, analogously similar to the adjustments a conductor might ask a player to make to the valve in an instrument in order to ensure harmony. This is different to calibration in the Classical model context where experts' forecasts are never adjusted but simply weighted according to their statistical accuracy. As outlined in the previous section, the legal and philosophical validity of this as an activity may be very context dependant. Building on ideas from Cox (1958) and Morris (1974); Clemen and Lichtendahl (2002) developed a model of expert overconfidence using past data to estimate, what they term, 'inflation factors' for assessed distributions post hoc. While some common models treat all experts as exchangeable, Clemen and Lichtendahl use hierarchical MCMC models which allow experts to be calibrated individually. Here, for simplicity we can imagine, the model has a parameter, α, for each expert which describes whether that expert displays consistent bias (i.e. continuously over/under estimating) on their best guess, (or 50% quantile) across forecasts. The MCMC algorithm then sequentially reviews experts' previous performance at forecasting, over a set of data known as the 'seed data', and infers the value for α (it will ultimately be a distribution rather than a point estimate). From this the decision maker can decide how to consider each experts judgements for future forecasts. The authors then extended this model to consider the other

decisions involving both groups and Bayesian networks in the field of adversarial risk analysis. Due to the focus of the EFSA guidance on the expert problem, a detailed review of group decision problems is not given in this chapter. One paper covering this topic is French (2011).

elicited quantiles and expert to expert correlation by creating further parameters, which represent the non-independence of experts.

Clemen and Lichtendahl did not explicitly consider the choice of variables used for calibration, though this is clearly important. As mentioned, the underlying assumption of all calibration techniques is that the behaviour experts' display on the seed variables is indicative of their final behaviour on the variables of interest. In particular, there are systematic biases, of a similar nature to those outlined by Kahneman et al. (1982), in consistent evidence which should be removed from the decision maker's analysis. The case for this is compelling but critically, only when the variables used for calibration are representative of the target variables of interest. A decision maker should not expect an expert's performance with relation to a weather forecast to be indicative of their ability to accurately assess the likelihood of a bolt breaking in a suspension bridge. Similarly, the data must be on a similar scale. Experts are notoriously inaccurate when assessing probabilities for extremely rare events and one would expect that behaviour seen here would not correlate with behaviour seen for more commonly occurring variables. Assessing the right variables to use for a calibration model remains a question, and something that should be researched further.

Although Clemen and Lichtendahl tackle the question of how we calibrate multiple experts whilst assessing the expert to expert correlation, they do not consider the issue of what a decision maker should do once she has received this data. Utilising the authors' methods a decision maker would be able to translate multiple experts elicited quantities into their unbiased counterpart's, however, how the decision maker would actively use these is not apparent. It would seem a shame to precisely calibrate experts but then for the decision maker to update her belief in a method that does not use this richness of information. To this extent, it is important to examine Bayesian methods of *aggregating* the data also.

More recently, Albert et al. (2012) proposed such a model. Their model is known as a Supra-Bayesian parameter updating approach. This outlines a class of models which assume that the aggregation represents the belief of an overarching rational, but hypothetical, decision maker, the Supra Bayesian, who has beliefs that can be represented by particular parameters. For example, they may believe that the output is Gaussian with an unknown mean and variance. The model then updates these parameters based on inference over the experts judgements (here, as usual, utilised as data). Similar to the calibration model, this method also considers expert judgements from indirect elicitation, i.e. rather than trying to elicit a mean and a variance for each expert's beliefs, expert's knowledge is elicited on more intuitive observables, such as quantiles, and the parameters then inferred. Here the inference is made by mapping the elicited quantiles (or similar) to a selected parameterised distribution, using distribution fitting. Different models often use different parametrisations for this. The parameters of the fitted distribution are assumed to represent the expert's underlying belief that often cannot be directly elicited due to the complexity of mentally processing these concepts.

The model that Albert et al. (2012) use is hierarchical in nature and captures correlation in an interesting way. One of the drivers of inter-expert correlation is

that experts may have had similar education or historic frames of references. The proposal from the authors of this paper therefore was to group experts together into homogeneity groups, where each group is defined by like-minded experts. Here the authors take like-minded to mean 'similar background or schools of thought', although do not go into specific detail on how this may be assessed. For now we will assume that experts can be appropriately grouped in some way, however we will return to this shortly. The aggregation model will then assume that each expert's beliefs are linked to that of the other experts in their group and the groups likewise are linked to each other. Each group, h, will have a parametrised distribution γ_h defined by the beliefs of its expert members. 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. 6.2. This is an extension of the simpler model outlined earlier.

The motivation for this expert partition is that rather than explicitly calculating the correlation for each expert, the grouping approach is used to appropriately weight the impact of each expert in the final model, offsetting over-confidence effects driven by correlation. The theory here is that past experience and knowledge is one of the underlying key drivers of this correlation. One of the advantages of this approach is that the hierarchical model can capture both the consensus and diversity between experts, and this is very compelling. As mentioned above, one of the areas not overtly tackled by the authors was how to support a decision marker or analyst in assigning experts to groups. The authors of this chapter are currently researching an

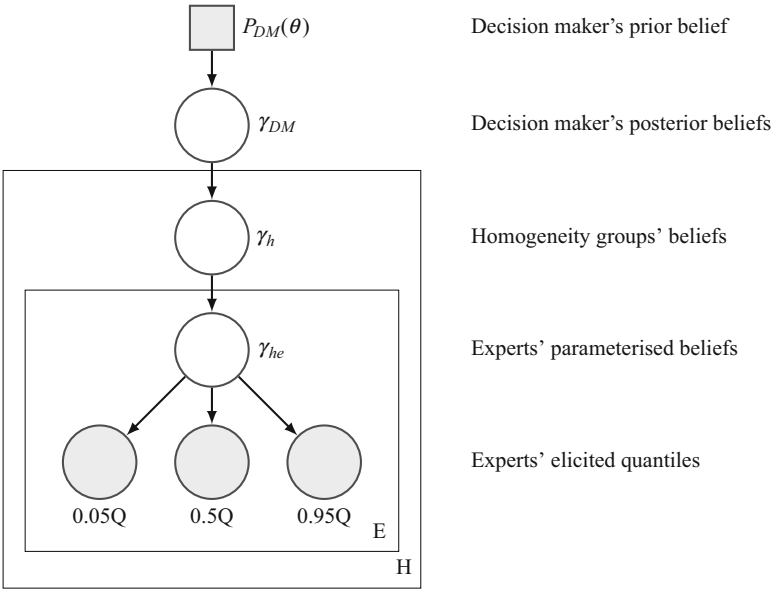


Fig. 6.2 A Bayesian Network for aggregation of expert judgement with homogeneity groups

algorithmic approach to this assignment utilising Dirichlet process mixture models, which are a form of clustering technique.

For Bayesian models there has also been much discussion about a more advanced set of properties that could be expected of a Bayesian SEJ model such as *external Bayesianity* (Madansky 1964; French 2011), in which the final decision of a group of Bayesians should in turn be perceived as Bayesian, or *marginalisation* (Genest 1984; French 1985). However we shall not discuss these in detail here.

Overall, with these recent developments, it appears that the goal of finding a, practical Bayesian framework for SEJ is not an impossibility and MCMC may be a critical design element. The more recent models outlined also show the potential for the Bayesian approach to take a significant step forward in being more context agnostic. There is, however, a number of complex idiosyncrasies that make these techniques challenging in a practical environment compared to the Classical model, in particular:

- Technical details being intractable to non-analytical decision maker's
- Complexity (and model overdependence) in setting the correct priors
- Overreliance on hard to gather calibration data.

Some of these issues can be resolved by research into the modelling techniques utilised, however, others can be better addressed by considering the processes and procedures that may need to be different for a Bayesian model of SEJ.

6.5 Practical Procedures

Of the major texts covering the processes and procedures surrounding structured expert judgment, the European Food Safety Authority's (EFSA 2014) guidance is possibly the most complete. Many of the topics and models are also outlined in other texts such as Meyer and Booker (1991), Cooke and Goossens (2000), and the Sheffield Method by O'Hagan and Oakley; but we shall take EFSA (2014) as an exemplar to discuss process and procedural issues from a Bayesian perspective.

It is important at this stage to note the role of EFSA,[7] the European Agency. It operates independently of both the European legislative, executive institutions and EU Member States, and is responsible for risk assessment in the area of European food safety. This is completely separate from risk management or policy making, and was legally established under the General Food Law—Regulation 178/2002. EFSA plays an important role in collecting and analysing data to ensure that risk assessment is supported by scientific information, including expert judgement, and then appropriately communicating this to both stakeholders, such as policy makers, and the public at large. Acting in this way, EFSA will be most interested in evidence and analysis for societal decisions, impacting the approaches and contexts

[7] https://www.efsa.europa.eu/en/aboutefsa.

in which EFSA operates SEJ. EFSA is a regulator and so deals with expert problems or, occasionally, textbook problems. In these contexts, EFSA would seem to be developing probability distributions that represent the views of a *rational scientist*.

The EFSA SEJ process starts with the formation of a Working Group. The Working Group comprises individuals accountable for the overall program of work. They are tasked with problem definition and the development of a risk assessment model. In undertaking these they will identify when limited evidence is available for some of the critical variables, deciding that there is a requirement to consult experts to fill such knowledge gaps.

At this stage, the Working Group will typically hand the program over to a second group, the Steering Group. The role of the Steering Group, *inter alia*, is to refine the parameters to be elicited and to identify the precise expert knowledge that is needed. Once these elements have been finalised, it is critical to select the experts themselves, which may have implications for the elicitation method used and thus the final aggregation model. The final decision on each of these elements lies with the Steering Group.

In practice, the selection of the experts is not a trivial matter and can be impacted by a number of variables quite outside considerations of their expertise on the parameters. Availability is obviously a critical factor; and for EFSA there are potentially political constraints factored into the decision. There may, for example, be quotas on attendance from EU member-states or other issues of representation. This constraint may impact the analysis as it potentially introduces a further risk of expert bias which may need to be controlled as part of the elicitation/modelling process. This issue is not just limited to EFSA nor to similar contexts, it is a common phenomenon in SEJ that experts may be assigned rather than selected. From a Bayesian perspective, were this assignation to happen, having a clear understanding of these affiliations may be critical for the creation of the homogenisation groups and the ultimate reduction of bias within the model.

In selecting the model, the Steering Group can choose from three approaches ratified by EFSA, these approaches are the Sheffield method, Cooke's method or a version of the Delphi method. Each of these versions have different requirements and the ultimate selection of the model will depend on factors such as geographical split of the experts, diversity of backgrounds, or simply time or skill requirements.

Following the selection of the elicitation method and the model to be used, the Steering Group will hand over to an Elicitation Group. The Elicitation Group will typically be more familiar with facilitation and elicitation, and will be accountable for training the experts to help them understand the process, the requirements necessary and to also help them be aware of their own biases and how to mitigate these. Following this the Elicitation Group will perform the elicitation and any subsequent modelling necessary. The information from here is then either handed back to the Steering Group or directly translated into a set of Post Elicitation documentation. Please see Fig. 6.3 for a simple visual of the process.

In a typical EFSA study, there are a wide variety of individuals involved throughout the process. It is imperative to ensure that common understanding of context and models flow through the groups. Documentation is a critical component to this. In other SEJ contexts, it is feasible that the process is much leaner and the

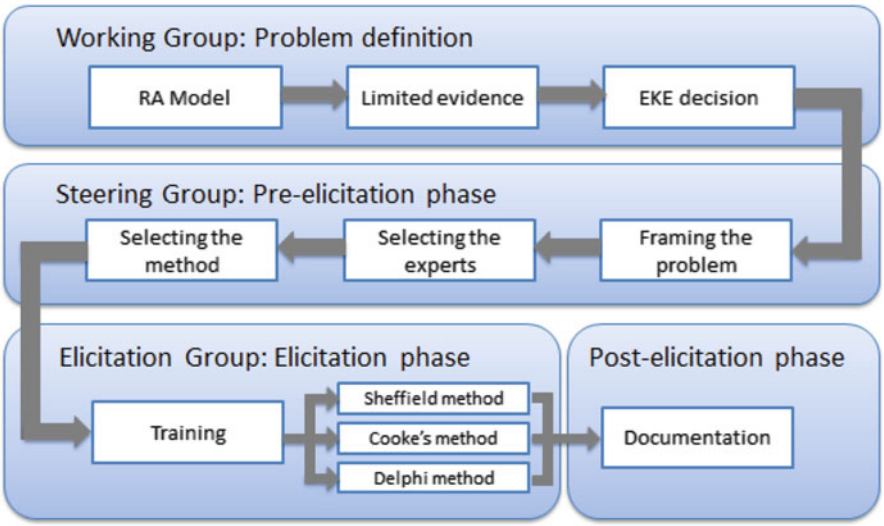


Fig. 6.3 The EFSA expert knowledge elicitation process, reproduced from EFSA 2014

Working Group and the Steering Group are compressed. Or in certain decisions, it is feasible that a decision maker will come directly to an analyst for insight that will ultimately lead the analyst to play the role of all three of the groups outlined here. In these circumstances the common understanding is much simpler to achieve, however the training and documentation requirements may be harder to implement with limited resource.

If a Bayesian adjustment of EFSA were considered, a number of other elements would need to be addressed. As highlighted, the key difficulty in the Bayesian approach is often the development of tractable likelihood models, $P_{DM}(\mathbf{Q}|A)$ or $P_{DM}(\mathbf{Q}|\theta)$, as these are the most mathematically complex elements. However, when considering the EFSA procedures the decision on who should own the prior for the model $P_{DM}(A)$ is an equally important consideration. In a fully subjective context, i.e. when there is a single decision maker who is ultimately accountable for the output of the SEJ, e.g. a commercial leader who is utilising SEJ to invest their own money, it is very reasonable for this individual to own the prior personally. EFSA is typically looking for the view of a rational scientist, and therefore potentially a Supra-Bayesian approach, as utilised in Albert et al. (2012) is necessary. This hypothetical Supra-Bayesian would evidently still need to have prior's assigned. There are a couple of areas of potential ownership:

- **The Working Group**—*Ultimately accountable for the output of the SEJ as it feeds into the risk assessment model, one option would be to have the Working group define and own the priors.*
- **The Steering Group**—*Closer to the refined parameters being elicited, the experts and the final method utilised, the Steering Group would also be a logical*

owner for the priors. If any quantitative assessment of the variable of interest was made during the initialisation of the pre-elicitation phase then this could be considered as a prior.

- **The analysts performing the modelling**—*If naive priors*[8] *were used and therefore limited knowledge encoded within the prior specification, the analysts could feasibly act as a proxy for the rational scientist.*
- **The experts themselves**—*Another source of priors would be the experts themselves or a subset of other experts. Utilising the experts for the prior however, would draw us closer to the group decision problem, rather than the rational scientist expert judgement problem typified in the EFSA model. This is not inherently a problem however does bring about a number of other constraints to be considered and blurs the boundary of the role of the expert vs. the decision maker. If this approach were to be taken, a very different set of processes would need to be considered. For an expert problem all knowledge from the expert's should be codified in the likelihood function.*

The most compelling of these options would appear to be the Steering group, as they represent a body close enough to the problem whilst still in a position of accountability. There are also multiple priors being assessed, there are priors over the variables of interest but also over the experts (and their potential correlations) themselves. It would be a considerable risk for the analyst to be accountable for these priors due to the potential impact on the output and the legal ramifications discussed before.

Regardless of the ownership, decisions need to be made on whether naive priors should be considered. Naive priors would focus the final output much more directly on the expert judgement, however, would clearly reduce the amount of data that could be encoded into the problem. The context itself here is important, in a fully subjective model with a specific decision maker it would be unwise to utilise naive priors as ultimately you are trying to update someone's belief in the light of expert opinion and the decision maker's belief is naturally a critical starting point to this. For EFSA and the rational scientist viewpoint it would appear sensible to aim for naive priors over the variables of interest. However, we would argue that priors over the experts should not be so naive. If for example we consider calibration; starting from the belief that experts are well calibrated and only recalibrating with significant evidence (where significance here is determined by the application of Bayes rule with a calibration data set) rather than starting as agnostic to calibration issues, would appear an appropriate decision for the Steering Group.

Another critical component of the EFSA process that must be analysed from a Bayesian perspective is documentation. One constituent of the EFSA guidance is often a shared evidence dossier. This dossier captures all of the known data regarding the parameters of interest, and the risk assessment model, ahead of the elicitation exercise and is shared with all of the experts. This is important for

[8]Naïve priors are very flat distributions which seek to represent complete lack of knowledge or something close to it.

transparency; and from an auditing perspective, it would appear to be unethical to with-hold evidence from an expert before they are due to make judgements that may impact critical decisions. The legal ramifications of a decision being recommended when data was withheld may be substantial. However, there is both a technical and philosophical issue with this approach. The technical issue is that in creating this evidence dossier the Steering group may inadvertently increase the correlation between experts as, by definition, they are given a shared body of knowledge from which to base their judgements. From a philosophical perspective there is also an issue with this approach as it makes assertions about the evidence base that may not be complete, indeed as we are engaging in a SEJ study it is incomplete by definition. Sharing this partial data with the experts may further bias the results, for example it increases the risk of the availability bias being demonstrated, or for experts to become increasingly overconfident. There are a couple of ways that this issue could be handled in the Bayesian model:

- **Evidence Dossier shared with experts pre-elicitation** as per the standard EFSA guidance—*Increases the risk of bias and correlation but ensures that no data is withheld.*
- **Evidence Dossier shared with experts during elicitation**—*It would in theory be feasible to elicit the experts knowledge before they see the Evidence Dossier and then perform a second elicitation after this has been shared. This would allow the analysts to directly ascertain the impact of the evidence dossier and consider this within the final recommendation, however, would put a significant burden on the elicitation process.*
- **Utilise the Evidence Dossier in prior definition**—*As empirical evidence, it would be potentially feasible for the evidence dossier to be used by the Steering Group, rather than the experts, in the definition of the priors. This would ensure that priors encoded knowledge, but only empirically generated knowledge, and would also help to ensure that the data is utilised in the process. Here the experts never see the evidence dossier and therefore there is no increase in cross expert correlation, however, the issue of data being withheld is reduced as any final recommendation will be net of any existing evidence.*

The decision of which approach to use in any analysis may again be context dependant.

A further decision to be made is; How to pass back qualitative knowledge along with the consensus distribution? The final output of any structured elicitation exercise should not just be the consensus distribution itself but also the qualitative knowledge experts utilised to inform their decision making. This qualitative knowledge can be critical in getting decisions ultimately implemented and to enrich/explain the outputs of the analysis. Any documentation requirements should consider these elements in addition to the distributions.

Finally, in addition to the processes and procedures in place, it is also important to consider the software utilised to support any elicitation or to perform any analysis. Currently there are a number of pieces of software available on a mixture of different platforms with various levels of validation. One element that is critical

to many decision making contexts is security and so for broad use it is probably important that software is not web-based. However, transparency and auditability in the software itself is paramount, in order to ensure that the usage is adopted in many different contexts. Work needs to be done to harmonise the existing approaches, and any new models as they develop, into a single toolkit for analysts and decision makers alike. Integrating the software in this way could provide the support necessary to further enhance the procedural guidance, such as EFSA, that is already available. We propose this 'meta-software' would further help to embed the use of structured expert judgement into currently untapped contexts.

6.6 Conclusions

This chapter has discussed some of the considerations needed when utilising a Bayesian approach to structured expert judgement and the impact these have on existing processes and procedures. It is evident much of the broad framework exists in the current literature, however, some of the more important nuances to this outline when considering a more subjective viewpoint, or indeed some of the more complex contexts, are yet to be considered.

Much more than simply a modelling approach, the Bayesian perspective provides opportunity to step back and consider the basic framework of probability a decision maker is utilising when undertaking a piece of structured expert judgement. Considering the subjective perspective with both aleatory and epistemological uncertainty in this way gives a broad view of the easily glossed over challenges facing decision makers.

Some major elements, we have proposed, to be researched further to develop a full subjective approach to structured expert judgement are:

- The role calibration plays in different SEJ contexts
- Correlation between (and within) experts and decision makers
- The role information, such as evidence dossiers, play in updating expert's beliefs during analysis
- Validated SEJ software

Whilst we have outlined some of the approaches that could be taken to incorporate these into the current guidelines, it is clear that there is much work to do.

One key element apparent in all of the discussion is the role that context plays in how these issues can and should be solved. What is appropriate in the context of producing a piece of analysis representing the viewpoint of a rational scientist for a societal decision is not the same as the truly subjective decision of a commercial stakeholder choosing to invest their own money wisely. As much of the literature available today has developed from current utilisations of structured expert judgement, particularly in the context of regulators, it often focuses on a single contextual viewpoint. The Bayesian paradigm through its fundamental view on the definition of probability and uncertainty appropriately raises philosophical

questions about other contexts and, through the modelling approaches available, provides options on how to tackle these.

As the use of MCMC for expert judgement evolves and research builds on the work of Clemen and Lichtendahl (2002) and Albert et al. (2012), Bayesian models are likely to become much more tractable than historically and provide legitimate options for analysts to tackle some of these more complex context dependant challenges. As this happens the Bayesian paradigm will become more than an interesting philosophical set of challenges but a viable modelling option and included alongside the other approaches outlined in the standard literature. We are not here today, but the immediate future is looking bright.

References

Albert I, Donnet S, Guihenneuc-Jouyaux C, Low-Choy S, Mengersen K, Rousseau J (2012) Combining expert opinions in prior elicitation. Bayesian Anal 7(3):503–532

Alexander DE (2014) Communicating earthquake risk to the public: the trial of the "L'Aquila Seven". Nat Hazards 72(2):159–1173

Bacharach M (1975) Group decisions in the face of differences of opinion. Manag Sci 22(2): 182–191

Barnett V (1999) Comparative statistical inference. Wiley, Chichester

Bedford T, Cooke R (2001) Probabilistic risk analysis: foundations and methods. Cambridge University Press, Cambridge

Clemen RT, Lichtendahl KC (2002) Debiasing expert overconfidence: a Bayesian calibration model. In: PSAM6: San Juan, Puerto Rico

Cooke RM (1991) Experts in uncertainty. Oxford University Press, Oxford

Cooke RM (ed) (2007) Expert judgement studies. Reliab Eng Syst Saf 93(5):766–768

Cooke RM, Goossens LHJ (2000) Procedures guide for structured expert judgement in accident consequence modelling. Radiat Prot Dosim 90(3):303–309

Cox DR (1958) Two further applications of a model for binary regression. Biometrika 45:562–565

De Finetti B (1974) Theory of probability. Wiley, Chichester

De Finetti B (1975) Theory of probability. Wiley, Chichester

Edwards W, Miles RF, Von Winterfeldt D (eds) (2007) Advances in decision analysis: from foundations to applications. Cambridge University Press, Cambridge

EFSA (2014) Guidance on expert knowledge elicitation in food and feed safety risk assessment. EFSA J 12(6):3734

French S (1980) Updating of belief in the light of someone else's opinion. J R Stat Soc A143:43–48

French S (1981) Consensus of opinion. Eur J Oper Res 7:332–340

French S (1985) Group consensus probability distributions: a critical survey (with discussion). In: Bernardo JM, DeGroot MH, Lindley DV, Smith AFM (eds) Bayesian Statistics 2. North-Holland, Amsterdam, pp 183–201

French S (2007) Web-enabled strategic GDSS, e-democracy and Arrow's theorem: a Bayesian perspective. Decis Support Syst 43:1476–1484

French S (2011) Aggregating expert judgement. Rev R Acad Cienc Exactas Fis Nat 105(1): 181–206

French S (2012) Expert judgment, meta-analysis, and participatory risk analysis. Decis Anal 9(2):19–127

French S (2013) Cynefin, statistics and decision analysis. J Oper Res Soc 64(4):547–561

French S, Argyris N (2016) Decision analysis and political processes. In preparation. https://www2.warwick.ac.uk/fac/sci/statistics/staff/academic-research/french/publications/davspolitics_ver1.pdf
French S, Rios Insua D (2000) Statistical decision theory. Arnold, London
French S, Maule AJ, Papamichail KN (2009) Decision behaviour, analysis and support. Cambridge University Press, Cambridge
Garthwaite PH, Kadane JB, O'Hagan A (2005) Statistical methods for eliciting probability distributions. J Am Stat Assoc 100(470):680–701
Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2013) Bayesian data analysis. Chapman and Hall, London
Genest C (1984) A conflict between two axioms for combining subjective distributions. J R Stat Soc Ser B Methodol 46(3):403–405
Gigerenzer G (2002) Reckoning with risk: learning to live with uncertainty. Penguin Books, Harmondsworth
Gosling JP (2017) SHELF: the Sheffield elicitation framework. In: Dias LC, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgment. Springer, New York. doi:10.1007/978-3-319-65052-4_4
Hockey GRJ, Maule AJ, Clough PJ, Bdzola L (2000) Effects of negative mood on risk in everyday decision making. Cognit Emot 14:823–856
Hora S (2007) Eliciting probabilities from experts. In: Edwards W, Miles RF, Von Winterfeldt D (eds) Advances in decision analysis: from foundations to applications. Cambridge University Press, Cambridge, pp 29–153
Jeffreys H (1961) Theory of probability. Oxford University Press, Oxford
Jouini MN, Clemen RT (1996) Copula models for aggregating expert opinions. Oper Res 44(3):444–457
Kadane JB (1993) Several Bayesians: a review (with discussion). Test 2(1–2):1–32
Kahneman D (2011) Thinking, fast and slow. Penguin/Allen Lane, London
Kahneman D, Klein G (2009) Conditions for intuitive expertise: a failure to disagree. Am Psychol 64(6):515
Kahneman D, Slovic P, Tversky A (eds) (1982) Judgement under uncertainty: heuristics and biases. Cambridge University Press, Cambridge
Kallen M, Cooke R (2002) Expert aggregation with dependence. Probabilistic safety assessment and management. Elsevier, Amsterdam
Lichtendahl KC (2005) Bayesian models of expert forecasts. PHD Thesis. Department of Business Administration, Duke University, Durham
Lichtendahl KC, Winkler RL (2007) Probability elicitation, scoring rules, and competition among forecasters. Manag Sci 53(11):1745–1755
Lichtenstein S, Fischhoff B, Phillips LD (1982) Calibration of probabilities: the state of the art to 1980. In: Kahneman D, Slovic P, Tversky A (eds) Judgement under uncertainty. Cambridge University Press, Cambridge, pp 306–334
Lin S-W, Bier VM (2008) A study of expert overconfidence. Reliab Eng Syst Saf 93:711–721
Lindley DV, Tversky A, Brown RV (1979) On the reconciliation of probability judgements (with discussion). J R Stat Soc A142:146–180
Madansky A (1964) Externally Bayesian groups. RM-4141-PR: RAND
Meyer MA, Booker JM (1991) Eliciting and analyzing expert judgment: a practical guide. Academic Press, London
Morris PA (1974) Decision analysis expert use. Manag Sci 20(9):1233–1241
Mumpower JL, Stewart TR (1996) Expert judgement and expert disagreement. Think Reason 2(2–3):191–211
O'Hagan A, Oakley JE (2004) Probability is perfect, but we can't elicit it perfectly. Reliab Eng Syst Saf 85(1):239–248
O'Hagan A, Buck CE, Daneshkhah A, Eiser R, Garthwaite PH, Jenkinson D, Oakley JE, Rakow T (2006) Uncertain judgements: eliciting experts' probabilities. Wiley, Chichester

Quigley J, Colson A, Aspinall W, Cooke RM (2017) Elicitation in the classical method. In: Dias LC, Morton A, Quigley J (eds.) Elicitation: the science and art of structuring judgment. Springer, New York. doi:10.1007/978-3-319-65052-4_2

Ramsey FP (1926) Truth and Probability. In: Braithwaite RB (ed) The foundations of mathematics and other logical essays. Brace and Co., Harcourt

Savage LJ (1972) The foundations of statistics. Dover, New York

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

Skjong R, Wentworth BH (2001) Expert judgement and risk perception. In: Proceedings of the eleventh (2001) international offshore and polar engineering conference. The International Society of Offshore and Polar Engineers, Stavanger

Smith JQ (2010) Bayesian decision analysis: principles and practice. Cambridge University Press, Cambridge

Taylor AD (1995) Mathematics and politics. Springer, New York

Wilson KJ (2017) An investigation of dependence in expert judgement studies with multiple experts. Int J Forecast 33:325–336

Wilson K, Farrow M (2017) Combining judgements from correlated experts. In: Dias LC, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgment. Springer, New York. doi:10.1007/978-3-319-65052-4_9

Winkler RL (1981) Combining probability distributions from dependent information sources. Manag Sci 27(4):479–488

Wiper MW, French S (1995) Combining experts' opinions using a normal-Wishart model. J Forecast 14:25–34

Chapter 7
A Methodology for Constructing Subjective Probability Distributions with Data

John Quigley and Lesley Walls

Abstract Our methodology is based on the premise that expertise does not reside in the stochastic characterisation of the unknown quantity of interest, but rather upon other features of the problem to which an expert can relate her experience. By mapping the quantity of interest to an expert's experience we can use available empirical data about associated events to support the quantification of uncertainty. Our rationale contrasts with other approaches to elicit subjective probability which ask an expert to map, according to her belief, the outcome of an unknown quantity of interest to the outcome of a lottery for which the randomness is understood and quantifiable. Typically, such a mapping represents the indifference of an expert on making a bet between the quantity of interest and the outcome of the lottery. Instead, we propose to construct a prior distribution with empirical data that is consistent with the subjective judgement of an expert. We develop a general methodology, grounded in the theory of empirical Bayes inference. We motivate the need for such an approach and illustrate its application through industry examples. We articulate our general steps and show how these translate to selected practical contexts. We examine the benefits, as well as the limitations, of our proposed methodology to indicate when it might, or might not be, appropriate.

7.1 Introduction

Our goal is to acquire a probability distribution consistent with an expert's belief about the true value of a quantity of interest. In this chapter we explain how to construct such a prior probability distribution using observed data by adopting an empirical Bayes method embedded within an elicitation process to achieve consistency between the distribution obtained and the judgement of an expert. Motivated by the need to elicit subjective distributions within real industry applications, the

J. Quigley (✉)
University of Strathclyde, Glasgow, UK
e-mail: j.quigley@strath.ac.uk

L. Walls
Department of Management Science, University of Strathclyde, Glasgow G4 0GE, UK
e-mail: lesley.walls@strath.ac.uk

L.C. Dias et al. (eds.), *Elicitation*, International Series in Operations Research & Management Science 261, https://doi.org/10.1007/978-3-319-65052-4_7

methodology is grounded in core theoretical principles and aims to provide a useful, scientifically sound approach.

Core to our reasoning is the consideration of the ways in which an expert might assess uncertainty through analogy with similar events. In this respect we adhere to the view expressed by David Hume (1748) who wrote that *"All our reasonings concerning matter of fact are founded on a species of analogy"*. Others have acknowledged the role of empirical data for similar events in making assessments of uncertainty. For example, Kahneman and Lovallo (1993) proposed using empirical data as a means of correcting for overconfidence and optimism bias which might exist when an expert is asked to express her subjective assessments directly. Inherent in their so-called outside view is the mapping between the observed histories of the similar events and the future histories of the events associated with the quantity of interest. Practically, such an approach can be operationalised in various ways, including as a read-across process as discussed in EFSA (2015). Earlier, Koriat et al. (1980) articulated three stages for elicitation of probability judgements from an expert: first, memory is searched for relevant information; second, evidence is assessed to arrive at a feeling of uncertainty; and third, the feeling has to be mapped onto a conventional metric. However, they recognised that an expert's lack of experience in performing the internal mapping between feeling and a metric might lead to a corresponding lack of reliability, and/or incoherence, in the probabilistic expression of uncertainty. This chapter contributes a methodology consistent with an outside view which builds upon the initial stages of a probability elicitation but avoids the need for an expert to make an internal mapping. We aim to systematically support an expert to perform an appropriate mapping by grounding an analogy assessment in domain knowledge to select relevant empirical data for similar events which can then be translated into a defensible subjective probability distribution.

We begin by describing selected industry examples where both the need to express uncertainty about a quantity of interest and the opportunity for an expert to match the event to be predicted with an analogous pool of events exists. By abstracting from these examples, and by drawing upon theory from the wider literature, we present general steps for eliciting a subjective probability distribution using empirical data. The rationale and activities involved in each step are explained. Examples of implementing our approach illustrate how the general principles can be applied. We conclude by examining the benefits and shortcomings of our proposed approach to provide some insight on when it can be useful, when it might not be applicable, and issues to consider during implementation.

7.2 On the Nature of the Problem

7.2.1 Motivating Industry Challenges

Let us consider two industry contexts. Both examples are simplifications of real issues for which probabilities are required for variables within models developed to

support management decision-making. Here, we focus only upon issues related to the expression of the prior probability distributions.

First, consider a situation where a supply chain manager has procured a new supplier and wishes to assess the uncertainty in the true non-conformance rate of the parts to be supplied as an input to modelling quality related decisions (Quigley et al. 2018). The manager is uncomfortable making subjective probability assessments because the concept of quantifying some outcome that will in time be observable is cognitively challenging. But she is able to match the new supplier with similar existing suppliers since all have been subject to the standard procurement process. Hence the manager is relatively more comfortable in making analogy assessments between suppliers in terms of characteristics that might impact their performance. This judgement guides the creation of a relevant data set for existing suppliers providing a comparator pool that can be used to estimate a prior distribution. Of course, the uncertainty in the non-conformance rate of the new supplier represented by the estimated prior distribution should be checked for consistency with the beliefs of the supply chain manager.

Now consider a context where a new engineering design for an aerospace system is being developed as a variant of an earlier generation product (Walls et al. 2006). Typically the designers match the functionality of the new design specification and existing products to assess what aspects of the existing designs can be transferred. In addition, innovations relating to technologies, materials, processes and such like are introduced to create a new system design. The designers are asked to provide estimates of the probabilities associated with key failure modes of the new system design as part of a reliability analysis which in turn impacts the development budget. As in our first example, the designers are not entirely comfortable in expressing subjective probabilities. In part, this is because their mind-set implies designs are created to function not to fail hence thinking through negative outcomes is challenging. But also, because assessing probabilities arising from the myriad of scenarios across which uncertainty might be manifested is cognitively complex. Since the designers naturally match the functionality of the new system to analogous existing system designs we build upon this natural comparison to obtain our probability assessments of the failure of the new system to function as required. We take as our primitive for expert judgement the engineering relationship between the new and heritage system designs so that we can select relevant operational experience data from earlier generation products for the latter to obtain an empirical prior distribution for given failure modes of the former.

7.2.2 Generalisation of the Problem

Abstracting these two industry contexts allows us to establish three common features of our elicitation problem.

First, we consider situations in which we are effectively anticipating data about events that might be realised in future and for which there exists observed data for analogous entities. For example, the number of non-conformances in future

parts delivered by a new supplier or the number of failure events in the future operational use of the new system design. In each situation data are available for existing suppliers or systems, and a data set associated with the new supplier or system will become available, at least in principle if not also in reality.

Second, we can articulate a set of models to explain the variability in the anticipated data set. That is, the data set that comprises the future event history for the quantity of interest that does not yet exist but might be realised. This model family is indexed by parameters to describe the variability in the data generating process (DGP) associated with the event history. For our examples, a simple probability model for the DGP could be a Poisson distribution parameterised by the underlying true rate. For the supplier non-conformance and the new system development examples, the Poisson model describes the count of the non-conformances and the count of failure events per unit time parameterised by the true non-conformance rate and the true failure rate respectively. The true rate is not known with certainty, therefore we can represent the uncertainty in the parameter using a prior distribution if we follow a Bayesian approach.

Third, we require expert judgement to specify the prior probability distribution representing the uncertainty in the quantity of interest. For example, the prior distribution provides a set of plausible values representing the uncertainty about the true non-conformance rate of the new supplier or the true failure rate of the new system design. The challenge is to elicit a prior distribution so that it is meaningful and defendable, making appropriate use of expert judgement.

7.2.3 Implications of Inference Principles for Elicitation

If we approach elicitation from a Bayesian perspective then we are effectively asking an expert to map her beliefs about the quantity of interest onto a mechanism where the uncertainty is fully understood. This mechanism can be conceptualised by, for example, chips or a probability wheel (Spetzler and Stael von Holstein 1975), all of which translate to asking questions during elicitation to obtain an answer to a question such as 'what is the probability of a non-conforming part being delivered by the new supplier?'. The elicitation intends to encourage an expert to think about a self-consistent betting regime. Take a simple probability wheel conceptualisation, as shown in Fig. 7.1. If an expert states there is a 50% chance that the next part delivered by the new supplier is a non-conformance then we could map this outcome to the white or black implying that an expert is effectively mapping her belief as a bet she is willing to take onto a mechanism whose stochastic characteristics are fully known.

But what happens when an expert more naturally makes analogies to her experience related to, say, past suppliers based on an assessment of similarity between characteristics believed to be influencing quality performance. Based upon the evidence of achieved performance for similar suppliers for whom empirical data are available, we can construct a class of plausible non-conformance rates for the new supplier. In this situation, an expert is essentially forming a comparator data set

representing the extent of her knowledge about the uncertainty in the true rate. More abstractly, we can say the expert needs to assess the characteristics of a DGP for the non-conformance rate of the new supplier so that a comparator pool of DGPs for which empirical data already exists can be identified. We argue that this matching of the DGPs for the new and similar existing suppliers represents the extent to which we can make reliable use of expert judgement. Achieving a match implies that the probability distribution representing the variation in the comparator pool allows us to empirically estimate the prior distribution for the true non-nonconformance rate.

Theoretically, we reason that if the comparator pool reflects the beliefs of an expert then, as the number of DGPs within the pool increases, the empirical distribution characterising the uncertainty in the quantity of interest will converge to the subjective prior distribution obtained through mapping to a probability mechanism that is fully understood; see Fig. 7.1. Practically, of course, constraints are likely to exist on the amount of experience which can be accumulated by an expert meaning that an infinite pool is infeasible which in turn implies that we lack complete understanding of the probability mechanism. If expert judgement is based on finite pools, or equivalently incomplete experience, then this leads us to question the general adequacy of a prior distribution elicited solely using subjective expert judgement. To address the challenges of some practical contexts, such as those discussed in our motivating examples, we propose an alternative approach that aims to make use of an expert's judgement as well as relevant empirical data with the goal of eliciting a meaningful prior distribution for parameter uncertainty. Our proposed approach is grounded in the method of empirical Bayes inference.

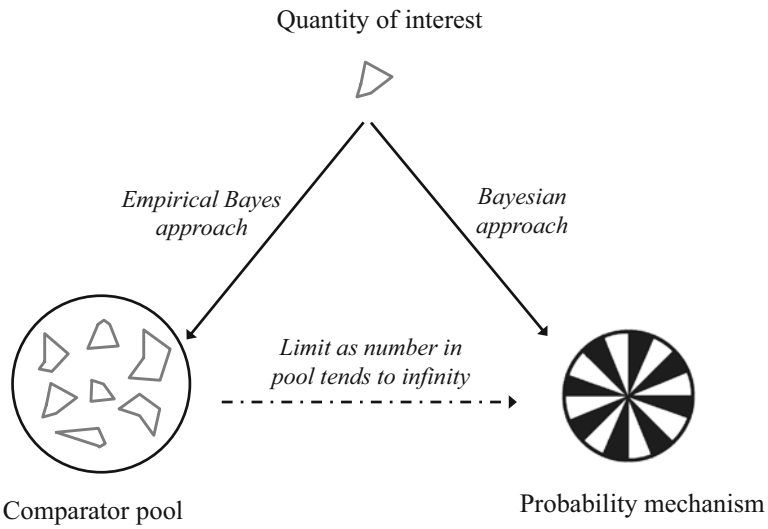


Fig. 7.1 Empirical Bayes and Bayesian reasoning for subjective probability elicitation

7.2.4 *Principles of Empirical Bayes Inference*

Figure 7.2 illustrates the concepts of empirical Bayes inference. Multiple data generating processes (i.e. the m DGPs) are required to form a comparator pool of data for the quantity of interest. Each DGP is described by a family of probability models for which empirical observations are available to support parameter estimation. We use the term family deliberately since the probability models are all of the same type (e.g. Poisson) but the parameter values of each distribution can differ to characterise the variation in each individual DGP. Importantly in our context, the empirical data across all DGPs are pooled to estimate the parameters of the prior distribution, which represents the variation in the comparator pool. For example, if the probability model family for the DGPs is a Poisson distribution parameterised by the non-conformance rate, then the empirical prior mean estimated by pooling data provides a point estimate of the true non-conformance rate of the new supplier while the full prior probability distribution characterises the uncertainty.

Although not the focus of this chapter, it is worth mentioning that Bayes theorem can be used to generate a posterior distribution by updating the prior distribution in light of empirical data for a given DGP, whether the DGP relates to the events for a new or an existing entity, such as a supplier. In general, the posterior estimate will be a weighted average of the comparator pool and the individual estimate, where the weighting depends on the degree of experience. Typically less weight is given to an individual and more weight to the pool for those DGP with limited histories, with greater weight given to an individual with more data.

In summary, empirical Bayes adopts the same basic steps as a Bayesian methodology by articulating a prior distribution and having the capability of updating

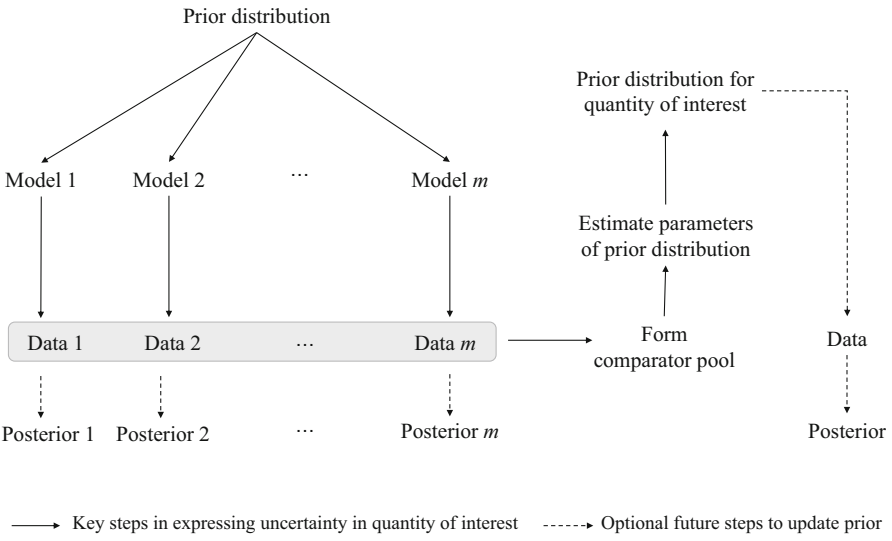


Fig. 7.2 Rationale of an empirical Bayes approach to obtain a prior distribution

this prior with data to generate a posterior distribution. The difference is that under empirical Bayes the prior distribution is estimated using observed data for a comparator pool while a full Bayes approach uses a subjective prior distribution. The roots of empirical Bayes reasoning can be traced to von Mises (1942), with Robbins (1955) formalising the terminology and providing the first serious study of the method within a non-parametric framework. Further details about empirical Bayes can be found in the seminal papers by Good (1965), Efron and Morris (1972), Efron and Morris (1973), Efron and Morris (1975), Efron et al. (2001). While Carlin and Louis (2000) as well as Efron (2012) provide introductory texts.

7.3 General Methodological Steps

We propose a five step approach to obtain the prior distribution using relevant empirical data, as shown in Fig. 7.3.

7.3.1 Characterise the Population DGP

We begin by identifying those factors characterising, what we call, the population DGP. This is the process generating the anticipated data or future events for the quantity of interest. This is an important step because it defines the criteria by which data sets (i.e. the sample DGPs) are subsequently selected for inclusion in the comparator pool used to construct the prior distribution.

The characterisation of the population DGP should be driven by problem domain experts, suitably facilitated by an analyst. An expert has an important role in this step because it is the expert who possesses substantial accumulated understanding of what is likely to influence the realisation of events and, with the support of the analyst, articulates the factors to provide the basis for similarity matching.

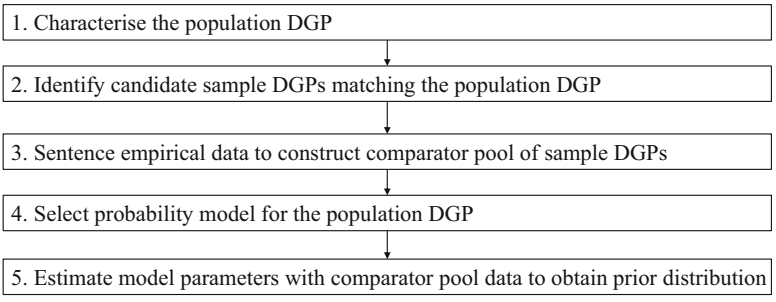


Fig. 7.3 Key steps in constructing a prior distribution using empirical data

The way in which we might approach characterisation of the population DGP can be considered a specific instance of the ideas inherent within the wider context of statistical sampling. Much has been written in the literature on, for example, survey sampling about the need to identify appropriate factors to define population characteristics to support sound inference and to share insights into how such factors might be identified within applications. See, for example, Cochran (1975).

There has also been consideration of this issue within the context of the so-called reference class problem (Reichenbach 1971). This is concerned with classifying an event such that appropriate data can be used to infer probability. According to the Oxford English dictionary a reference class within the context of probability theory and the philosophy of science is *"the class of entities sharing a property with respect to which a theory or a statement of probability is framed"*. Cheng (2009) has examined the challenges associated with identifying reference classes in legal practice where it is acknowledged that a finite number of possible (i.e. sample) DGPs exist for a given case and a key question is *"how does one choose the comparison group?"*. Although the paper is framed as inference within an adversarial context, with perhaps information asymmetry between the two opponents, many points raised have more general currency providing examples of defining the population characteristics, including some where an apparent lack of consideration of the appropriate factors to define the population DGP has resulted in misleading inference.

7.3.2 Identify Candidate Sample DGPs Matching Population

The argument underpinning the approach proposed by Cheng (2009) is that *"reference class-style reasoning is equivalent to using a highly simplified form of regression modelling"* where the factors characterising the population DGP can be switched on/off for candidate data sets effectively providing a means of making a relative assessment of relevance against a set of criteria. Cheng (2009) also points out that in practice the goal is not to find the optimal class but simply to find the best available data sets to make reasonable and timely inference. This is an important point in relation to our purpose since we are also likely to be constrained by the availability of a finite number candidate data sets. Also, unlike other types of statistical sampling, we are not in a position to collect primary data to match the characteristics of the target population. Instead we are matching the characteristics of the population DGP with secondary sources of data that already exist. Hence a formal means of matching the similarity between the population and the candidate sample DGPs (i.e. available data sets) in terms of the factors characterising the former is important and necessary.

7.3.3 Sentence Empirical Data to Construct Sample DGPs

Once the preferred data sources have been selected, the events recorded should be scrutinised in collaboration with a domain expert to assess the representativeness of the records for the type of experience to which our anticipated DGP will be exposed and to sentence these records, if required. The intention is to create a data set that is not only appropriate in terms of its similarity matching to the population characteristics but is also relevant in terms of the events and the circumstances under which these have been realised. The nature of sentencing can vary with application contexts and the associated modelling. For example, sentencing can include selecting records for events within a sub-set of the realised data set to form a representation of the anticipated experience or simply to screen out events that have been realised under unusual circumstances that are not representative.

More formally, we can reason through an assessment between the population and sample DGPs as follows. Although the data sets formed to create candidate sample DGPs are heterogeneous with respect to their stochastic characteristics (e.g. means and standard deviations), the expert should not be able to meaningfully discriminate between these DGPs based on any information other than their realisations. Care must be taken with the data analysis since the realisations within any DGP will be correlated as belonging to the same DGP, but the sets of event data records between DGPs are assumed independent. Confirming the suitability of the data records essentially requires checking that the predictive distributions for each DGP are independent and identically distributed. This can be achieved by conceptualising as a comparison of order statistics. Let ${}_jX_{i:n}$ denote the ith smallest value from a sample of n records from the jth DGP where $j = 0$ denotes the DGP associated with the quantity of interest for which an estimate of uncertainty is to be made. The comparator pool of sample DGPs will be appropriate, if an expert can confirm that based on the covariate information only the following statement is true:

$$\Pr\left(Min\left({}_0X_{i:n}, {}_1X_{i:n}, \ldots, {}_mX_{i:n}\right) = {}_jX_{i:n}\right) = \Pr\left(Min\left({}_0X_{i:n}, {}_1X_{i:n}, \ldots, {}_mX_{i:n}\right) = {}_kX_{i:n}\right), \quad \forall i, j, k, n.$$

In words, based on the reference factors used to characterise the population DGP, the minimum of any order statistic is equally likely to be generated from any of the sample DGPs; this is true for all order statistics and for all possible sample sizes. Practically this implies that when assessing the order statistics an expert may simply reflect upon whether the extremes and the typical values of the comparator data sets are appropriate for the quantity of interest.

7.3.4 Select Probability Model for Population DGP

The family of probability models considered suitable for describing the population DGP will be largely determined by the context so that it supports suitable inference, not only in terms of meaningful parameters but also in terms of mathematical and computational implementation.

For the two motivating examples, we indicated that the Poisson distribution is an appropriate simple model to describe the variation in the count of events and hence capture the aleatory uncertainty as the within-process variation. Since the prior probability distribution predicts the epistemic uncertainty in the true rate then choosing a conjugate parametric form leads naturally to the Gamma distribution to model the between-process variation across the comparator pool of sample DGPs.

Therefore it is important to select a model family that allows coherent representation of the variation both within and between the DGPs to articulate both the aleatory and epistemic uncertainties, even though it is the latter that is of primary interest to us in the elicitation context.

7.3.5 Estimate Model Parameters to Obtain Prior Distribution

Statistical inference to estimate the parameters of the model for the population DGP can be conducted using standard approaches such as Maximum Likelihood or Method of Moments (e.g. Klugman et al. 2012). The mathematics of the inferential procedure will depend upon the parametric form of the probability models. For example, Quigley et al. (2007) provides mathematical details of the statistical inference methods for the Poisson-Gamma model family.

The parameter estimates obtained using the data in the comparator pool formed from the sample DGPs allow the prior probability distribution to be fully specified.

7.4 Example Applications of the Elicitation Process

Two examples are presented. Both relate to industrial applications of risk and reliability analysis for which the quantity of interest relates to the frequency of events over time. We have deliberately selected examples where related probability models are chosen for the population DGP since it allows us to show how different application considerations give rise to adaptation and customisation of the general methodology. Each example presents distinct challenges in relation to the characterisation of the population DGP, the identification and sentencing of empirical data to create sample DGPs, and the method selected to estimate model parameters for the given the probability models. We present the examples in order of their relative complexity of the emergent elicitation issues. For this reason we focus our discussion on the distinctive elements of each example even though the elicitation for both examples did require careful consideration of each step.

7.4.1 Assessing Uncertainty in Supplier Quality

A project aimed to model the risk in supplier performance for a manufacturer of complex, highly engineered systems reliant on an extensive, international supply chain for parts and sub-assemblies. The modelling problem under consideration involved supporting decisions about whether, or not, to develop a supplier given only information gained about quality from company standard contracting and procurement processes. Quigley et al. (2018) describe the wider modelling methodology and results. Here we focus upon the elicitation of the subjective distribution representing the uncertainty in the quality performance of the new supplier, where quality is measured by the true non-conformance rate associated with parts delivered from the supplier to the manufacturer.

7.4.1.1 Characterise the Population DGP

To characterise the population DGP, we need to identify the reference factors in partnership with a suitably qualified expert. Taking an expert to be a person(s) with substantive experience in relation to the event for which uncertainty is to be assessed, then the natural set of experts for this problem are those staff within the manufacturing company with qualifications and experience in managing the supply chain and production operations.

As is common more generally (e.g. Slack et al. 2016), the manufacturer organises its parts supply base into coherent commodity groups each of which correspond to classes of technologies and processes. Such a classification allows managers to share the responsibility for the procurement and development of a set of suppliers within a given market. Importantly, it also implies that the manufacturer has already considered classification of parts in terms of common factors that are likely to influence the nature of the functional specification and hence the opportunity to conform (or not) with that requirement as a consequence of the type of part being supplied.

Much has been written about the types of factors affecting supplier quality and the risks associated with supply chain performance (Nagurney and Li 2016; Sodhi and Tang 2012; Talluri et al. 2010; Zhu et al. 2007). Hence secondary information about the possible types of factors which might influence the new supplier quality performance is available to the analyst leading the elicitation. Such information can be useful in preparing to elicit those factors which are considered by an expert to be influential for the case under consideration.

So who is our expert and how do we identify the factors believed to be important in characterising the population DGP? Our expert is a supply chain manager who possesses the experience of the day-to-day management of the supply base within the company as well as wider expertise in managing operations in similar organisational contexts. In this sense our expert is suitably qualified to share his knowledge and experience during the elicitation. Through multiple conversations

taking the form of semi-structured interviews, supported where appropriate with diagramming techniques, we have surfaced the expert's beliefs about influencing factors and the relationships between them. Factors identified include the nature of the part technology, design, production and shipping, including type, complexity and scale, as well as the nature of the supplier experience, capability, capacity and location.

7.4.1.2 Identify Candidate Sample DGPs Matching Population

Next we identify empirical data sets in terms of their match to the population characteristics as defined by the reference set of influential factors.

To manage operations, the manufacturer has databases containing empirical records associated with supplier and part details as well as their transactional data for events related to the placing and receiving of parts ordered for engineering projects, including the quality of parts received at goods inwards. More generally, such databases or enterprise resource planning (ERP) systems are core to managing operations (Gallien et al. 2015). They can be extensive both spatially, in terms of part/supplier coverage, as well as temporally, given the dynamic nature and scale of manufacturing production. This means that in terms of matching and subsequent sentencing of empirical data sets, we need to consider the records to be used in terms of both coverage of 'similar' event histories for suppliers and also the relevant time window in order to obtain a reliable predictor of the uncertainty in the true non-conformance rate of the new supplier.

In our application, our choices about possible matches to the population DGP includes event history data for a super-set of all suppliers, a set of suppliers within the commodity group to which the new supplier belongs, a sub-set of this commodity group defined by those suppliers/parts possessing common identifiable factors. We have used the commodity group data as the basis for our candidate samples from the population DGP because this best matches those factors believed by our expert to most influence the supplied part quality. The commodity grouping confounds the influential effects of part technology and processes on the opportunity to deviate from conforming to functional specification. We discounted the other two alternatives mentioned for the following reasons. Using a super-set of all suppliers mixes multiple groupings each with different degrees of opportunity and so would tend to overestimate the uncertainty in the true non-conformance rate of the new supplier. Using a sub-set of suppliers within the commodity group might underestimate the uncertainty since the reduction could only be based on recorded factors such as geographical location, which experts judge to be less influential than other factors such as supplier production capacity and loading which are not directly observable.

While our decision to select particular data sets has been based upon the judgement elicited from the domain expertise of the supply chain manager, we have also been able to explore the degree of historical influence of certain recorded factors on the variation in the observed non-conformance rate of existing suppliers using,

for example, regression modelling. Although not an exhaustive analysis since the covariate information is incomplete, the findings of such data analysis can help us to challenge and to elicit judgements from an expert.

The choices we make in selecting data sets will ultimately affect the number of sample DGPs we use to estimate the prior distribution. For example, taking the commodity group of 35 suppliers as a baseline, then by definition there will be more (less) candidate sample DGPs in the super-set of all suppliers (sub-set of the commodity group). Obviously, the number of sample DGPs, as well as the amount of data in each, will impact the degree of sampling error and hence inference.

7.4.1.3 Sentence Empirical Data to Construct Sample DGPs

Having selected the candidate data associated with the existing suppliers, we now require to finalise the set of event records for each supplier in order to form the comparator pool of sample DGPs to be used for inference.

Two types of data sentencing are needed. First, to choose the relevant records from past event data. Second, to cleanse the selected records to deal appropriately with any anomalies whether they arise due to data recording errors or unusual circumstances affecting the suppliers. The latter is standard statistical data preparation, therefore we focus discussion on the former.

Since the purpose of selecting the data records is to form a distribution representing the uncertainty in the unknown true non-conformance rate, we need to consider historical events for existing suppliers only insofar as they are likely to be reliable predictors of the future for the new supplier. Hence again expert judgement will be vital in assessing the relevance of choosing data from different time horizons. In our application, data are recorded daily but management reports use summaries on weekly, monthly, annual windows associated with different purposes bringing a tendency for the expert to anchor upon conventional time frames. Given the length of our engineering procurement projects, which last several years during which there is turnover in the supplier base, we elect to use time windows defined on annual basis on our initial sentencing of the data. Of course, there can be a tension between the relevance of the time windows selected and sample size given that focussing on the recent past implies a shorter sampling history than had we chosen a longer time horizon. However this is a trade-off that needs to be made since relevance of the selected events over time is preferred to simply more event data per se.

In our application we agree upon a data set to represent the sample DGPs that includes the number of non-conformances over the specified annual time intervals for 35 similar suppliers. Although not shown in its raw form, there is a degree of heterogeneity in the data from the comparator pool and this is used to capture the distribution from which the new supplier's 'future experience' can be considered to be randomly selected.

In assessing the order statistics between the DGP associated with the true non-conformance rate of the new supplier and the candidate sample DGPs for the existing suppliers, it is sufficient to assess whether the minimum rate for the new

supplier is equally likely to be from any of the existing suppliers, if we are assuming a Poisson-Gamma probability model. However, to assess the parametric distributional assumptions requires the expert to reflect upon the order statistics more fully as described in Sect. 7.3.3. For example, if an expert identifies that one supplier is much more volatile than another but each have similar median performances, then this would indicate the distributional assumptions are in question.

7.4.1.4 Select Probability Model for Population DGP

We use a Poisson-Gamma probability model because it provides a flexible family capable of representing a wide class of patterns of uncertainty and, as a conjugate of the Poisson, computations are easily supported (Carlin and Louis 2000). Given the prior is estimated empirically it is also possible to check the statistical fit of this assumed model family by, for example, comparison of the observed and expected percentiles of the fitted predictive distribution.

Figure 7.4 shows an annotated version of the empirical Bayes approach, originally given in Fig. 7.2, for this supplier non-conformance rate application.

More formally, denote the number of non-conformances $N_i(t_i)$ accumulated by time t_i for the ith supplier to be conditionally independently Poisson distributed with mean $\lambda_i t_i$. We follow an empirical Bayes methodology, whereby a two stage hierarchical model is assumed, such that the rate for each supplier, $\Lambda_i, i = 1, 2, \ldots m$, is treated as though independent and identically distributed (i.i.d.) from a continuous prior distribution, the form of which is assumed to be Gamma with shape parameter α and scale parameter β:

$$\Lambda_i \overset{i.i.d}{\sim} G(\alpha, \beta)$$

$$N_i|\Lambda_i = \lambda_i \overset{indep}{\sim} Po(\lambda_i t_i).$$

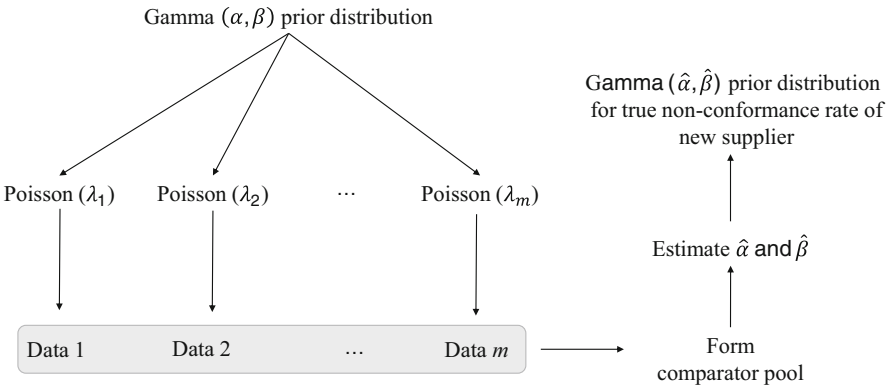


Fig. 7.4 Empirical Bayes reasoning for the Poisson-Gamma probability model for supplier non-conformance rate

7.4.1.5 Estimate Model Parameters to Obtain Prior Distribution

The parameters of the prior distribution, α and β, are estimated using the empirical data in the sample DGPs by calculating the predictive distribution, which then forms the basis for the likelihood function for the model. For our Poisson-Gamma model the predictive distribution takes the form of the Negative Binomial distribution (Greenwood and Yule 1920):

$$P\left(N_i\left(t_i\right)=n_i|\,\alpha,\beta\right)=\int_0^{\infty}\frac{\left(\lambda_i t_i\right)^{n_i}e^{-\lambda_i t_i}}{n_i!}\frac{\beta^{\alpha}\lambda_i{}^{\alpha-1}e^{-\beta\lambda_i}}{\Gamma(\alpha)}d\lambda$$
$$=\frac{\Gamma(n_i+\alpha)}{\Gamma(\alpha)n_i!}\left(\frac{\beta}{\beta+t_i}\right)^{\alpha}\left(\frac{t_i}{\beta+t_i}\right)^{n_i},\quad \alpha>0,\ \beta>0,\ t_i>0,\ n_i=0,1,2,\ldots$$

Following Arnold (1990), a likelihood function for the data can be constructed by taking the product of the predictive probability functions for the *i*th supplier evaluated at each of the associated realisations of non-conformance events for that supplier:

$$L\left(\alpha,\beta\right)=\prod_{i=1}^{m}\frac{\Gamma\left(n_i+\alpha\right)}{\Gamma\left(\alpha\right)n_i!}\left(\frac{\beta}{\beta+t_i}\right)^{\alpha}\left(\frac{t_i}{\beta+t_i}\right)^{n_i}.$$

Thus the Type 2 (Good 1976) Maximum Likelihood Estimators (MLE) of the pool parameters, denoted by $(\hat{\alpha},\hat{\beta})$, can be obtained as confidence regions for the parameters.

Figure 7.5 shows the form of the Gamma prior Probability Density Function (PDF) obtained for the data in our comparator pool of sample DGPs and an

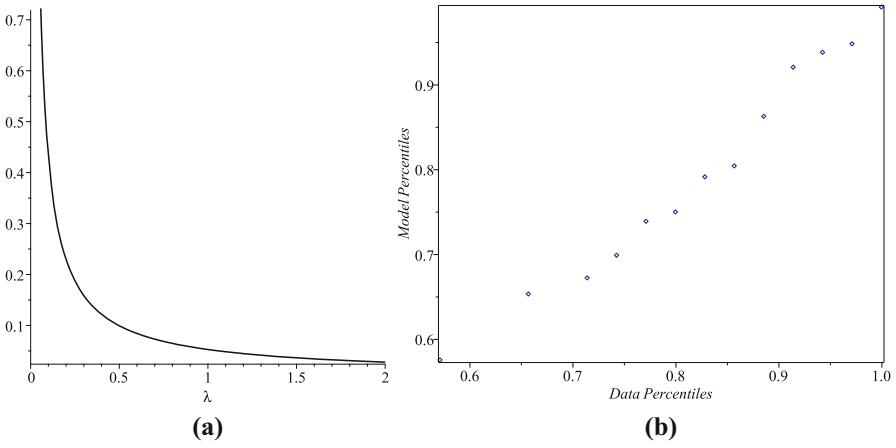


Fig. 7.5 (**a**) Estimated prior PDF for λ, the true non-conformance rate of the new supplier and (**b**) Fit of the Poisson-Gamma model to comparator pool data based on the predictive distribution and empirical percentiles

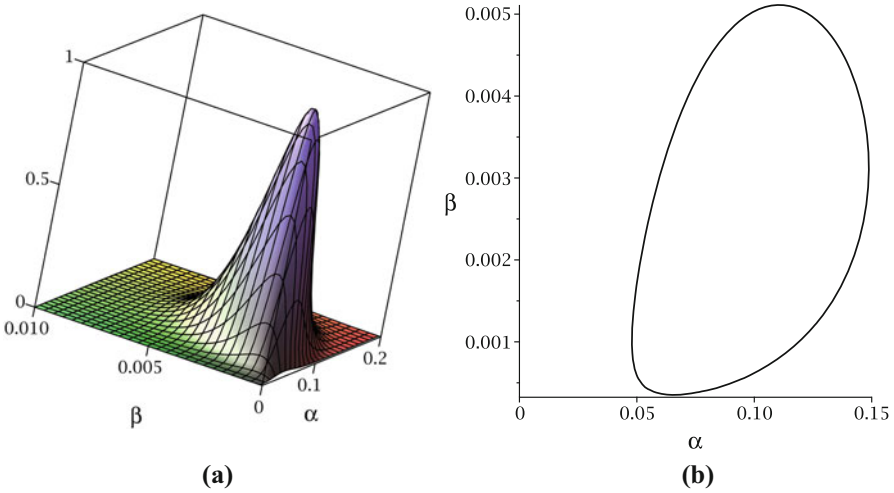


Fig. 7.6 (**a**) Relative Likelihood function and (**b**) 95% confidence region for pool parameters

assessment of the statistical adequacy of the probability model for the data used. The prior distribution in Fig. 7.5a has a long right tail with a prior mean of nearly 50 non-conformances per unit time. The plot of observed and expected percentiles based on the predictive distribution in Fig. 7.5b indicates a reasonable fit of the Poisson-Gamma model to the empirical data given that the points fluctuate around the 45° line. Thus the data used is consistent with the probability model selected for the population DGP.

Figure 7.6 illustrates the relative likelihood function and the associated 95% confidence region for the pool parameters estimated from the 35 sample DGPs. In Fig. 7.6a the peak corresponds to the Maximum Likelihood Estimates (MLE) and is assigned a value of one from which the likelihood of any combination of parameter values are measured relatively. Figure 7.6b shows that, based on the 95% confidence region, α is between 0.048 and 0.148 while β is between 0.00035 and 0.00511. Moreover, the parameter estimates are not independent since some the coverage of some pairings are not within the confidence region, in particular the high values of α coupled with the low values of β.

Figure 7.7 provides a pointwise 95% tolerance interval for the prior Cumulative Distribution Function (CDF). The long right tail of the distribution is evidenced by the steep climb of the CDF followed the relatively flat growth. The MLE of the CDF provides an estimate of the probability that the true rate, λ, is less than a specified value. For example, although not discernible in the plot, there is a 0.4 probability that λ will be less than 0.01. More apparently, as λ increases to 0.1 and again to 1 then the corresponding cumulative probabilities rise to 0.49 and 0.6, respectively.

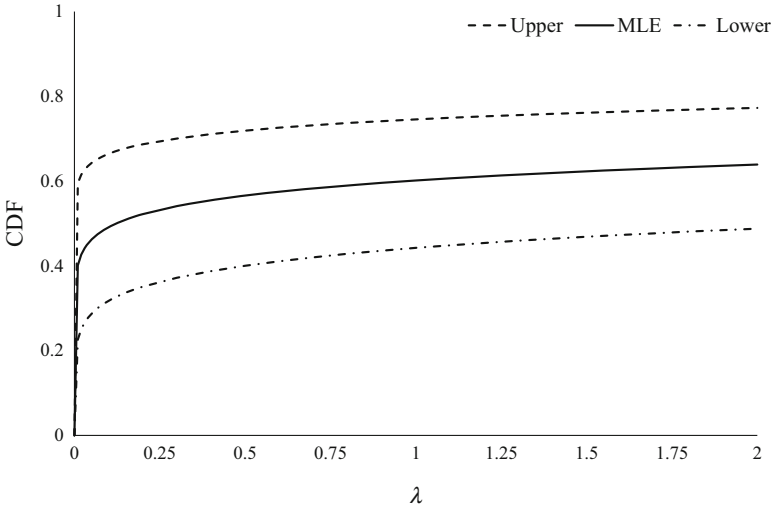


Fig. 7.7 Piecewise 95% tolerance interval for the CDF of new supplier true non-conformance rate

More interesting is the width of the tolerance intervals. For example, over the range of λ values between 0.01 and 1 the width remains relatively constant between 0.36 to 0.30, respectively; that is, approximately one third. This has implications for the degree of uncertainty in the prior distribution. In this example, our tolerance intervals allow us to acknowledge the uncertainty in the prior distribution estimated from the 35 sample DGPs in the comparator pool.

Although the prior distribution is estimated empirically using the data selected by expert judgement, we also require the expert to assess whether the estimated prior distribution adequately represents his beliefs about the uncertainty in the true non-conformance rate of the new supplier. In our application we make such an assessment by providing the expert with visual feedback, say, in the form of the prior distribution plot. Since the methodology has been developed for estimating the prior parameters from data, it is also relatively easy to provide alternative prior distributions based on different selection of data, as might correspond to different matching of sample DGPs from the population. For example, we have shown the findings based on selection of a pool of suppliers from the same commodity group. But it is also possible to generate equivalent plots using, say, the super-set of all suppliers or a sub-set of commodity group suppliers to present an expert with alternative prior distributions representing different degrees and patterns of uncertainty. Such a tactic provides a form of internal consistency checking between the explicit reasoning about the influential factors affecting the uncertainty in the non-conformance rate and the representation of these beliefs in the form of a prior probability distribution.

7.4.2 Assessing Uncertainty About Reliability of an Engineering Design

Our second example is based on a project to model the reliability of an engineered unit during its design and development phase. The unit will be part of a new generation aircraft. The ultimate purpose of modelling is to support decisions about the efficient allocation of resources to grow the reliability performance of the unit design to meet its required specification (Walls and Quigley 1999; Walls et al. 2006; Johnston et al. 2006; Wilson and Quigley 2016). The modelling approach adopted requires elicitation of the sources of uncertainty regarding any design weaknesses and the time to their realisation as failures if not removed or mitigated. The design under consideration is a variant of an established product family and so the manufacturer has extensive operational data on performance of earlier generations of the unit type. Such data contains information about all life events for each unit within a fleet, including entry into service, failure and maintenance events.

In order to identify possible weaknesses of the new unit design, structured expert judgement is elicited from relevant engineers to both express their concerns and to quantify the uncertainties about the existence of these concerns as subjective probabilities. The process supporting this subjective elicitation is given in Walls and Quigley (2001) while reflections on the practice of implementation are given in Hodge et al. (2001). Specifically for this example, a representative selection of thirty engineers have been interviewed, including designers, programme managers, as well as specialists in components, environmental test, procurement, and manufacture. These engineers have identified their concerns and assessed their chance of occurrence in system operation resulting in a subjective Poisson prior distribution with means ranging from approximately 3–11 across different classes of engineering concern.

Our focus in this chapter is upon the expression of an empirical prior distribution for the epistemic uncertainty associated with the time to realisation of engineering concerns as failures which can be estimated from relevant observational data from variants of the unit design already in service. A different expert to those involved in sharing engineering judgement about the nature of concerns is involved in providing judgement about the selection of the empirical data to be used to model the failure occurrences within specified time intervals. The expert working with the analyst to develop the empirical priors assumes a more systems level view of the new unit than those engineers who had provided judgements about the nature of epistemic uncertainties in relation to the concerns about the new design. The expert assuming the role in empirical data preparation is an experienced technical engineer with a breadth and depth of experience of the product family. Earlier he supported the facilitation of the subjective judgement from domain experts about design concerns from a systems perspective and so provides a link in interpreting the engineering detail about specific design issues with the observational data available for product families.

Quigley and Walls (2011) describe the full methodology for combining the subjective prior distribution on engineering concerns with empirical prior distributions to support reliability growth decision making. Here we consider the application steps in constructing the prior distribution only.

7.4.2.1 Characterise the Population DGP

The nature of the engineering concerns are pivotal to the characterisation of the population DGP since these concerns capture the potential for types of failure to occur due to a mismatch between the conceptual design 'strength' and the 'stresses' to which it will be exposed. Engineering concerns may relate to, for example, choices about electronic component rating, material characteristics, manufacturing processes, topology and so on. More generally concerns relate to aspects of the design, manufacture, operation and maintenance where opportunities for stressors to challenge the intended functionality of the unit might arise.

There are, of course, more tangible factors that might characterise the population DGP in the form of the specified requirements of the unit design. Such requirements will articulate the function, environment, duration as well as other influential features of the design specification. It is based on such factors that design engineers might select a base design from an existing product family in order to develop a new variant (Pahl and Beitz 2013). While such factors can also aid characterisation of the population DGP, it might be too naïve to consider them only since they effectively represent the factors that drive the choices of the designers in developing a new unit. It is the consequences of these design and other choices in engineering the unit that give rise to concerns.

In essence, the concerns represent the epistemic uncertainties of the engineers about the ability of the new design to function as intended in its operational environment. The nature of how the concerns will be realised as failures provides a means of characterising a sub-population DGP which is needed because each type of concern will be associated with a distinct pattern of realisation. For example, if the electronic components are insufficient for the operating stresses then this concern is likely to be realised early in service as a form of shock failure, whereas material characteristics may imply a faster rate of degradation than intended, resulting in a failure later in service but before the anticipated lifetime of the unit and so sooner than desired.

The elicitation of engineering concerns is very important because it allows us to understand the possible effects of failures that might be realised due to the causal reasoning from design choices through to operational functioning. This understanding allows us to define the reference factors relevant to each class of concern in terms of their temporal realisation as failures and so specify the characteristics of the population DGP at a sub-population level.

7.4.2.2 Identify Candidate Sample DGP Matching Population

As mentioned, the company has operational data on life events for related products within the unit family. For earlier generations of the unit design, no elicitation of engineering concerns had been formally conducted although other forms of reliability analysis are available which provide insight into anticipated failure modes and why these did or might have occurred. To identify our candidate data sets we need to consider the concerns elicited for the new design and the equivalent data for past heritage designs given the relative similarity between design variants in terms of the consequences of the choices about externalities of function, form and environment so that we understand the relative opportunities for vulnerabilities to exist and to be experienced.

In our application, we identify several existing unit designs for which there are data sets offering candidate sample DGPs. However, there is not a one-to-one similarity match between the full set of concerns, and the reasons for these concerns, between the new unit design and the existing units. This is not unexpected given we have characterised the population DGP at the sub-population level. In this context, the sources and coverage of our data sets for the candidate sample DGPs can vary for different sub-populations depending on the classification of the engineering concerns.

Figure 7.8 summarises the principles underpinning the formation of the sample DGP for this example. Each concern class in the centre of the diagram represents a sub-population DGP defined in terms of common reference factor settings for the engineering concerns. The links between the individual concerns and the classes

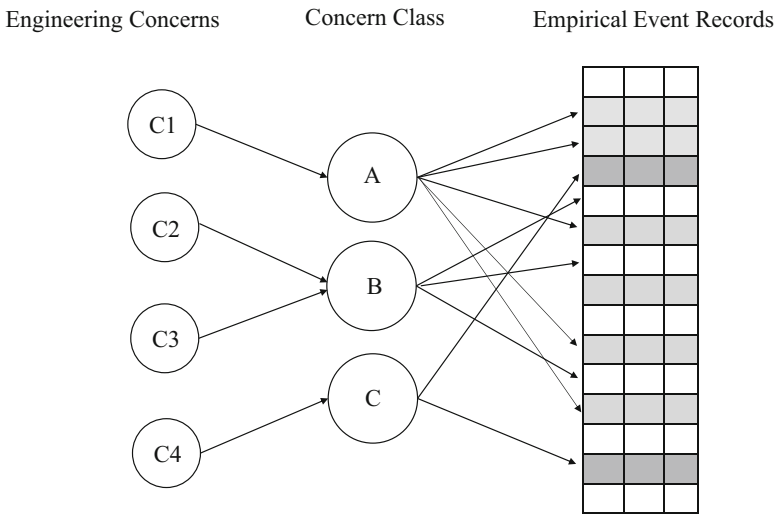


Fig. 7.8 Conceptual relationship between engineering concerns for new unit design and event history records for related designs

represent the mapping between the engineering judgement and their expression as reference factors for the similarity matching with heritage unit designs which have empirical data. The data records shown on the right side are in the form of event histories with rows corresponding to accumulated time to an observed failure and columns giving covariate information, such as heritage unit and failure mode. The shading of the records indicates the distinct data sources selected for different heritage units that are candidate sample DGPs. The links between the concern classes and individual events represents the failure data records to be included in the sample DGPs. If a class contains only one concern for which there is a match between the relevant reference factors and the observed failure event codings, then this can be conceptualised by mappings such as C1 $\rightarrow$ A and C4 $\rightarrow$ C. A class might be formed if individual concerns can be meaningfully grouped in terms of their likely pattern of realisations through time, as shown by the mapping C2 $\cup$ C3 $\rightarrow$ B. For example, specific unit build vulnerabilities can be grouped together if the pattern of realisation of the resulting shock failures due to manufacturing issues are assessed by the engineering experts to be the same.

7.4.2.3 Sentence Empirical Data to Construct Sample DGPs

So far we have built up our argument in terms of defining our population DGP in order to match suitable sample data sets. For this reason we show the mapping from concerns to classes to empirical records in Fig. 7.8. However, as we acknowledged earlier, sentencing data is a craft built upon scientific principles, hence we also need to take into account the state of the empirical data sets into consideration during the process of constructing the sample DGPs.

In our example, the empirical data sets included the individual unit reference, accumulated flight hours, date stamp, number of flight cycles, type of aircraft, operator, fault type, failure mode code, failure effect code, text description of event occurrence, amongst others. Having data that describes the context, the nature and a classification of that event is not atypical in a reliability engineering context (e.g. Cooke 1996). In particular, the classification of events is embedded in the manner in which much historical failure event data has been stored and shared both within organisations and at industry sector levels (e.g. Rausand and Hoyland 2004). Although it can be convenient to use the standard classes within the empirical data set to define the classes of engineering concerns, we urge caution in simply automatically back-fitting. It is important to define classes grounded in the nature of the engineering concerns for the new design for which the prior distribution, and ultimately the reliability, are assessed. Even with historical data, rich information can be found in event descriptions to form sets of records that match to appropriate classes, which might be a sub-set of the standard grouping of events. As an example, consider a situation with two distinctive engineering concerns articulated in relation to some electronic components in the unit. One concern might relate to the geometry of one component's siting and another concern might relate to the material properties of another component. These concerns, should they exist as real

problems, are reasoned to be realised in different ways since the former will be likely to occur more quickly as it will be vulnerable to operating stresses within a flight cycle, while the latter might be realised more slowly since events are more likely to occur as experience is accumulated between flight cycles. The empirical data set categorises all events related to the electronic components together and so mixes the time to failure distributions that relate to the concerns. If sufficient information is available from the textual description then the records within the electronics components categories can be partitioned into more appropriate classes that better match the population characteristics of the concerns.

It is possible that using empirical records from past units to assess the times to failure of some engineering concerns is judged to be inappropriate by the engineering experts. This might occur when there are novel aspects of the design for which reasoning through the physical science of the failure mechanisms might provide a better assessment of uncertainties. Within the context of probabilistic risk analysis for engineering design, Fragola (1996) introduces the notion of *"tolerance uncertainty"* which relates to this issue. Tolerance uncertainty corresponds to an engineering expression of the relevance of historical failure data in relation to an anticipated failure mode for a new design so that credible choices are made about the selection of relevant data for analysis. Following this logic, we are effectively arguing that if the empirical data on observed events for related designs are judged by the engineering experts to be tolerable assessments of the anticipated occurrences of failures due to an engineering concern for a new design then the empirical data can be selected to form the prior distribution. However, empirical data should not be used if it is judged by the engineering experts to be intolerable since this implies an alternative source, such as subjective assessments of uncertainty based on understanding of the underlying science supplemented by engineering analysis and test data, are arguably more justifiable.

Focusing upon the use of empirical data only, then like our first example, choices also need to be made about issues relating to the boundaries of data in terms of time and coverage as well as treatment of data anomalies. In this example we need to consider the inclusion or exclusion of data from particular units within the fleet for the existing design that is to be used to inform the prior. Some units might be spares and so experience long periods in storage followed by short periods of intensive use and so have unusual operational profiles compared with the majority of units which will be operated on aircraft in very similar flight patterns. Also, choices need to be made about the time windows over which empirical data will be selected. In this context there will be considerable relative stability for long periods given the nature of the certification and operational use of aircraft, however there can be scheduled upgrades which roll out part design changes across the fleet and so should be taken into consideration if it affects particular engineering concerns.

For our example, we have used an empirical source data relating to over 400 heritage units and extracted records relating to events occurring over several years. For this stage of the modelling we work closely with the engineering expert who possesses the expertise and extensive experience in the design process and technology together with the responsibility for managing the reliability development

programme. The empirical data selection and sentencing is led by the analysts who drive the methodological approach but the choices made are based upon the judgement of our expert. Ultimately we have created a data set containing the times to first occurrence of events within each of eight classes relating to the engineering concerns surfaced.

We also partition the operating time horizon into five intervals with natural breakpoints corresponding to the accumulated flying hours at nominal inspection periods associated with units of different ages. This choice was made for a combination of engineering and modelling reasons. The engineers are most interested in the likelihood of failures occurring during stages of a unit life, while the analysts are thinking ahead to candidate probability models which will be consistent with the data and the wider purpose of analysis. Further, since we have partitioned time into five mutually exclusive intervals, the expert has to assess the equivalence of the probabilities for each interval as a means of operationalising the assessment of the equivalency of the order statistic distributions for each DGP in the comparator pool, as described in Sect. 7.3.3.

7.4.2.4 Select Probability Model for Population DGP

The reliability in this example is taken to be a measure of the duration of unit failure free operating time and is parameterised by both the engineering concerns and their time to realisation. More formally we can write this as follows. Let J denote the number of concern classes, N_j represent the number of concerns in class j that will be realised as failures and let I denote the number of mutually exclusive and exhaustive partitions of the distribution of times to realisation of concerns. Then the prior distribution is sought on the $(I \times J)$ matrix, denoted by $\underline{P}$, whose (i, j) element, denoted by p_{ij}, represents the probability that a concern associated with class j will be realised in the ith epoch. Hence the probability that a unit will not fail by time t_0, denoted by T_u, conditioned on the matrix $\underline{P}$, and the vector $\underline{N} = \{N_i, \ldots N_J\}$ is given by:

$$P\left(T_u > t_0 \,|\underline{N}, \underline{P}\right) = \prod_{j=1}^{J} \left(1 - \sum_{i=1}^{t_0} p_{ij}\right)^{N_j}.$$

A multinomial distribution provides a simple and reasonable model to describe the sampling variation in the number of failures within time partitions of the concern classes. Each interval is assigned a parameter to measure the chance that a failure arising due to a concern would be realised in that time interval and the set of probabilities for any failure class are constrained to lie within a simplex. Further, the vectors of probabilities across classes are assumed to be independent and be Dirichlet distributed. We seek the empirical prior on these Dirichlet distributions; one for each class.

7.4.2.5 Estimate Model Parameters to Obtain Prior Distribution

A likelihood function to obtain Type 2 MLE for a concern class can be derived by first taking the product of all multinomial distributions for each sample DGP in the comparator pool and subsequently taking the expectation with respect to the Dirichlet prior.

Let m_{ik} denote the observed number of failures realised in time period i from the kth sample DGP created after sentencing the relevant historical data and let $\underline{M}$ denote the corresponding matrix of data. The likelihood function for the kth DGP, which is a function of the vector $\underline{P_k} = (p_{1k}, \ldots p_{Ik})$, can be expressed as:

$$L_k(\underline{P_k}) = \begin{pmatrix} \sum_{i=1}^{I} m_{ik} \\ m_{1k,} \ldots, m_{1k} \end{pmatrix} \prod_{i=1}^{I} p_{ik}^{m_{ik}}.$$

Following Ng et al. (2011), we assume the conjugate prior of the multinomial distribution to be the Dirichlet distribution of the form:

$$\pi\left(p_{1,} \ldots, p_I\right) = \frac{\Gamma\left(\sum_{i=1}^{I} a_i\right)}{\prod_{i=1}^{I} \Gamma\left(a_i\right)} \prod_{i=1}^{I} p_i^{a_i - 1}, p_i \geq 0, \sum_{i=1}^{I} p_i = 1,\ a_i > 0.$$

By taking the expectation of the likelihood equation with respect to the Dirichlet prior distribution, the new likelihood is obtained as a function of the parameters in the prior distribution and is given by:

$$L\left(a_{1,} \ldots, a_I\right) = \prod_{k=1}^{K} \frac{\Gamma\left(\sum_{i=1}^{I} a_i\right)}{\prod_{i=1}^{I} \Gamma\left(a_i\right)} \cdot \frac{\Gamma\left(a_i + m_{ik}\right)}{\Gamma\left(\sum_{i=1}^{I} a_i + m_{ik}\right)},\ a_i > 0$$

from which the Type 2 maximum likelihood estimates (MLE) of the a_i can be obtained.

Table 7.1 gives the Type 2 MLE of a_i together with the empirical prior mean proportion of failures in each time period and the proportion of failures observed in each of the eight classes corresponding to engineering concerns. Note that the empirical Bayes inference does not impose a monotonic function on the form of the prior distribution.

Table 7.1 Estimates of empirical Dirichlet prior distribution parameters for unit concern classes

Time interval	MLE of a_i	EB prior estimates of $\underline{P}$	Observed proportion of events in classes							
			C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
1	6.74	0.28	0.34	0.20	0.00	0.30	0.50	1.00	0.33	0.18
2	3.27	0.14	0.09	0.00	1.00	0.07	0.00	0.00	0.00	0.08
3	8.08	0.34	0.27	0.20	0.00	0.50	0.50	0.00	0.33	0.49
4	4.66	0.20	0.10	0.60	0.00	0.13	0.00	0.00	0.33	0.25
5	1.03	0.04	0.19	0.00	0.00	0.00	0.00	0.00	0.00	0.00

7.5 Summary and Conclusions

We have proposed a methodology for elicitation that aims to preserve the character of early stage judgements by using empirical data, where possible, to express a probability assessment of uncertainty consistent with an expert's beliefs. Our rationale is based upon the premise that expertise does not reside in the stochastic characterisation of the events, but rather upon other problem features to which an expert can relate her domain knowledge. Thus we map the quantity of interest to an expert's experience when there are associated data sets to support the quantification of uncertainty. Empirical Bayes inference is used to estimate the prior probabilities with the relevant observational data to provide a distribution representing the epistemic uncertainty about the quantity of interest.

We contribute a methodology consistent with an outside view of uncertainty assessment as discussed by Kahneman and Lovallo (1993). Our approach avoids imposing conformance upon an expert when assessing uncertainties probabilistically. Therefore it is capable, in principle, of overcoming some biases acknowledged to exist when an expert makes a subjective assessment through an internal mapping to an assumed probability mechanism.

7.5.1 Methodological Steps

Table 7.2 summarises our methodology in the five key steps, which can be summarised by the acronym CISSE corresponding to the initial verbs associated with the purpose of each step. The tasks involved in translating the general steps to an application are described to provide an analytical guide. Cross-references to the choices made for the two example applications are provided for illustration. Specifically, the role of the expert within each step is shown to highlight how subjective judgement is kernal to obtaining a meaningful prior distribution estimated using empirical data.

Table 7.2 CISSE methodology to construct a subjective probability distribution with data

Step	Acronym	Objective	General description	Expert role	Example 1	Example 2
1	C	Characterise the population DGP	Define the reference factors believed to characterise the future history of the quantity of interest	Provide domain knowledge to specify reference factors	Identify factors influencing new supplier quality measured by non-conformances	Identify factors influencing realisation of engineering concerns as failures
2	I	Identify candidate sample DGPs matching population	Match empirical data for entities with observed event histories to the future history of the quantity of interest based on the reference factors	Advise on candidate data sets and reasons for matching	Match to data sets for existing suppliers	Match to data sets for related designs
3	S	Sentence empirical data to construct sample DGPs	Make choices about which events for the selected data sets provide relevant empirical data for the comparator pool	Assess relevance of observed events to possible future events	Choose records for matched suppliers	Choose records for matched concern classes
4	S	Select probability model for population DGP	Choose appropriate probability model for the population to which the future and observed histories are believed to belong	Advise on credibility of model assumptions	Select Poisson-Gamma model for non-conformances	Select multinomial-Dirichlet model for realisation of failures in time windows
5	E	Estimate model parameters with sample DGPs to obtain prior distribution	Obtain empirical prior representing the future history for the quantity of interest estimated using the data for the comparator pool	Verify empirical prior is valid expression of uncertainty in quantity of interest	Estimate prior empirically and check consistency with expert's belief about uncertainty in new supplier non-conformance rate	Estimate prior empirically and check consistency with expert's belief about uncertainty in time to failure of new engineering design

7.5.2 Effect of Sample Size on Prior Distribution

Given our reliance on empirical data, there is an obvious question relating to the impact of 'sample size' on estimation and hence upon the representation of uncertainty in the prior distribution. Heuristically we can appreciate that there are two competing sample size effects. The relationship between these effects on inference might be complicated but we can reason through the effects of the choices we make in steps 2 and 3 by considering the effect of the length of a sample DGP and the number of sample DGPs separately.

Firstly, as the number of sample DPGs increases then the sampling variation in estimating the parameters of the prior distribution will reduce. This implies that the confidence regions for the comparator pool parameters will be tighter, and the associated tolerance intervals of credible prior distributions consistent with the empirical data will be narrower, when a larger number of sample DPGs are selected to match the population DGP. For example, we showed analysis of the sampling variation on the pool estimates based on the 35 suppliers used in our first example application. Had we used a larger (smaller) number of data sets providing equivalent similarity matches then we would expect the tolerance intervals to be narrower (wider) than those shown in Fig.7.7.

Secondly, as the history of a sample DGP increases then this will primarily reduce parameter estimation error associated with that individual DGP with only a marginal error reduction in estimates for the comparator pool. For example, an empirical Bayes estimate of the non-conformance rate for an individual supplier will be affected more by changes in the length of the event history for that supplier than the corresponding estimates based on the comparator pool which provides the prior distribution for the true rate of non-conformance of a new supplier.

We emphasise that our reasoning is limited to consideration of the mutually exclusive effects of the number and length of sample DGPs. However, it is important to appreciate such sample size effects because of the resulting implications for the degree of uncertainty inferred in the empirically constructed prior distribution. It is possible, as shown for our first example, to quantify the effects of sampling error allowing us to appreciate the implications for the assessment of uncertainty.

7.5.3 Caveats and Challenges

We acknowledge some caveats associated with our approach. Importantly, it will only be feasible in situations where it is possible to construct a comparator pool of data consistent with an expert's articulation of the reference factors defining the population DGP. This might not always be the case. For example, radical innovations leading to very novel engineering designs in a reliability context, or long term predictions in a supply chain management context are problem contexts for which our approach is less credible. More generally, if no candidate sample

DGPs can be identified then constructing a prior through our proposed empirical approach should not be pursued. Even when comparator pools do exist then the analyst has considerable responsibility in ensuring that the data used are relevant and defensible given the impact of making choices about candidate data sets and forming relevant sample DGPs. A formal means of allowing an expert to assess the credibility of the empirical prior provides a degree of mitigation against this risk.

It is well known that empirical Bayes inference improves as comparator pool homogeneity increases (e.g. Efron 2012; Carlin and Louis 2000). Here we have constructed sample DGPs through a process involving subjective expert judgement. It is possible to scientifically aid the homogenisation process by including homogenisation factors within the probability model. See, for example, Quigley et al. (2011) who examine the role of expert judgement to specify homogenisation factors.

In our example applications we have illustrated the choices made during elicitation using the type of empirical data available at the time of analysis. Both contexts considered relate to scenarios where extensive data already exists but has not being fully utilised to understand the degree of uncertainty associated with the quantities of interest relevant to engineering development and operational decision-making. There are potentially interesting challenges affecting data selection and sentencing with more extensive or unstructured data that might be available in future. For example, in a reliability context many engineering systems are fitted with many sensors implying more empirical data is available for covariates (Meeker and Hong 2014) that may relate to the reference factors that define the population characteristics. Such explanatory data from sensors and other automated data collection might be used to support more effective and/or efficient formation of comparator pools.

Since we have proposed and illustrated how to construct a 'subjective' probability distribution using data, we conclude by emphasising the importance of engaging an expert in key steps. The nature of our approach also requires us to examine the roles of both the subject domain experts and the analytical experts because both make choices that impact the prior probability distribution obtained. The analyst makes choices in our methodology, as indeed in any elicitation process, in relation to issues such as who are experts, how should they be engaged and how should their judgements be credibly expressed. However we also require the analyst to be actively engaged in data preparation, probability model and inference method selection. Most importantly, where possible, we are not asking an expert to express his or her uncertainty about some event of interest as a subjective probability. Rather we advocate using the subject domain expertise to structure the characteristics of the population DGP for the quantity of interest and to be involved at the key stages of matching candidate data sets, sentencing records and assessing the credibility of probability distributions, both in terms of any underlying assumptions and the resulting profile of uncertainty. Our goal is to construct empirically a probability distribution that is consistent with the subjective assessment of uncertainty about a relevant quantity by the expert.

Acknowledgements We would like to thank the many engineers and managers from various companies who have been involved in challenging and evaluating our approach in practical decision-making contexts. Their engagement has helped us develop our scientific thinking into an operational process.

References

Arnold S (1990) Mathematical statistics. Prentice-Hall, Englewood Cliffs

Carlin BP, Louis TA (2000) Bayes and empirical Bayes methods for data analysis. Chapman & Hall/CRC, Boca Raton

Cheng EK (2009) A practical solution to the reference class problem. Columbia Law Rev 109(8):2081–2105

Cochran W (1975) Sampling techniques. Wiley, New York

Cooke RM (1996) The design of reliability databases Part 1 - review of basic design concepts. Reliab Eng Syst Saf 51(2):137–146

Efron B (2012) Large-scale inference: empirical Bayes methods for estimation, testing, and prediction, vol 1. Cambridge University Press, Cambridge

Efron B, Morris C (1972) Limiting the risk of Bayes and empirical Bayes estimators - Part II: the empirical Bayes case. J Am Stat Assoc 67(337):130–139

Efron B, Morris C (1973) Stein's estimation rule and its competitors - an empirical Bayes approach. J Am Stat Assoc 68(341):117–130

Efron B, Morris C (1975) Data analysis using Stein's estimator and its generalizations. J Am Stat Assoc 70(350):311–319

Efron B, Tibshirani R, Storey JD, Tusher V (2001) Empirical Bayes analysis of a microarray experiment. J Am Stat Assoc 96(456):1151–1160

EFSA (2015) Scientific opinion on the risks for public health related to the presence of chlorates in food. EFSA J 13(6):4135

Fragola JR (1996) Risk management in US manned spacecraft: from Apollo to Alpha and beyond. In: Perry M (ed) Proceedings of the product assurance symposium and software product assurance workshop, EAS SP-377, European Space Agency, pp 83–92

Gallien J, Mersereau AJ, Garro A, Mora AD, Vidal MN (2015) Initial shipment decisions for new products at Zara. Oper Res 63(2):269–286

Good IJ (1965) The estimation of probabilities. Research monograph, vol 30. MIT Press, Cambridge, MA

Good IJ (1976) The Bayesian influence, or how to sweep subjectivism under the carpet. In: Foundations of probability theory, statistical inference, and statistical theories of science, Springer Netherlands, New York, pp 125–174

Greenwood M, Yule GU (1920) An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. J R Stat Soc 83(2):255–279

Hodge R, Evans M, Marshall J, Quigley J, Walls L (2001) Eliciting engineering knowledge about reliability during design-lessons learnt from implementation. Qual Reliab Eng Int 17(3): 169–179

Johnston W, Quigley J, Walls L (2006) Optimal allocation of reliability tasks to mitigate faults during system development. IMA J Manag Math 17(2):159–169

Kahneman D, Lovallo D (1993) Timid choices and bold forecasts: a cognitive perspective on risk taking. Manag Sci 39(1):17–31

Klugman SA, Panjer HH, Willmot GE (2012) Loss models: from data to decisions. Wiley, New York

Koriat A, Lichtenstein S, Fischhoff B (1980) Reasons for confidence. J Exp Psychol Hum Learn Mem 6(2):107–118

Meeker WQ, Hong Y (2014) Reliability meets big data: opportunities and challenges. Qual Eng 26(1):102–116
Nagurney A, Li D (2016) Competing on supply chain quality. Springer, Berlin
Ng KW, Tian GL, Tang ML (2011) Dirichlet and related distributions: theory, methods and applications, vol 888. Wiley, New York
Pahl G, Beitz W (2013) Engineering design: a systematic approach. Springer, Berlin
Quigley J, Walls L (2011) Mixing Bayes and empirical Bayes inference to anticipate the realization of engineering concerns about variant system designs. Reliab Eng Syst Saf 96(8):933–941
Quigley J, Bedford T, Walls L (2007) Estimating rate of occurrence of rare events with empirical Bayes: a railway application. Reliab Eng Syst Saf 92(5):619–627
Quigley J, Hardman G, Bedford T, Walls L (2011) Merging expert and empirical data for rare event frequency estimation: pool homogenisation for empirical Bayes models. Reliab Eng Syst Saf 96(6):687–695
Quigley J, Walls L, Demirel G, MacCarthy B and Parsa M (2018) Supplier quality improvement: the value of information under uncertainty. Eur J Oper Res 264(3):932–947
Rausand M, Hoyland A (2004) System reliability theory: models, statistical methods and applications. Wiley, New York
Reichenbach H (1971) The theory of probability. University of California Press, Berkley
Robbins H (1955) An empirical Bayes approach to statistics. In: Proceedings of the third Berkley symposium mathematical statistics and probability 1, University of California Press, Berkley, pp 157–164
Slack N, Brandon-Jones A, Johnston R (2016) Operations management, 8th edn. Pearson
Sodhi MS, Tang CS (2012) Managing supply chain risk. Springer, Berlin
Spetzler CS, Stael von Holstein CAS (1975) Exceptional paper-probability encoding in decision analysis. Manag Sci 22(3):340–358
Talluri S, Narasimhan R, Chung W (2010) Manufacturer cooperation in supplier development under risk. Eur J Oper Res 207(1):165–173
von Mises R (1942) On the correct use of Bayes' formula. Ann Math Stat 13(2):156–165
Walls L, Quigley J (1999) Learning to improve reliability during system development. Eur J Oper Res 119(2):495–509
Walls L, Quigley J (2001) Building prior distributions to support Bayesian reliability growth modelling using expert judgement. Reliab Eng Syst Saf 74(2):117–128
Walls L, Quigley J, Marshall J (2006) Modeling to support reliability enhancement during product development with applications in the UK aerospace industry. IEEE Trans Eng Manag 53(2):263–274
Wilson KJ, Quigley J (2016) Allocation of tasks for reliability growth using multi-attribute utility. Eur J Oper Res 255(1):259–271
Zhu K, Zhang RQ, Tsung F (2007) Pushing quality improvement along supply chains. Pest Manag Sci 53(3):421–436

Chapter 8
Eliciting Multivariate Uncertainty from Experts: Considerations and Approaches Along the Expert Judgement Process

Christoph Werner, Anca M. Hanea, and Oswaldo Morales-Nápoles

Abstract In decision and risk analysis problems, modelling uncertainty probabilistically provides key insights and information for decision makers. A common challenge is that uncertainties are typically not isolated but interlinked which introduces complex (and often unexpected) effects on the model output. Therefore, dependence needs to be taken into account and modelled appropriately if simplifying assumptions, such as independence, are not sensible. Similar to the case of univariate uncertainty, which is described elsewhere in this book, relevant historical data to quantify a (dependence) model are often lacking or too costly to obtain. This may be true even when data on a model's univariate quantities, such as marginal probabilities, are available. Then, specifying dependence between the uncertain variables through expert judgement is the only sensible option. A structured and formal process to the elicitation is essential for ensuring methodological robustness. This chapter addresses the main elements of structured expert judgement processes for dependence elicitation. We introduce the processes' common elements, typically used for eliciting univariate quantities, and present the differences that need to be considered at each of the process' steps for multivariate uncertainty. Further, we review findings from the behavioural judgement and decision making literature on potential cognitive fallacies that can occur when assessing dependence as mitigating biases is a main objective of formal expert judgement processes. Given a practical focus, we reflect on case studies in addition to theoretical findings. Thus, this chapter serves as guidance for facilitators and analysts using expert judgement.

C. Werner (✉)
Department of Management Science, University of Strathclyde, Glasglow, UK
e-mail: christoph.werner@strath.ac.uk

A.M. Hanea
CEBRA, University of Melbourne, Parkville, VIC, Australia
e-mail: ahanea@unimelb.edu.au

O. Morales-Nápoles
Faculty of Civil Engineering and Geosciences, Technological University Delft, Delft, The Netherlands
e-mail: O.MoralesNapoles@tudelft.nl

L.C. Dias et al. (eds.), *Elicitation*, International Series in Operations Research & Management Science 261, https://doi.org/10.1007/978-3-319-65052-4_8

8.1 Introduction

Probabilistic modelling of uncertainties is a key approach to decision and risk analysis problems. It provides essential insights on the possible variability of a model's input variables and the uncertainty propagation onto its outputs.

Typically, uncertainties cannot be treated in isolation as they often exhibit dependence between them which can have unanticipated and (if not properly modelled) possibly misleading effects on the model outcome. Therefore, modelling dependence of uncertainties is an area of ongoing research and several modelling approaches have been developed, serving different purposes and allowing for varying levels of scrutiny. A common challenge with regards to model quantification is a lack of relevant historical data while simplifying assumptions, such as that of independence, are not justifiable. Then, the only sensible option for quantifying a model is by eliciting the dependence information through expert judgement. This is even necessary when relevant data on the marginal probabilities are available.

A structured approach to eliciting multivariate uncertainty is encouraged as it supports experts to express their knowledge and uncertainty accurately, hence producing well-informed judgements. For instance, cognitive fallacies might be present when experts assess dependence which can inhibit the judgements' accuracy. Therefore, mitigation of these fallacies is a main objective of an elicitation process. Further, a structured process addresses other questions which affect the reliability of the elicited result and hence model outcome, such as aggregating various judgements. Lastly, a formal process makes the elicited results transparent and auditable for anyone not directly involved in the elicitation.

8.1.1 Objective and Structure of the Chapter

Complementary to the case of eliciting univariate uncertainty, this chapter's objective is to outline the main elements of formal expert judgement processes for multivariate uncertainty elicitation. This is done by discussing theoretical and empirical findings on the topic, though the reader should note that fewer findings are available for eliciting joint distributions than for the elicitation of univariate quantities.

The structure of this chapter is as follows. In the remainder of this section we introduce a definition of dependence for the subjective probability context which establishes a common language and understanding of the key concept discussed here. In Sect. 8.2, the importance of formal expert judgement processes is discussed and an overview of the necessary adjustments for dependence elicitation is given. This provides the reader with the scope of the topic. Section 8.3 outlines the heuristics and biases that might occur when eliciting dependence. Then, Sect. 8.4 discusses the preparation of an elicitation (or the pre-elicitation stage) which for instance entails the choice of the elicited forms and the training of experts.

In Sect. 8.5, we present considerations for the actual elicitation phase, including structuring and decomposition methods as well as the quantitative assessment. In Sect. 8.6, we review required alterations of the process for the post-elicitation stage, such as when combining the expert judgements. Finally, Sect. 8.7 concludes the chapter by summarising the main points addressed and discussing the status-quo of this research problem.

8.1.2 Dependence in the Subjective Probability Context

In this chapter, we use the terms *dependence* and *multivariate uncertainty* interchangeably and in a general sense. They contrast the specific association measures (or dependence parameters) that quantify a dependence model and are therefore often used as elicited variables. When discussing dependence in a general sense, we refer to situations with multiple uncertain quantities and when gaining information about one quantity, we change the uncertainty assessments for the others. More formally, we say that two uncertain quantities X and Y are independent (for experts) if they do not change their beliefs about the distribution of X after obtaining information about Y. This is easily extended to higher dimensions in which all quantities are independent of one another if knowing about one group of variables does not change experts' beliefs about the other variables. It follows that dependence is simply the absence of independence.

Note that dependence in a subjective probability context is a property of an expert's belief about some quantities so that one expert's (in-)dependence assessment might not be shared with another expert who possesses a different state of knowledge (Lad 1996).

8.2 Structured Expert Judgement Processes: An Overview

The necessity for a structured and formal process when eliciting uncertainty from experts, such as in form of probabilities, has been recognised since its earliest approaches. For instance, it has been acknowledged in the area of Probabilistic Risk Analysis (PRA) which comprises a variety of systematic methodologies for risk estimation with uncertainty quantification at its core (Bedford and Cooke 2001). From a historical perspective, main contributions in PRA have been made in the aerospace, nuclear and chemical process sector. Hence, after expert judgement was used only in a semi-formal way in one of the first full-scale PRAs, the original Reactor Safety Study[1] by the US Nuclear Regulatory Commission (USNRC 1975),

[1]The study is also known as WASH-1400 and as the *Rasmussen Report* due to Norman Carl Rasmussen. At that time, the use of expert opinion for assessing uncertainties was often viewed

major changes towards a more scientific and transparent elicitation process were made in the subsequent studies, known as NUREG-1150 (USNRC 1987; Keeney and Von Winterfeldt 1991). When reflecting on the historical development of PRA, Cooke (2013) highlights the improvements made through a traceable elicitation protocol as a newly set standard and main achievement for expert judgement studies.

Another pioneering contributor to formal approaches for expert judgement is the Stanford Research Institute (SRI). The Decision Analysis Group of SRI similarly acknowledged the importance of a formal elicitation process when eliciting uncertainty from experts. Therefore, they developed a structured elicitation protocol that supports a trained interviewer through a number of techniques to reduce biases and aid the quantification of uncertainty (Spetzler and Staël von Holstein 1975; Staël von Holstein and Matheson 1979).

Following from these early contributions, various proposals for formal expert judgement processes have been made and its various components were further developed. While not one particular step-by-step process to follow exists given the varying and particular objectives of each elicitation, there is agreement regarding which high level steps are essential. Fairly complete elicitation protocols are for instance presented in Merkhofer (1987), Morgan and Henrion (1990), Cooke and Goossens (1999), Walls and Quigley (2001), Clemen and Reilly (2014) and EFSA (2014). Even though these references explicitly address the case of eliciting a univariate quantity, they serve as guidance for our purpose of presenting and discussing the considerations for eliciting dependence.

The elicitation of dependence follows historically from advances made for eliciting univariate uncertainty and an overview of the historical development of expert judgement in risk analysis is presented in Cooke (2013). This development is not surprising given that marginal distributions need to be specified (at least implicitly) before any dependence assessment can be made. Furthermore, univariate quantities are (typically) more intuitive to assess. Whereas some findings for eliciting univariate uncertainty are still applicable in the multivariate case, for other parts of the process adjustments need to be made. Figure 8.1 shows the main elements of elicitation processes with the modifications that are necessary when eliciting dependence.

Regarding the different roles during an elicitation, in this chapter we consider the situation of a specific decision or risk analysis problem that is of importance for a *decision maker*. *Experts* assess the uncertainty on the variables without any responsibility for the model outcome or consequences of the later decision. The experts are chosen based on their substantive (also subject-matter) expertise, meaning they are experts on the particular topic of the decision problem. This implies that the experts might not have normative expertise, thus they are not

highly sceptical, however a main challenge was that until then no nuclear plant accident had been observed. Therefore, the report, together with its use of expert opinion, was only revived due to the Three Mile Island accident (1979). After the incident, the report's results were prescient. In particular, the inclusion of human error as a source of risk made the case for expert judgement.

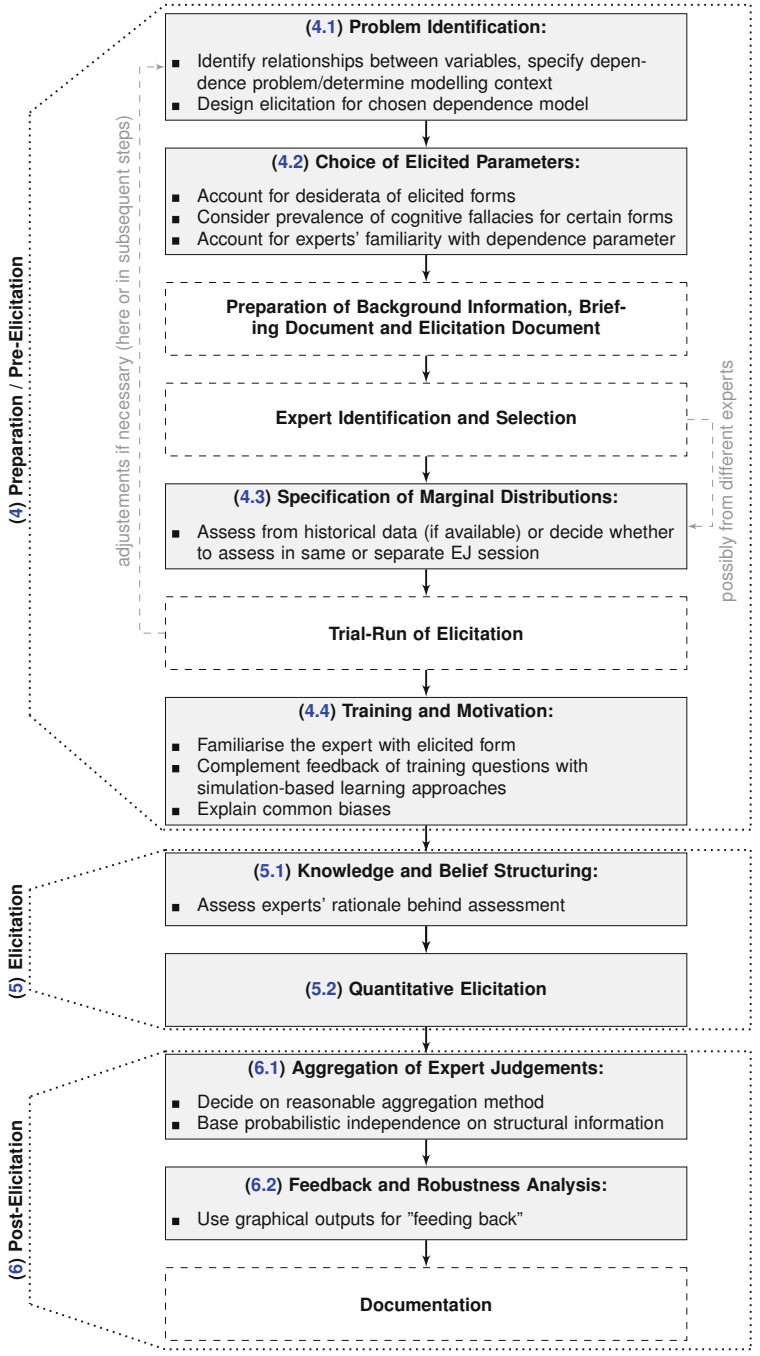


Fig. 8.1 Overview of the expert judgement process adjusted for eliciting dependence (steps discussed in this chapter are in grey)

statistical or probabilistic experts. The *facilitator*, who manages the actual elicitation part of the overall process, might be either the same person as the decision maker or an independent third type of attendee at the elicitation workshop. The facilitator clarifies any questions from the experts. An *analyst* on the other hand is usually in charge of the whole process. This includes the preparation of the elicitation and the finalisation of results afterwards. Such a situation with a given, formulated problem and clearly defined roles is often the case, however other ones are possible. French (2011) discusses various elicitation contexts and their potential implications.

We regard an elicitation as successful if we can be confident that the experts' knowledge is captured accurately and faithfully, thus their uncertainty is quantified through a well-informed judgement. However, the assessments' reliability might be still poor if little knowledge about the problem of interest prevails. This often implies that there is high uncertainty in the area of the decision problem overall.

8.3 Biases and Heuristics for Dependence Elicitation

In this section, we review main findings from the behavioural judgement and decision making literature on assessing dependence as psychological research shows that experts are not guaranteed to act rationally when making such assessments. Hence, the goal of this section is to raise awareness of departures from rationality in the hope to minimise them in the elicitation. Briefly, rationality implies that experts make assessments in accordance with normative theories for cognition, such as formal logic, probability and decision theory. Irrationality, on the other hand, is the systemic deviation from these norms. While this definition suffices here, the topic is much more complex and a critical debate on the concept of rationality can be found in Stanovich and West (2000) and Over (2004). In contrast to normative theories that describe how assessments ought to be made, descriptive research investigates how assessments are actually made. This relates directly to our earlier definition of a successful elicitation (Sect. 8.2) that states our aim of eliciting accurate and faithful assessments from experts. In other words, a successful elicitation aims at mitigating a range of potential biases.

For expert judgement, in particular two types of biases, cognitive and motivational, are of importance as they can distort the elicitation outcome severely.

Cognitive biases refer to the situation in which experts' judgements deviate from a normative reference point in a subconscious manner, i.e. influenced by the way information is mentally processed (Gilovich et al. 2002). This bias type occurs mainly due to *heuristics*, in other words because people make judgements intuitively by using mental short-cuts and experience-based techniques to derive the required assessments. The idea of a heuristic proof was used in mathematics to describe a provisional proof already by Pólya (1941), before the term was adopted in psychology, following Simon (1957) with the concepts of *bounded rationality* and *satisficing*.

Motivational biases may deviate experts' judgements away from their true beliefs. In other words, experts ought to make the most accurate judgements regardless of the implied conclusion or outcome, yet they do not. Motivational biases happen consciously and depend on the experts' personal situations. For instance, social pressures, wishful thinking, self-interest as well as organizational contexts can trigger this type of biases (Montibeller and Von Winterfeldt 2015). Given that motivational biases are not different for univariate and multivariate uncertainty assessments we will not consider them in our review in Sect. 8.3.2.

Regarding the mitigation of biases, a motivational bias can be addressed in a technical way by introducing (strictly proper) scoring rules or as well by the direct influence of a facilitator who encourages truthful answers. A cognitive bias is mainly counteracted through training of experts, decomposing and/or structuring the experts' knowledge prior to the quantitative elicitation as well as a sensible framing of the elicitation question(s). The latter also entails the choice of the elicited form.

Over the last 40 years, the number of newly identified heuristics and biases has increased tremendously. Nevertheless, only a few findings are available for the case of assessing dependence. We present these findings in the remainder of this section and Table 8.1 provides an overview. For discussions on some main univariate biases, we refer to Kynn (2008) and Montibeller and Von Winterfeldt (2015).

As can be seen in Table 8.1, most identified heuristic and biases that are applicable for the case of multivariate uncertainty concern conditional assessments, such as conditional probabilities. While conditionality is a common way to conceptualise probabilistic dependence, it is shown that in addition to the explicit fallacies (as introduced in the following), understanding and interpreting conditional forms

Table 8.1 Main biases and heuristics for dependence elicitation

	Name	Reference(s)[a]	Description	Originates with	Suggested Remedies
conditional	Confusion of the inverse	Meehl and Rosen (1955), Eddy (1982), Dawes (1988), Hastie and Dawes (2001)	Experts confuse conditional probabilities of P(X|Y) with its inverse P(Y|X)	representativeness heuristic, causal interpretation, "non-natural" base-rates	elicit frequency formats (if possible) (O'Hagan et al. 2006, Meder and Gigerenzer 2014), structure rationale/relationships, include graphical aids (see Fountain and Gunby (2011)
	Causality heuristic	Ajzen (1977), Tversky and Kahneman (1980)	Experts overestimate P(X|Y) when perceiving causal relationship, i.e. Y *causing* X	causal interpretation, base-rate neglect	avoid single, focal scenarios as experts' rationale, evoke alternative scenarios, use experts with different backgrounds (Wright and Goodwin 2009)
	Insufficiently regressive prediction	Kahneman and Tversky (1973)	Experts *translate* one scale to the other, not adjusting for imperfect association	representativeness heuristic, predictive interpretation	specify reference class with central tendency and variability → assess individual case → adjust/calibrate (O'Hagan et al. 2006)
	Bayesian likelihood bias	Edwards (1965), DuCharme (1970)	Experts are more conservative than Bayes' Theorem implies	representativeness heuristic, base-rate neglect	decompose into assessing priors (odds) and likelihoods (ratios) (Montibeller and Von Winterfeldt 2015)
conjunction	Confusion of joint and conditional probabilities	Einhorn and Hogarth (1986)	Experts confuse joint and conditional probabilities	causal/temporal interpretation	address semantic misunderstandings in training, structure rationale/scenarios/ functional relationships
	Conjunction fallacy	Ajzen (1977), Tversky and Kahneman (1980)	the conjunction of X and Y is judged as more probable than X and Y individually	causal interpretation, base-rate neglect	demonstrate probability logic (Montibeller and Von Winterfeldt 2015)
concordance	Cell A strategy	Smedslund (1963), Allan (1980), Kao and Wasserman (1993)	Experts overvalue joint presence of variables (in bivariate assessment)	predictive interpretation	clarify underlying assumptions, such as rarity assumption
general	Illusory correlation	Chapman and Chapman (1969), Eder et al. (2011)	Experts base assessment on false (pre-existing) belief about relationships	availability bias, causal interpretation	as for availability: provide probability training, counter-examples, relevant statistics (if available) (Montibeller and Von Winterfeldt 2015)

[a] only main and/or original references listed

remains a challenge in today's statistics and probability education (Díaz et al. 2010). An explanation for this difficulty comes from Carranza and Kuzniak (2009) who note that a main focus of probability education is on frequentist approaches to probability together with (idealised) random experiments, such as coin tosses. Regarding conditional probabilities, such a position is however problematic as with equally likely cases, reducing the subspace has no clear impact on the equal probabilities. With a subjective view on probability (Sect. 8.1.2) on the other hand, a conditional probability is more intuitive as one simply revises judgements given new information that has become available (Borovcnik and Kapadia 2014).

8.3.1 Causal Reasoning and Inference

Before we address in detail the biases from Table 8.1, recall that we are interested in the experts' ability to assess dependence in accordance with Sect. 8.1.2. Usually this is done through specifying a dependence parameter and we address the choice of an elicited form in Sect. 8.4.3. While emphasizing that assessing dependence, e.g. as a correlation, is not the same as claiming a causal relationship, we consider experts' mental models about causal relationships as a main determinant for their assessments (despite the missing statistical noise). Therefore, we briefly address findings of behavioural studies on causal reasoning and inference first.

The concept of causation itself is highly debated[2] and its discussion is out of scope here, yet it is proposed that in most situations people believe that events actually have causes. In other words, their belief is that events mainly occur due to causal relationships rather than due to pure randomness or chance (Hastie 2016). Moreover, it is argued that people have systematic rules for inferring such causal relationships based on their subjective perception (Einhorn and Hogarth 1986). They then update their mental models of causal relationships continuously and might express summaries of causal beliefs in various forms, such as serial narratives, conceptual networks or images of (mechanical) systems (Hastie 2016).

Due to incomplete knowledge and imperfect mental models, we emphasize the concept of probabilistic causation (Suppes 1970). A formal framework that has been used widely for representing probable causes in fields such as statistics, artificial intelligence, as well as philosophy of science and psychology, is a probabilistic (causal) network. The topic of causation within probabilistic networks is however not without criticism and generates debate. Extensive coverage of this topic is given in Spirtes et al. (2000), Pearl (2009) and Rottman and Hastie (2014).

[2]There has been ongoing philosophical debate about the meaning of causation. While some refuted the concept of causation in science altogether (Russell 1912), others focused on specific aspects. For us, probabilistic causation (Suppes 1970) and its perception/inference are of interest. Hume (1748/2000) proposes one of the most established accounts for that. He proposes a (unobservable) causal mechanism which is inferred through the regularity of an effect following a cause.

A first type of information for inferring a probabilistic causal relationship is the set of necessary and sufficient conditions that constitute a presumed background of no (or only little) causal relevance (i.e. they are not informative for inference), but which need to be in place for an effect to happen. These conditions are known as *causal field*. For instance, when inferring the cause(s) of someone's death, being born is a necessary and sufficient condition, nevertheless it is of little relevance for establishing a causal explanation (Einhorn and Hogarth 1986). The causal field is a key consideration when structuring experts' beliefs about relationships as it relates to model boundaries and determines which events should be included in a graphical (or any other) representation of the system of interest. We discuss structuring beliefs in Sect. 8.5.1.

Another type of information that is assumed to be in place for making causal inferences is summarised as *cues-to-causality*. Most of these origin with Hume (1748/2000) and comprise temporal order, contiguity in time and space, similarity, covariation, counterfactual dependence and beliefs about the underlying causal mechanism as seen by events' positions in causal networks (Hastie 2016). Generally, the presence of multiple cues decreases the overall uncertainty, even though conflicting cues increase it. The way in which these cues are embedded in the causal field and how both types of information together shape one's causal belief is shown by Einhorn and Hogarth (1986) with the following example:

> Imagine that a watch face has been hit by a hammer and the glass breaks. How likely was the force of the hammer the cause of the breakage? Because no explicit context is given, an implicitly assumed neutral context is invoked in which the cues-to-causality point strongly to a causal relation; that is, the force of the hammer precedes the breakage in time, there is high covariation between glass breaking (or not) with the force of solid objects, contiguity in time and space is high, and there is congruity (similarity) between the length and strength of cause and effect. Moreover, it is difficult to discount the causal link because there are few alternative explanations to consider. Now imagine that the same event occurred during a testing procedure in a watch factory. In this context, the cause of the breakage is more often judged to be a defect in the glass.

This simple example shows that by changing the contextual factors while keeping the cues constant, someone's causal belief can change rather dramatically.

The ways in which these types of information influence a causal perception are important for the remainder of this section as experts' causal beliefs and inferences often serve as candidate sources for several biases.

8.3.2 Biased Dependence Elicitation: An Overview

In the following, the main cognitive fallacies that can occur when eliciting dependence, as shown in Table 8.1, are presented in more detail. In addition to introducing the examples that the original researchers of the different biases propose, we illustrate each bias with a simplified example from the area of project risk assessment. Explaining all biases with the same example allows for a better comparison between their relevance and the context in which they apply.

Suppose, we manage a project with an associated overall cost. The project's overall cost is determined by various individual activities which are essential for the project completion and which each have their own cost. We denote the cost of an individual activity by c_a and when we distinguish explicitly between two different activities, we do so by indexing them as 1 and 2, so as c_{a_1} and c_{a_2}. It follows that we are interested in modelling and quantifying the dependence between the individual activities' costs and the dependence's impact on the overall cost. Note that assuming independence between the activities might severely underestimate the likelihood of exceeding some planned overall cost. In order to better understand the dependence relationships, we take for instance into account how the individual activities can be jointly influenced by environmental and systemic uncertainties. In this simple example, we consider whether (and if yes, how) such uncertainties impact the activities' costs, e.g. due to affecting the durations of certain activities. The duration or time an activity takes is represented by t_a. A main area of research in PRA that focuses on modelling implicit uncertainties, which have a joint effect on the model outcome but that are not well enough understood to consider these factors explicitly, is common cause modelling. For an introduction, see Bedford and Cooke (2001).

Confusion of the Inverse A common way of eliciting dependence is in form of conditional judgements, such as conditional probabilities (Sect. 8.4.2). A main bias for conditional forms of query variables is the *confusion of the inverse* (Meehl and Rosen 1955; Eddy 1982; Dawes 1988; Hastie and Dawes 2001). Villejoubert and Mandel (2002) provide a list of alternative names proposed in the literature. For that, a conditional probability $P(X|Y)$ is confused with $P(Y|X)$. In our project risk example, this might happen when considering the time that an activity takes and whether this influences its own (but also other activities') cost. When eliciting the conditional probability $P(c_a \geq v|t_a \geq w)$ where v and w are specific values, an expert might confuse this with its inverse, $P(t_a \geq w|c_a \geq v)$.

An empirical research area in which this fallacy has been studied more extensively is medical decision making. It is shown that medical experts often confuse conditional probabilities of the form $P(\text{test result}|\text{disease})$ and $P(\text{disease}|\text{test result})$. In a pioneering study, Eddy (1982) reports this confusion for cancer and positive X-ray results. More recently, Utts (2003) lists the confusion of the inverse among the main misunderstanding that "educated citizens" have when making sense of probabilistic or statistical data. Further, Utts (2003) outlines several cases in which being prone to this fallacy has led to false reporting about risk in the media.

One explanation for confusing the inverse is attributed to the similarity of X and Y. Therefore, some researchers suggest that this bias is linked to the better known *representativeness heuristic* (Kahneman and Tversky 1972; Kahneman and Frederick 2002). For that, people assess the probability of an event with respect to essential characteristics of the population which it resembles. For dependence assessments this implies that experts regard $P(X|Y) = P(Y|X)$ due to the resem-

blance or representativeness of X for Y and vice versa (O'Hagan et al. 2006). For instance a time-intensive project activity might resemble a cost-intensive one and vice versa.

Another explanation for this fallacy is related to neglecting (or undervaluing) base-rate information (Koehler 1996; Fiedler et al. 2000). Generally, the *base-rate neglect* (Kahneman and Tversky 1973; Bar-Hillel 1980) states that people attribute too much weight to case-specific information and too little (or no) to underlying base-rates, i.e. the more generic information. With regards to confusing the inverse, Gavanski and Hui (1992) distinguish between *natural* and *non-natural* sampling spaces. A natural sampling space is one that is accessed more easily in one's memory (this may or may not be the sample space as prescribed by probability theory). In the fallacy's classical example of P(test result|disease) for instance, the sample space of "people with a disease" often comes to mind easier than that of "people with a certain test result", such as "positive", given that the latter can span over several types of diseases. Similarly in our project risk example, for $P(c_a \geq v | t_a \geq w)$ an expert ought to regard the activities exceeding a certain duration before thinking of the activities within this subspace that exceed a certain cost. However, the sample space of activities exceeding a specified cost might be more readily available so that from this the proportion of the activities exceeding a certain time is considered.

A last suggested source for the inverse fallacy stems from experts' (potentially) perceived causation between X and Y. Pollatsek et al. (1987) attribute a potential confusion between conditionality and causation to similar wordings such as "given that" or "if". Remember that temporal order is important for determining the cause(s) and the effect(s) of two or more events. For instance, Bechlivanidis and Lagnado (2013) show how causal beliefs influence the inference of their temporal order and vice versa, i.e. how temporal order informs causal beliefs. Thus, when eliciting the dependence between two activities' durations, experts might confuse $P(t_{a_1} \geq w | t_{a_2} \geq w)$ with its inverse if the durations are not easily observed, e.g. due to lagging processes, and the first completed activity is seen as causing the other.

In the medical domain, in which this confusion has been observed most often, we note that for P(test result|disease) the test result is observed first (in a temporal order) even though the outbreak of the disease clearly preceedes in time. Therefore, the cause is inferred from the effect. This is a situation in which Einhorn and Hogarth (1986) see the confusion of the inverse very likely to occur, even though temporal order has no role in probability theory. By some researchers, this is called the *time axis fallacy* or *Falk phenomenon* (Falk 1983). Another interesting example from medical research concerns the early days of cancer research and the association between smoking and lung cancer. While it is now established that smoking causes lung cancer, some researchers have also proposed the inverse (Bertsch McGrayne 2011). Indeed, the question of whether a certain behaviour leads to a disease or whether a disease leads to a certain behaviour can be less clear. A potential confusion of the inverse is then subject to the expert's belief on the candidate cause.

Causality Heuristic The close connection between conditional assessments and causal beliefs can be the source of another cognitive fallacy. In a pioneering study, Ajzen (1977) coined the term *causality heuristic*, claiming that people prefer causal information and therefore disregard non-causal information, such as base-rates with no causal implication. Other researchers (e.g. Bes et al. 2012) have since then confirmed this preference for causal information. At a general level, the causality heuristic relates to causal induction theories in contrast to similarity-based induction (Sloman and Lagnado 2005). For instance, Medin et al. (2003) found that people regarded the statement "bananas contain retinum, therefore monkeys do" as more convincing than "mice contain retinum, therefore monkeys do" which shows that the plausibility of a causal explanation can outweigh a similar reference class.

In the context of conditional assessments, it is noteworthy that people assess a higher probability for $P(X|Y)$ when a causal relation is perceived between X and Y, even though according to probability theory, a causal explanation should make no difference in the assessment (Falk 1983). This is shown further by people's preference to reason from causes to effects rather than from effects to causes (Hastie 2016). As a result, causal relations described as the former are judged as more likely than the latter even though both relations should be equally probable. For our example of assessing $P(c_a \geq v|t_a \geq w)$, we therefore need to consider whether experts perceive a causal explanation and how it influences the assessment outcome.

In an experimental study, Tversky and Kahneman (1980) asked subjects whether it is more probable that (*a*) a girl has blue eyes if her mother has blue eyes?, (*b*) a mother has blue eyes if her daughter has blue eyes?, or (*c*) whether both events have equal probability? While most participants (75) chose the correct answer (*c*), 69 participants responded (*a*) compared to 21 that chose (*b*). Whether this result can be fully attributed to the role of participants' perception of causation is however questionable given other possible influences on the assessments such as semantic difficulties (Einhorn and Hogarth 1986). Nevertheless, it is an indicator for how experts are led by preferences about perceiving a conditional relation (which might contradict the elicited one) once they regard the variables as causes and effects.

While sometimes being regarded as a different bias, the *simulation heuristic* (Keren and Teigen 2006) affects judgements in a very similar manner. Here, the premise is that conditional probability judgements are based on the consideration of if-then statements. This is an idea originating with Ramsey (1926) and his "degree of belief in p given q", roughly expressing the odds one would bet on p, the bet only being valid if q is true. Hence, it is proposed that for assessing a conditional probability, $P(X|Y)$, one first considers a world in which Y is certain before assessing the probability of X being in this world. The simulation heuristic states then that the ease with which one mentally simulates these situations affects the probability judgement. People often compare causal scenarios and tend to be most convinced by the story that is most easily imaginable, most causally coherent and easiest to follow. However, they then neglect other types of relevant information together with causal scenarios that are not readily available for their conception.

Insufficiently Regressive Prediction A fallacy that might occur when people interpret a conditional form as a predictive relation is *insufficiently regressive prediction*. Kahneman and Tversky (1973) show that when assessing predictive relationships, people do not follow normative principles of statistical prediction. Instead, they "merely translate the variable from one scale to another" (Kahneman and Tversky 1973). In the project risk example, when predicting an activity's cost from its duration, e.g. through conditional quantiles, experts might simply choose the value of the cost's *i*th quantile based on the time's *i*th quantile. This is problematic as typically there is no perfect association between the variables. Hence, people do not adjust their assessment for a less than perfect association between the variables. O'Hagan et al. (2006) give an example of predicting the height of males from their weight while assuming a correlation of 0.5 between the variables. Then, for a male who is one standard deviation above the mean weight, the best prediction for his height should only be 0.5 standard deviations above the mean height. However, people tend to assess the prediction too close to one standard deviation above the mean height.

A common explanation for this fallacy is again the representativeness heuristic. Regarding one variable representative for the other, e.g. viewing tall as representative for being heavy or a time-intensive project activity as representative for a cost-intensive one, experts disregard the aforementioned imperfect association.

As shown in Sect. 8.4.2, eliciting conditional quantiles is one common way to elicit dependence information.

Bayesian Likelihood Bias Research investigating experts' conditional assessments in the context of intuitively using Bayes' Theorem[3] formulated what is named (by some) the *Bayesian likelihood bias* (DuCharme 1970). Bayes' Theorem is proposed as a normative rule for revising probabilities given new evidence. The fallacy is that people are too conservative in their assessment (Edwards 1965), at least for certain framings (see Kynn (2008) for a critical discussion on this fallacy). The univariate equivalent is the *conservatism bias*. It refers to the finding that higher probabilities are underestimated while lower ones are overestimated, i.e. assessments vary less from the mean and avoid extreme values. For $P(c_{a_1} \geq v | c_{a_2} \geq v)$, experts might make too conservative assessments in light of new information about another activity's cost. In a pioneering study by DuCharme (1970), participants assessing the probability of a person's gender given the height, $P(\text{gender}|\text{height})$, tended to underestimate the number of tall men and overestimate the number of tall women.

[3]Bayes' Theorem is named after Thomas Bayes (1701–1761) who first proposed it. Since then it has been further developed and had its impact in a variety of problem contexts (see Bertsch McGrayne 2011 for a historical overview). In its simplest form, for events X and Y, it is defined as $P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$ whereas $P(Y) \neq 0$.

Confusion of Joint and Conditional Probabilities A cognitive fallacy that might be present when assessing dependence for events occurring together, i.e. the conjunction of events, such as in a joint probability assessment is the *confusion of joint and conditional probabilities*.

Consider the framing of the elicitation question: "What is the probability of $c_{a_1} \geq v$ and $c_{a_2} \geq v$?" While a more precise framing (specifying that we elicit the joint probability) or eliciting a joint probability still framed differently (see Sect. 8.4.3) would be helpful, it is important to note that from the view of probability theory, when using the word "and", we would expect the expert to assess $P(c_{a_1} \geq v \cap c_{a_2} \geq v)$, i.e. the conjunction of the events. However, it is shown that this is often interpreted differently. For some people "and" implies a temporal order (which has no role in probability theory), so they assess the conditional probability of $P(c_{a_1} \geq v | c_{a_2} \geq v)$ instead (Einhorn and Hogarth 1986). This fallacy is closely related to the confusion of the inverse for which one explanation is based as well on an implicit influence of temporal order.

Conjunction Fallacy A more extensively studied bias that is relevant when eliciting the conjunction of events is the *conjunction fallacy* (Tversky and Kahneman 1983). In experiments, subjects assessed the probability of a conjunction of events $P(X \cap Y)$ as more probable than its separate components, i.e. $P(X)$ or $P(Y)$, despite its contradiction to probability theory. For instance, when Lagnado and Sloman (2006) asked participants which of the following two statements is more likely: (*a*) a randomly selected male has had more than one heart attack, and (*b*) a randomly selected male has had more than one heart attack and he is over 55 years old, (*b*) was judged more probable than (*a*) by most participants. Similarly, experts in our project risk example might assess $P(c_a \geq v \cap t_a \geq w)$ as more probable than $P(c_a \geq v)$ or $P(t_a \geq w)$ separately.

As with the confusion of the inverse, a suggested source for the conjunction fallacy is the representativeness heuristic. However, while this is the most common explanation, it is not without criticism and numerous other candidate sources for this fallacy exist (Costello 2009; Tentori et al. 2013). For example, another explanation is the aforementioned causality heuristic. Hence, the constituent events are related through a causal explanatory variable. The additional information that constitutes the subset is then judged as causally relevant, as e.g. in our earlier examples being over the age of 55 is seen as causally relevant for having a heart attack, and an activity exceeding a certain duration for exceeding a certain cost.

In the context of assessing conditional probabilities, Lagnado and Shanks (2002) discuss the conjunction fallacy through the related concept of *disjunction errors*. People assess the conditional probabilities through subordinate and superordinate categories. For example in their example, a subordinate category, Asian flu, was regularly judged as more probable than its superordinate category, flu, given a set of symptoms. A possible explanation is based on a predictive interpretation for the conditional probability. Participants view the symptoms as more predictive for the subordinate category and base their likelihood judgement on it.

Cell a Strategy Some research focuses on interpreting and assessing dependence as the concordance of events whereas this is based on a frequency (or cross-sectional) interpretation for the event pairs. In other words, it explicitly requires a population to draw from. At the most general level, this relates to people's ability to assess dependence in form of the "perhaps simplest measure of association" (Kruskal 1958), the quadrant association measure. It gives the probability that the deviations of two random variables from (for instance) their medians have regularly the same signs, i.e. positive or negative. This is closely related to assessing a concordance probability which is introduced in Sect. 8.4.3.

In some situations this is the way how people perceive association between (binary) variables and a research stream that investigates this form of dependence perception is *associative learning* (Mitchell et al. 2009). A common cognitive fallacy is the *cell A strategy* (Kao and Wasserman 1993) which is named like this for reasons that will become apparent.

While certain activities are highly standardised and performed similarly across numerous projects, it is still rather an idealised case to serially observe whether or not the duration of the same activity exceeded a certain value for j projects with $j = 1, 2, \ldots, J$, i.e. whether $t_{a,j} \geq w$ or $t_{a,j} < w$, before obtaining this information for its cost. Despite its idealisation, this is how experts would perceive dependence in this case. Similarly in his pioneering study, Smedslund (1963) worked with medical experts and the variables referred to symptoms and diseases. The experts were given information about the presence or absence of a disease following information on the presence or absence of a symptom and then assessed its correlation.

This information can be ordered within four quadrants. The upper left corresponds to the presence of both variables, the lower right shows the joint absence and the remaining two quadrants relate to one variable being present while the other is absent. Whereas in normative theory, all four quadrants should be equally informative, it is found that people focus on the joint presence of both variables disproportionally in relation to the observed frequencies, so that this quadrant has a larger impact on the assessment. This quadrant has also been called cell A when labelling the four quadrants from A to D[4] which explains the name of this fallacy. It suggests that subjects fail to use all relevant information available and in fact, a preference order exists in form of $(X_+, Y_+) > (X_+, Y_-) \approx (X_-, Y_+) > (X_-, Y_-)$ (McKenzie and Mikkelsen 2007). Mandel and Lehman (1998) offer two explanations. The first considers the frequencies (or observations) per quadrant as a sample from a larger population and assumes presence is rare ($P < 0.5$) while absence is common ($P > 0.5$). Then a joint presence is more informative to judge a positive relationship in contrast to joint absence. In other words, it would be more surprising to observe a joint presence rather than a joint absence. The second

[4] When, $+$ indicates the presence of variables X and Y, and $-$ their absence, the quadrants can be presented as: $\begin{array}{l|l} A : (X_+, Y_+) & B : (X_+, Y_-) \\ \hline C : (X_-, Y_+) & D : (X_-, Y_-) \end{array}$.

explanation relates to hypothesis testing and since the quadrant of joint presence is evidence in favour of the hypothesis, this is again (typically) more informative in contrast to both non-joint quadrants that are evidence against it.

Illusory Correlation A cognitive fallacy that is not subject to the specific form of an elicited variable but applies at a general level is known as *illusory correlation*. For this, experts assess that two uncorrelated events show a (statistical) dependence or the correlation is (at least) overestimated. Note that this bias is a systematic deviation that experts may make consistently and not simply a false belief that one expert has but not another. Illusory correlation can be present due to prior beliefs that people have about the co-occurrence of events so that a statistical dependence is expected even though actual observations/data do not confirm this.

In their pioneering research in psychodiagnostics, a field of psychology studying the evaluation of personality, Chapman and Chapman (1969) found that medical experts assessed an illusory correlation for the relation of symptoms and personality characteristics. The phenomenon of assuming a correlation where in fact no exists was since then confirmed in different settings and experiments (Eder et al. 2011) and explains various social behaviours, such as the persistence of stereotypes (Hamilton 2015).

One explanation for the (false) expectation of a correlation is that it is triggered by the availability bias. This bias implies that people are influenced considerably by recent experiences and information that can be recalled more easily (Tversky and Kahneman 1973). For instance, one might be overvaluing the recent observation of a co-occurrence of two events by regarding it as a commonly observed co-occurrence. In our project risk example, this could apply when having recently observed a project delay before seeing its cost exceeding a certain value and regarding this co-occurrence as a frequent observation for similar type of projects. Another source of this fallacy is attributed to pre-existing causal beliefs (Bes et al. 2012). In this regard, the prior belief about the correlation stems simply from a false belief about an underlying causal mechanism, as shown in the causality bias.

8.3.3 Implications of Biases for the Elicitation Process

After having presented the main biases that are relevant for eliciting dependence from experts in various forms, we briefly outline the implications that these findings have for the design of the elicitation process.

One finding is that various biases are triggered from the different possible ways that experts might interpret a dependence relationship. In particular, for conditional forms of elicitation, such as conditional probabilities, it is crucial for a facilitator to understand whether the experts might assess the conditional relationships based on similarity/representativeness, causation (e.g. temporal order), or predictive power. As shown, each of these different interpretations can have an effect on the amount and type of information that experts take into consideration when making

assessments. In other words, each of the interpretations biases the outcome of an elicitation in a certain way. While more research is necessary to understand how different interpretations are triggered and affect an assessment, we highlight the importance of structuring experts' knowledge and beliefs about a dependence relationship qualitatively, prior to the quantitative elicitation. This ensures that the decision maker and the experts have the same understanding about the dependent variables and more insight about experts' interpretation might be provided. Further, it helps experts to clarify their own understanding and interpretation. This is essential for ensuring confidence in the resulting elicitation outcome as well as for supporting transparency and reproducibility of the expert judgement process.

In addition, the different interpretations and their implications should be addressed in a training session for the experts, in which misunderstandings, such as semantic ones, are resolved. Then, common pitfalls, such as confusing conditional statements and conjunction of events, can be avoided.

Another finding is that several of the presented fallacies originate with (and are closely linked to) more common biases that are not only observed when assessing dependence, e.g. the representativeness heuristic, base-rate neglect and availability bias. For these, research has addressed debiasing methods through alternative framing of elicitation questions, eliciting variables in various forms and training. Montibeller and Von Winterfeldt (2015) discuss and give an overview to debiasing methods. Further, Table 8.1 lists specific debiasing techniques for the discussed biases.

8.4 Elicitation Process: Preparation/Pre-elicitation

As can be seen in Fig. 8.1, the elicitation process starts already before actually interacting with any experts. The different elements of the preparation (or pre-elicitation) phase ensure that the decision maker's problem is addressed properly and in accordance with the underlying model for which the right variables need to be quantified by suitable experts. In addition, the choices made in this phase allow the experts to assess the uncertain variables as intuitively as possible. In the following, we present the various elements of the this part in more detail.

8.4.1 Problem Identification and Modelling Context

The first step in an elicitation process is the identification of the actual problem at hand in accordance with the decision maker or stakeholder. This step has been termed for instance *background* (Clemen and Reilly 2014) or *preparation* (O'Hagan et al. 2006) and includes typically not just the definition of the elicitation's objective but also the identification of the variables of interest.

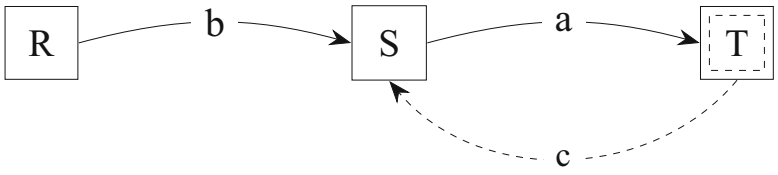


Fig. 8.2 Schematic representation of modelling and elicitation context

When drawing conclusions from one of the earliest experiences on formal processes for probability elicitation, Spetzler and Staël von Holstein (1975) referred to this step as the *deterministic phase*. They describe it as the part of the modelling process in which relevant variables are identified and their relationships are determined before uncertainty assessment is considered (in the *probabilistic phase*).

Likewise for dependence elicitation, a main consideration during this part of the process is to design the elicitation in accordance with the underlying dependence model. A multivariate stochastic model might be pre-determined by the decision maker or is decided upon at this point in accordance with the analyst. In this regard, a broad variety of dependence models exists and their applicability is subject to particular problem situations as they serve different purposes and allow for varying degrees of scrutiny. Werner et al. (2017) review the elicitation for several dependence models and discuss how decisions in the modelling context are related to the elicitation by outlining elicitation strategies for three different, broad dependence modelling situations which are shown in Fig. 8.2.

At this general level, we have a vector of output variables T which depends deterministically on the vector of stochastic variables S in the model. Further, R represents auxiliary variables that are used to evaluate the uncertainty on S. Through the solid arrows uncertainty is propagated as they show the deterministic relationships between variables. Before we provide an illustrative example, note that it is common for there to be dependence between the output variables arising from the functional dependence in arrow (a), in particular when we cannot regard the variables in S as stochastically independent and hence have to model and assess dependence on S.

The first modelling context (a) refers to modelling the dependence relationships in S *directly* before the uncertainty is then propagated through the model (arrow (a)) to T. This is the predominant approach in the literature with common models, such as *Bayesian (Belief) nets* (BNs) (Pearl 1988, 2009), *copulas* (Joe 2014) as well as parametric forms of *multivariate distributions* (Balakrishnan and Nevzorov 2004) and *Bayes linear methods* (Goldstein and Wooff 2007). Given that later in this chapter we will discuss examples in which dependence is elicited for the two former models, we briefly define them here. A BN consist of a directed acyclic graph in which random variables are described by nodes while arcs represent the qualitative dependence relationships between the variables. The direct predecessors/successors of a node are called parent and child nodes accordingly and a BN is quantified

(for example in the discrete case) by assessing for every child node its conditional probability distribution given the state of its parent nodes. With a different modelling focus, a copula might be used to model dependence. Due to Sklar (1959), any multivariate distribution function can be decomposed into its marginal distribution functions and a function which is known as the copula. This can be reversed, meaning that any combination of univariate distribution functions through a copula is a multivariate distribution function. Various common copulas belong to either one of two main families, the *Elliptical* or *Archimedean* one. A main difference is that copulas in the former family are radially symmetric while this is not true for the copulas in the latter, implying a main difference for modelling.

In modelling context (*b*), a set of auxiliary variables is introduced. This is helpful if it allows an easier quantification of the multivariate uncertainty, for instance in the case of too little knowledge for direct modelling and therefore being more comfortable to quantify the uncertainty on the auxiliary variables. In fact, one might chose these so that they can be considered stochastically independent and the dependence in S arises from the complex relationships between the variables in R and S as shown with arrow (*b*). A common modelling type for this context is a regression model.

The last modelling context is (*c*). For that, we consider an alternative set of the output variables (see dotted node) given that a direct assessment of S is too difficult, but the dependence structure must satisfy reasonable conditions on the output variables which are easier to understand and quantify. The alternative set is not identical to T as otherwise we would simply assess its uncertainty directly. The multivariate model is then determined through backward propagation of the uncertainty on S as shown by arrow (*c*). The arrow is dotted to indicate the key difference to the solid arrows. The backward-propagation problem has no unique solution (or even no solution) so that criteria, such as maximum entropy methods, need to be used to select a unique solution, which can then be used to forward-propagate from S to T for looking at other output contexts. A common model type in this situation is Probabilistic Inversion (Kurowicka and Cooke 2006).

For context (*b*) and (*c*), we extend the model (beyond the variables strictly needed to specify the dependence) in order to simplify the necessary understanding of the underlying factors determining the multivariate uncertainty. This influences (or is even determined by) the experts' knowledge on the particular problem. As aforementioned in Sect. 8.3.2, in PRA several methods have been developed for capturing and incorporating implicit uncertainties that are not well enough understood to consider these factors explicitly.

We illustrate the different choices that can be made in the modelling and elicitation context (Fig. 8.2) and how these choices are influenced by the ease with which we can quantify the multivariate uncertainty with our earlier, simplified example from the area of project risk management (Sect. 8.3.2). Recall, we are managing a project which has an overall cost. This is represented by the output variable T (or vector of variables when managing several projects). The overall cost is determined by individual activities, which are important for the project's completion, and each have their own associated costs. The costs of these individual

activities are given by S. If we now want to model the stochastic dependence between these activities' costs, a first option is by doing so directly. The models that are often used for this are the ones mentioned earlier for modelling context (*a*). If the direct modelling of the cost elements is not satisfactory in terms of its outcome, we have the choice to include explanatory variables R, which might help us in understanding the relationships better, and for which we can quantify the uncertainty in the cost easier. The models that are used here are from modelling context (*b*). For our project, environmental uncertainties can be included as explanatory variables if we believe that they (partly) influence the project cost. Lastly, modelling systemic impacts of the project, such as the (un-)availability of qualified staff, can be necessary to capture some subtle dependencies which have been excluded in the earlier modelling contexts. For that, we use modelling techniques from context (*c*). With these, we model the distribution of the overall cost (or features of it) separately which leads to a changed model for the previous joint distributions (as modelled within (*a*) and (*b*)). Similarly, modelling context (*c*) can be applicable if we model a more complex situation with various projects. Then, we can assess the uncertainty for one project and propagate the uncertainty back to the activity costs S and obtain a better understanding about the overall costs of the other projects in T. The underlying idea is that we only ever specify parts of the joint distribution and hence might choose modelling techniques from other contexts to add to our understanding.

The implication for the remainder of the process is that the choices in the different modelling contexts are determined by the level of understanding about the dependencies to be modelled and therefore formulate our variables of interest. These in turn, define the applicability of elicited forms for a satisfactory representation of the experts' information in the model. Therefore, decisions on the model strongly affect the choice of which dependence parameter to elicit as discussed next.

8.4.2 Choice of Elicited Parameters

The next step in the preparation phase is the choice of an appropriate elicited form for the dependence information. Werner et al. (2017) review commonly elicited dependence parameters extensively with regards to the modelling context (Sect. 8.4.1) as well as the assessment burden for experts. These two considerations for choosing an elicited form formulate already main desiderata for this choice, however more are worth discussing.

While some desiderata are the same as for eliciting univariate uncertainty, others are of particular concern when eliciting multivariate quantities. Two desiderata that stem from the univariate case, are: (1) a foundation in probability theory, and (2) the elicitation of observable quantities. A foundation in probability theory ensures a robust operational definition when representing uncertainty. Observable quantities are physically measurable, and having this property may increase the credibility and defensibility of the assessments (Cooke 1991). Moreover, the form

of the elicited variable should allow for a low assessment burden. Kadane and Wolfson (1998) emphasise practicality in this regard. The elicited variables should be formulated so that experts feel comfortable assessing them while their beliefs are captured to a satisfactory degree. For the former, the elicited parameter should be kept intuitively understandable and for the latter, the information given by the experts should be linked (as directly as possible) to the corresponding model. When eliciting dependence, it might be preferred (for instance due to a potential reduction in the assessment burden) to elicit a variable in a different form than the one needed as model input, in which case we need to transform the elicited variable. Then, it is important to measure and control the degree of resemblance between the resulting assessments (through the model) and the dependence information as specified by the expert (Kraan 2002). The transformation of dependence parameters is typically based on assumptions about their underlying bivariate distribution. For instance, when transforming a product moment correlation into a rank correlation, the most common way assumes bivariate normality (Kruskal 1958). Similarly, when transforming a conditional probability into a product moment correlation, we might assume an underlying normal copula (Morales-Nápoles 2010). A potential issue is that positive definiteness is not guaranteed (Kraan 2002), leading to the next desideratum which is coherence. Coherence means that the outcome should be within mathematically feasible bounds. If it is not, it might need to be adjusted such that it still reflects the expert's opinion (as good as possible). Another solution to incoherence is to fix possible bounds for the assessment a priori, even though this can severely decrease the intuitiveness of the assessment. Both solutions are rather pragmatic and show why forms of elicited parameters that result in coherent assessments while being intuitive should be preferred. A last desideratum relates to the (mathematical) aggregation of numerous expert judgements (Sect. 8.6.1). When combining expert judgements, it is desirable to base this combination on the accuracy of experts' assessments measured by performance against empirical data. Therefore, an easily derived dependence parameter from related historical data based on which we can measure such performance is preferred. While there is no query variable that fulfils all of these desirable properties, the desiderata serve as guidance for which elicited parameter to choose under certain circumstances. For instance, an analyst might choose an elicited form that corresponds directly to the model input given a familiarity of the experts with the dependence parameter, therefore having intuitiveness ensured.

At a broad level, most elicited forms can be categorised into *probabilistic* and *statistical* representations. Table 8.2 outlines some main elicited forms in more detail.

We note that the majority of approaches for eliciting dependence fall under the probabilistic umbrella. Probabilistic forms have two main advantages: they (usually) elicit observable quantities and they are rooted in probability theory. Moreover, they are the direct input into various popular models, such as discrete BNs (Pearl 2009, 1988) and its continuous alternative (Hanea et al. 2015). For instance, Werner et al. (2017) found in a review of the literature on dependence elicitation and modelling that 61% of case studies, in which dependence was elicited, a BN was used for

Table 8.2 Overview of elicited forms

	Name	Definition	Framing	Assessment		
			"Consider the variable pair, X and Y. [...]"	independence	positive	negative
probabilistic	Conditional Probability	$P(X > x_i \mid Y \geq y_i)^a$	"[...] Suppose now that Y is observed above your/its median value. What is the probability that X lies also above your/its median value?"	$P(X > x_i)$	$\in (P(X > x_i), 1]$	$\in [0, P(X > x_i))$
	for higher dimensions	$P(X > x_i \mid Y_1 > y_{1,i}, Y_2 > y_{2,i})$	"[...] Suppose that not only Y_1 but also Y_2 is observed above your/its median value. What is now your probability that X is above your/its median value?"	$P(X > x_i \mid Y_1 > y_{1,i})$	as above	as above
	Joint Probability	$P(X \leq x, Y \leq y)$	"[...] What is the probability that both are within the lower (upper) k^{th} percentage of their respective distributions?"	$F_X(x)F_Y(y)^b$	$F_X(x)$ or $F_Y(y)$	towards 0
	Concordance Probability	$\frac{\sum_{a=1}^{n-1}\sum_{b=a+1}^{n} 1_{C*}((x_a,y_a)(x_b,y_b))}{\binom{n}{2}}$ $C* = (x_a - x_b)(y_a - y_b) > 0$	"Consider two independent draws, (x_a, y_a) from population a and (x_b, y_b) from population b. Given $x_a > y_a$ holds for a, what is your probability that $x_b > y_b$ holds for b?"	0.5	towards 1	towards 0
	Expected Conditional Quantiles	$E(F_X(x) \mid Y = y_i)$	"[...] Given the value for Y has been observed at its i^{th} quantile, y_i. What is Xs value in terms of its quantile?"	0.5	towards max $[F_Y(y), 1 - F_Y(y)]$	towards min $[F_Y(y), 1 - F_Y(y)]^c$
verbal/statistical	Direct Correlation	e.g. Kendall's τ, Spearman's ρ or Pearson's ρ^*	"[...] What is the (rank) correlation between them?"	$\rho = 0$	$\rho = 1$	$\rho = -1$
	Verbal	e.g. $\rho = \frac{S_{X,Y}-4}{3}$	"[...] Assess the strength of their relationship as: *strong positive* , *positive*, *slightly positive*, *neutral*, *slightly negative*, *negative* or *strongly negative*."	$S_{X,Y} = 4$	$S_{X,Y} = 7$	$S_{X,Y} = 1$

[a] i refers to i^{th} quantile, e.g. $i = 0.5$ for the median; [b] i.e. the product of the cumulative distributions; [c] if $F_Y(y)$ is above its median

modelling the dependence. The predominant form for the elicited parameter was a conditional probability (point estimates and quantile estimates).

A potential issue with the forms elicited in the probabilistic approaches, such as conditional and joint probabilities, is that they are regarded as non-intuitive and cognitively difficult to assess. Clemen et al. (2000) compare their assessment with other approaches, such as the direct assessment of a correlation coefficient, and found that conditional and joint probabilities were among the worst performances for coherence and in terms of accuracy against empirical data, i.e. not well-calibrated. In particular, joint probability assessments seem cognitively complex.

This is even true for independence assessments which are (typically) among the easier judgements to express. A further concern is the assessment of a conditional probability with a higher dimensional conditioning set, as discussed in Morales-Nápoles (2010) and Morales-Nápoles et al. (2013). The growing conditioning set poses a challenge for experts and this method is (in its current form) difficult to implement. Similarly, expected conditional quantiles (percentiles) are difficult to assess as they require the understanding of location properties for distributions together with the notion of regression towards the mean (Clemen and Reilly 1999).

As a more accurate and intuitive probabilistic way to assess dependence, concordance probabilities have been proposed (Gokhale and Press 1982; Clemen et al. 2000; Garthwaite et al. 2005). A requirement, which may restrict the variables of interest that can be elicited in this way, is the existence of a population to draw from and a certain familiarity with the population.

Alternatively to eliciting probabilistic forms, we can ask experts to assess dependence through statistical dependence measures. While theoretical objections, such as non-observability (Kadane and Wolfson 1998), persist for the elicitation of moments and similarly cross-moments, they seem to perform well with respect to various desiderata (other than theoretical feasibility). For instance, the direct elicitation of a (rank) correlation coefficient is shown to be accurate and intuitive in some studies (Clemen and Reilly 1999; Clemen et al. 2000; Revie et al. 2010; Morales-Nápoles et al. 2015), even though some research is not in agreement with this finding (Gokhale and Press 1982; Kadane and Wolfson 1998; Morgan and Henrion 1990). The contrasting opinions may arise from the difference in normative expertise that the experts in the studies have or as well from the difference in the complexity of the assessed relationships. For example, in the studies which conclude that eliciting a correlation coefficient is accurate and intuitive, the assessed correlations are on rather simple relationships, such as height-weight, or as well on relationships between stocks and stock market indices. This suggests that regarding relationships for which experts have a certain familiarity and maybe even some knowledge about historical data, the direct statistical method is indeed advantageous. Support for this conclusion comes from findings of weather forecasting. Here, experts obtained frequent feedback on correlations which allowed them to become accurate assessors (Bolger and Wright 1994). Neurological research concludes similar findings after evaluating the cognitive activity in a simulation game where participants obtained regular feedback on correlation assessments (Wunderlich et al. 2011).

An indirect statistical approach is the assessment of dependence through a verbal scale that corresponds to correlation coefficients (or other dependence parameters). Clemen et al. (2000) for example provide a scale with seven verbal classifiers. Generally, verbal assessment is seen as intuitive, directly applicable and has therefore enjoyed further consideration. Swain and Guttman (1983) introduce the Technique for Human Error Rate Prediction (THERP) which uses a verbal scale for assigning multivariate uncertainty between human errors. Since its introduction, THERP has been developed extensively in the field of human reliability analysis (HRA) and it has been applied in various industries (see Mkrtchyan et al. 2015 for a review on modelling and eliciting dependence in HRA).

Further, some BN modelling techniques, originating with noisy-OR methods (Pearl 1988), make use of verbal scales. For instance, in the ranked nodes approach, random variables with discretised ordinal scales are assessed by experts through verbal descriptors of the scale (Fenton et al. 2007).

While these are the main approaches for eliciting a dependence parameter, note that when quantifying some models, such as parametric multivariate distributions and regression models, more commonly so called *hyperparameters* are elicited. They allow (through restructuring) for eliciting (mainly) univariate variables.

For a more detailed and comprehensive review of the elicitation methods and elicited forms mentioned above together with some additional ones, see Werner et al. (2017).

8.4.3 Specification of Marginal Distributions

Before dependence can be elicited, the marginal distributions for the variables of interest need to be specified. In some situations, this information is available from historical data and we can simply provide the experts with this data (if they do not know it already). If this is not the case however, we need to elicit the information on the marginal distributions prior to eliciting dependence. This is important as otherwise the experts base their dependence assessments on different beliefs.

Consider for instance, we elicit dependence from experts in a conditional form. If the marginal distributions have not been specified formally, each expert will base their assessment on their own implicit judgement and as a result each assessment will be conditional on different marginal probabilities. While this leads to dependence assessments which are not comparable and therefore cannot be combined for model input, the implicitly specified marginal probabilities are also likely to lack the scrutiny that a formal elicitation process would allow for. In other words, even if eliciting multivariate uncertainty only from a single expert, a formal process for specifying the marginal distributions is still highly encouraged to ensure less biased and better calibrated assessments. Note that if we omit the specification of the marginal distributions, experts might even refuse to assess dependence as they regard the process as flawed.

Various expert judgement methods exist to elicit univariate quantities (as presented elsewhere in this book) and the process is similarly complex as the one presented here. This is an important remark as we need to decide whether all (univariate together with multivariate) variables are elicited in the same session or whether this is done separately. Eliciting all variables in one session is likely to be tiring for the experts while arranging two separate elicitation workshops might be challenging in terms of availability of experts and organisational costs.

8.4.4 Training and Motivation

Training and motivating are likely to improve elicitation outcomes for various reasons, one of which being the effort to mitigate motivational and cognitive biases (Hora 2007). Recall from Sect. 8.3.2 that although it is possible for experts to have an intuitive understanding of probabilistic and/or statistical dependence parameters, psychological research shows that interpreting and assessing dependence is often cognitively difficult and results may be distorted. Therefore, we try to counteract the influence of biases and a main approach to achieve this is to train and motivate experts. As aforementioned, motivational biases are not specific to quantifying multivariate uncertainty and are therefore not discussed in this chapter. Consequently, we will further consider only training (not motivating) experts.

Generally, a training session serves to familiarise the experts with the form in which the query variables are elicited by clarifying its interpretation. For univariate

quantities this (typically) includes introducing the experts to particular location parameters, such as the quantiles of a marginal distribution. This ensures that these are meaningful to the experts and they feel comfortable assessing them. Further, experts are made aware of the main cognitive fallacies that might affect their assessments so that they can reflect on them and make a well-reasoned judgement by taking a critical stance. While this ability is an important characteristic of someone's statistical literacy (Gal 2002), we emphasise a pragmatic approach to training experts as even experienced statisticians often have difficulties with such critical examining and reasoning.

For assessing multivariate uncertainty, the objectives are similar. As concluded in Sect. 8.3.3, main determinants of cognitive biases when assessing dependence are the different interpretations of the elicited forms (in particular of the conditional form). Recall that causal, predictive as well as similarity-based interpretations have a misleading influence on assessments. Therefore, a first focus of an effective training is on explaining the correct interpretation of the dependence parameter to be elicited. This involves an emphasis on the probabilistic and statistical features, such as randomness, in contrast to causal, predictive as well as similarity-based relationships. For instance, causal relationships are often regarded as deterministic, i.e. if Y is understood as the cause of X, then it follows that $P(X|Y) = 1$ as X is always present when Y is present. However, $P(X|Y) = 1$ is not claiming a causal relationship and we might need to account for other factors that affect X and Y (Díaz et al. 2010). As aforementioned, the confusion of the inverse as well as the causality heuristic (Sect. 8.3.2) are two main biases that can be explained by such a misleading interpretation. In this regard, some researchers have mentioned their concern about the language that is used in many statistics textbooks to teach fundamental concepts such as independence (Díaz et al. 2010). For instance, the phrase "whenever Y has no effect on X" is used to explain that two variables, X and Y, are independent and their joint distribution is simply the product of their margins. However, for many experts, the term "effect" might imply a causal relationship. This shows that training on the elicited form should also address any semantic misunderstandings at this step of the elicitation process.

In the same manner, we can address the other misinterpretations. For example, in order to avoid that conditional assessments are based on similarity, i.e. resemblance of X for Y, we should stress that the assessments might also be influenced by other factors. As such, a specific outcome, such as a certain diagnosis, can be *typical* for a certain disease but still unlikely (O'Hagan et al. 2006).

While probabilistic reasoning is commonly included in school curricula, its teaching is often done through formula-based approaches and neglects real-world random phenomena (Batanero and Díaz 2012). Therefore, it is common that experts hold misconceptions on probabilistic/statistical reasoning which are hard to eradicate. In fact, they might even consider this kind of reasoning as counterintuitive. A possibility to enhance a better understanding of these concepts might be to complement the practice of forming probability judgements and providing feedback on training questions (as commonly done before elicitations) with simulation-based approaches. There is empirical evidence that multimedia supported learning

environments successfully support students in building adequate mental models when teaching the concepts of correlation (Liu and Lin 2010) and conditional probability (Eichler and Vogel 2014).

Once the experts are familiar with the elicited form and its correct interpretation, an additional focus of the training session is on outlining the common biases as identified in Sect. 8.3.2. This allows the experts to obtain a better conceptual understanding and we can address potential issues more specifically, such as recognising that a conditional probability involves a restriction in the sample space, distinguishing joint and conditional probabilities or as well distinguishing the inverses.

8.5 Elicitation Process: Elicitation

After the preparation/pre-elicitation phase is concluded, the actual elicitation starts. Note that this is the phase in the overall process in which the facilitator works interactively with the experts, first when supporting experts to structure their knowledge and beliefs (or rationale), and second when eliciting the uncertain variables quantitatively. We will explain both steps in more detail below.

8.5.1 Knowledge and Belief Structuring

Neglecting existing knowledge and data that can be relevant for an assessment is another reason for biased elicitation outcomes in addition to misinterpreting the elicited form (Sect. 8.3.3). However, experts often have cognitive difficulties in *exploring* the underlying sample space to a satisfactory degree. Therefore, they need support for making better use of their knowledge and beliefs, a procedure we call *structuring* or which is also known as *knowledge evocation* (Browne et al. 1997). Apart from mitigating biases, structuring experts' knowledge and beliefs about a joint distribution prior to eliciting dependence quantitatively is essential for ensuring confidence in the later assessment as well as for supporting transparency and reproducibility of the expert judgement process. In fact, when quantifying multivariate uncertainties, identifying the factors that are relevant to the particular problem is a main outcome of the structured expert judgement process. In other words, knowledge structuring allows for obtaining an insight into the details of experts' understanding about the dependence relationships, thus their *rationale*.

Howard (1989) views this step of probability elicitation as the most challenging one in the process. This is due to people possessing knowledge about uncertain events or variables which is composed of many fragmented pieces of information, often all being of high relevance. Further, people typically know more than they think, therefore neglecting this step could result in less informative judgements.

Structuring knowledge might be part of a hybrid approach to dependence modelling in which qualitative, structural information about dependence relationships is specified first, before probabilistic quantification is considered. Typically, graphical models are used to reduce the cognitive load on experts' short term memory, even though other structuring methods, such as directed questions (checklist-based approaches) have been proposed (Browne et al. 1997). Some commonly used graphical models are *knowledge maps* (Howard 1989), *event and fault trees* (Bedford and Cooke 2001), *influence diagrams*[5] (Shachter 1988; Howard and Matheson 2005) and BNs (Sect. 8.4.1). Note that we can nevertheless also include a structuring part when quantifying a dependence model with experts which offers no such a graphical representation. In this case, rather than including the result of knowledge structuring in the actual model, we use it solely for supporting the experts. That being said, when reviewing the literature on eliciting dependence in probabilistic modelling, Werner et al. (2017) found that the dependence model, which is used most often together with expert judgement, is in fact a BN. A reason for its popularity is likely that it allows for an intuitive graphical representation. According to Zwirglmaier and Straub (2016), deriving the structure of a BN can be achieved in four ways. First, the structure can be specified through transforming existing probabilistic models of the problem, such as event and fault trees. Such a transformation is straightforward as the necessary structural information is already given in the existing models and it can be sensible as BNs are more flexible. Second, a BN structure can be inferred from some empirical or physical model. Third, the structure can be built based on existing historical data and fourth, it can be elicited from experts. The last way is of most interest for us as it is a common situation that not only the probabilistic information needs to be elicited from experts, but also the qualitative relationships (Pollino et al. 2007; Flores et al. 2011). Further, it corresponds directly to the knowledge structuring part of the process.

Zwirglmaier and Straub (2016) propose to begin the structural elicitation with identifying the relevant variables and to achieve this, they refer for instance to organized interviews (Hanea and Ale 2009). Then, the actual arcs are elicited, either interactively (as we describe below) or through reusable patterns of structures (Fenton and Neil 2013). Last, they deal with unquantifiable variables (e.g. through proxies).

As mentioned before, one way to derive the graphical structure is by eliciting the experts' input on these interactively (Norrington et al. 2008). One advantage of such an interactive procedure is that it allows (typically) for discussion among experts about the justification of nodes and arcs. In other words, pre-existing knowledge is challenged and elaborated on if necessary. Further, experts obtain a greater ownership of the model which they structured themselves so that they are more comfortable in quantifying it later on. A potential difficulty, which needs to be considered, is that the consensus on the final model structure might have been

[5]In the literature on event trees and influence diagrams, the idea of *decomposition* is often mentioned as it describes a "divide and conquer" technique (Hora 2007) that allows to ease the assessment in particular of conditional probabilities (see e.g. Kleinmuntz et al. 1996).

achieved by a dominating expert who dictated the result or due to group-think, i.e. without critical evaluation. Regarding these potential issues, Walls and Quigley (2001) suggest to elicit a structure from each individual expert, whenever there is a concern about not capturing the opinion of less confident experts. Aggregating diverse structural information coherently through rules (as opposed to consensus) is discussed in Bradley et al. (2014). While for hybrid dependence models a combined graphical structure is necessary, in terms of knowledge structuring it is also of interest how sharing knowledge and rationales among experts affects a later assessment. For instance, Hanea et al. (2017) integrate group interaction in a structured protocol for quantitative elicitation as it is shown to be beneficial in assessment tasks.

Besides the initial structuring step, Henrion (1989) mentions the potential necessity to refine a model structure during the actual quantification. In particular, the violation of conditional independence is of concern. By definition of a BN, the successor nodes (children) are conditionally independent given their parents. If this is not the case when observing the final model, an additional node is required. Pearl (1988) regards conditional independence therefore as a guiding principle as where it fails, further clarification about an assumed, hidden variable is needed.

8.5.2 Quantitative Elicitation

After structuring experts' knowledge and beliefs about the factors that influence the variable(s) of interest, the quantitative assessment follows. This step of the process is also named *encoding* (Spetzler and Staël von Holstein 1975). In this step, experts assess the variable(s) of interest in the form that was chosen to be appropriate with respect to various desiderata (Sect. 8.4.3).

The main considerations herewith are similar to those of eliciting univariate uncertainty. Likewise, we need to decide on how much interaction between the experts we allow for (we address the aggregation of assessments in Sect. 8.6.1). Further, at least one facilitator is present to answer questions regarding the understanding of the query variables. Prior to the session, experts should have received a briefing document which helps them to familiarise themselves with the purpose and structure of the elicitation (Cooke and Goossens 2004).

As there are no differences to univariate uncertainty elicitation in this part, we devote the remainder of this sub-section to illustrating an exemplary assessment which has been used similarly in an actual dependence elicitation problem. Morales-Nápoles et al. (2015) and Morales-Nápoles et al. (2016a) elicit and quantify dependence between rain amount and rain duration in the Netherlands through conditional exceedance probabilities. The elicited results are used as model input for quantifying parametric copulas. Modelling dependence in this way informs resilience analysis for critical components of road networks, such as tunnels and road sections. The aim of this analysis is to improve the understanding about the effects of extreme rainfall for the development of probabilistic models in reliable infrastructure risk analysis.

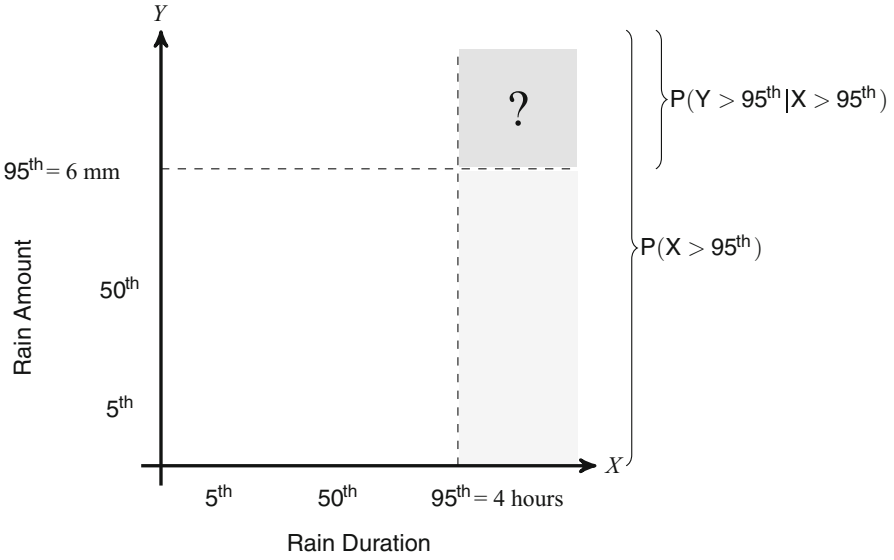


Fig. 8.3 Exemplary elicitation question with visualisation

Figure 8.3 shows a way of presenting experts with the elicitation question:

> For Rotterdam, NL, consider all samples for which the rain duration in *hours* (X) is larger than its 95th quantile (4 h). What is the percentage of this set of samples, for which the rain amount in mm (Y) is also larger than its 95th quantile (6 mm)?
>
> This can be expressed as $P(Y \geq$ 95th quantile$|X \geq$ 95th quantile) or likewise as $P(Y \geq 6\,\text{mm}|X \geq 4\,\text{h})$.
>
> Please provide your assessment: _____

The inclusion of a visualisation can be helpful for experts to obtain a better understanding about the framing of the elicitation question.

8.6 Elicitation Process: Post-elicitation

The last phase in the overall elicitation process (Fig. 8.1) is the post-elicitation part. The two main steps that are of importance here are aggregating the assessments of various experts and providing feedback to the experts. We address both steps in more detail below.

8.6.1 Aggregation of Expert Judgements

In order to capture a broad perspective on the uncertainties that we model and quantify, we (usually) elicit judgements from a variety of experts. Therefore, a main aspect of the post-elicitation phase is the aggregation (or combination) of the assessments from several experts.

As in the univariate case, a distinction at a broad level is made between *behavioural* and *mathematical* (or algorithmic) aggregation methods. The first type aims at reaching consensus so that the outcome is a single assessment upon which the group of experts has agreed. This might be achieved within a group elicitation session or through methods, such as Delphi (Rowe and Wright 2001). Given that these methods are the same as for univariate elicitation, they are not further discussed here. Recall however that a potential shortcoming of these methods (in the univariate as well as multivariate case) is that the consensus might be reached through one expert dominating the elicitation discussion or even dictating the elicitation's outcome (French 2011).

For aggregating judgements mathematically, in particular two approaches are common. The first is the *Bayesian approach* which allows for modelling quality aspects of individual expert distributions, for example overconfidence. The second approach is a *pooling function* which is typically seen as more robust and easier to use (Hora and Kardeş 2015).

For Bayesian aggregation, we apply Bayes' Theorem (Sect. 8.3.2) while regarding the expert judgements as data. If we are interested in an event or unknown quantity x, we elicit its probability or set of quantiles and obtain the experts' individual prior opinions, $f_{0,e}(x)$ for experts $e = 1, 2, \ldots, E$. We denote the set of elicited distributions as $\underline{D} = (f_{0,1}(x), \ldots, f_{0,E}(x))$, and get the combined posterior distribution for x, $f_{1,DM}(x|\underline{D})$ through $f_{1,DM}(x|\underline{D}) \propto f_{0,DM}(x)L_{DM}(\underline{D}|x)$. It is then necessary to elicit the likelihood function of observing $\underline{D}$ given x, i.e. $L_{DM}(\underline{D}|x)$ (Wilson 2017). A Bayesian aggregation model which has been used more commonly is Mosleh and Apostolakis (1986).

A pooling function on the other hand assigns weights to individual assessments to derive a weighted combination of the experts' judgements. The weights are either equal for each expert or they reflect an expert's competence or performance (in terms of statistical accuracy, if empirical data can be used for measuring this). For equal as well as performance-based weighting, all weights are non-negative and sum to one. A commonly used pooling function is linear averaging, for which the combined assessment is $DM_{(f_1(x),\ldots,f_n(x))} = \sum_{e=1}^{E} w_e f_e(x)$, with w_e being the weight of expert e. Alternatively, other pooling methods exist, such as logarithmic pooling, for which the combined assessment is defined as $DM_{(f_1(x),\ldots,f_n(x))} = k \prod_{e=1}^{E} f_e(x)^{w_e}$ where k is a normalising constant.

Linear pooling functions originate with Stone (1961) and DeGroot (1974) and the legitimacy of their application from an axiomatic perspective is primarily based on *event-wise independence* (or the weak set-wise function) and *unanimity preservation* (Aczél and Wagner 1980; McConway 1981; Dietrich and List 2016).

The first axiom implies that the collective probability of an event is only determined by the individual probabilities for that specific event (and not that of other ones). Unanimity preservation holds that if all experts give the same assessment, then this will be the collective one.

For aggregating dependence assessments, mainly linear pooling functions have been used (Werner et al. 2017), which is why we address them in more detail. Before we discuss these however, note that a possible concern with mathematical aggregation in the multivariate case is that not all dependence assessments are preserved. For instance, a linear combination of correlation matrices is still a correlation matrix, however conditional independencies such as in a BN are not preserved. Further, an axiomatic issue might be that of preserving *probabilistic independence* which ensures that if all experts regard two variables as (conditionally) independent, then this is preserved in the combined assessment. For several pooling functions (e.g. linear as well as logarithmic ones) this is problematic. However, it might be argued that unless independence assessments are also based on structural judgements (Sect. 8.5.1), i.e. they are not purely accidental, this normative constraint is questionable (Bradley et al. 2014). Note that this is a question of whether one regards dependence information as fully represented by probabilistic (un-)conditional dependence or only in addition to structural judgements in form of graphical representations (such as in BNs). As we have emphasised in Sect. 8.5.1 that structural information should be elicited either within the same modelling framework or separately, the independence axiom is not of concern and we regard linear pooling methods as applicable for dependence information.

Equal Weighting One option to set weights in a linear pooling function is by equally weighting all assessments (simple average). When eliciting correlation parameters directly, overall accuracy improved in that way through adding experts (Winkler and Clemen 2004). The authors tested the robustness by removing/adding experts and found that the mean absolute error (MAE) decreased when the number of experts increased.

Performance-Based Weighting Alternatively, Winkler and Clemen (2004) also showed that taking the average of only the top performing cohort of experts (in terms of lowest MAE) instead of the whole set of experts reduces the overall MAE further. This finding is consistent with expert judgement studies for univariate quantities (Cooke and Goossens 2008) and therefore motivated the idea of using a measure of calibration to assess experts' performance in terms of statistical accuracy as a score for multivariate assessments. Before we introduce this score, note that there is an indication that a common calibration method for univariate expert judgements (Cooke 1991) might not be feasible for aggregating dependence assessments (Morales-Nápoles et al. 2013).

The first and only calibration score for multivariate assessments (according to the authors' knowledge) is the dependence calibration score introduced in Morales-Nápoles and Worm (2013) which is based on the *Hellinger distance*. In order to assess this score (similar to *Cooke's Classical* model (Cooke 1991)) seed variables known to the facilitator but not the experts are elicited in addition to the target

variables. Then, two bivariate copulas f_C (a copula model used for calibration purposes) and f_E (a copula estimated by expert opinions) are used to derive the Hellinger distance, H, which is defined as:

$$H(f_C, f_E) = \iint_{[0,1]^2} \sqrt{\frac{1}{\sqrt{2}}(\sqrt{f_C(u,v)} - \sqrt{f_E(u,v)})^2} \mathrm{dudv}$$

In Abou-Moustafa et al. (2010) an overview of different distances between distributions is given. If the distributions are Gaussian, these distances can be written in terms of the parameters of the Gaussian distributions (i.e. the mean and covariance matrix). Under the Gaussian copula assumption, H may be parametrised by two correlation matrices:

$$H_G(\Sigma_C, \Sigma_E) = \sqrt{1 - \frac{\det(\Sigma_C)^{1/4}\det(\Sigma_E)^{1/4}}{(\frac{1}{2}\det(\Sigma_C) + \frac{1}{2}\det(\Sigma_E))^{1/2}}}$$

Here Σ_C is a correlation matrix used for calibration purposes and Σ_E the one estimated by experts. The *d-calibration* or dependence calibration score is:

$$D = 1 - H$$

The score is 1 if an expert's assessments correspond to the calibration model exactly. Conversely, it differs from 1 as the expert's opinion differs from the calibration model. Under the Gaussian assumption, i.e. when using H_G, the score approaches 1 as Σ_E approximates Σ_C element-wise and it decreases as H_G differs from H_C element-wise. A score equal to zero means that at least two variables are linearly dependent in the correlation matrix used for calibration purposes and the expert fails to express this. Or contrary to this, an expert expresses perfect linear dependence between two variables when this is not the case. For more details, see Morales-Nápoles et al. (2016b). In the same paper (Morales-Nápoles et al. 2016b), the method discussed in Morales-Nápoles and Worm (2013) is extended by using the Hellinger distance to compare a Gumbel copula generated from precipitation data with a copula constructed from experts' assessments of tail dependence between rain amount and duration in Rotterdam and De Bilt, in the Netherlands. The experts' assessments are obtained by a similar framing as shown in Sect. 8.5.2 and varying the elicited quantiles, e.g. 50th and 95th (see Morales-Nápoles (2010) for more details). An overview of the results in given in Table 8.3.

In this study, the combination of expert opinions based on the dependence calibration score outperforms individual expert opinions as well as weighting experts equally. In fact, the equal weights approach does not give satisfactory results. We observe that the performance-based aggregation is much closer to the actual empirical rank correlation. Further, it was noticed that experts with highest calibration scores for univariate assessments are not necessarily the experts with the highest dependence calibration score.

Table 8.3 Dependence calibration results based on rank correlation, Gaussian (H_G) and Hellinger (H) distance (Morales-Nápoles et al. 2016b)

	Rotterdam	De Bilt	Rotterdam	De Bilt
Name	X > 0.95	X > 0.95	X > 0.5	X > 0.5
$1 - H_G$				
Expert 1	0.809	0.812	0.894	0.897
Expert 2	0.889	0.892	0.766	0.769
Expert 3	0.960	0.963	0.853	0.856
Expert 4	0.746	0.769	0.960	0.963
Expert 5	0.832	0.812	0.979	0.982
Expert 6	0.733	0.736	0.730	0.733
Expert 7	0.787	0.790	0.730	0.733
Expert 8	0.809	0.812	0.894	0.897
$1 - H$				
Expert 1	0.822	0.825	0.900	0.903
Expert 2	0.895	0.899	0.784	0.787
Expert 3	0.962	0.965	0.862	0.865
Expert 4	0.767	0.787	0.962	0.965
Expert 5	0.843	0.825	0.980	0.983
Expert 6	0.756	0.759	0.753	0.756
Expert 7	0.802	0.805	0.753	0.756
Expert 8	0.822	0.825	0.900	0.903
Calibration score				
Equal weighting	0.814	0.817	0.837	0.841
Performance-based weighting	0.960	0.963	0.979	0.982
Rank correlation (result)				
Equal weighting	0.264	0.264	0.326	0.326
Performance-based weighting	0.578	0.578	0.608	0.608
Realisation	0.622	0.617	0.622	0.617

In order to combine dependence assessments, experts are weighted according to their dependence calibration score. Similar to the univariate case, a cut-off level is established, either chosen by the facilitator or by optimising the performance of the combination. If an individual expert falls below this level, their score will be unweighted for the pooling function.

8.6.2 Feedback and Robustness Analysis

Similar to eliciting univariate uncertainty, one of the final steps of the dependence elicitation process is testing the robustness of elicited results and providing feedback to the experts after a combined assessment has been constructed. While

this procedure is not much different for the multivariate case, it should be noted that many dependence models produce graphical outputs, such as scatter plots. Depending on the experts' understanding of the graphical output and their willingness to examine such outputs, it might be possible to feedback such a visualisation and assess their agreement with it.

8.7 Conclusions

In this chapter, we have presented the main considerations for eliciting multivariate uncertainty from experts. As shown, there are several important adjustments that are necessary when eliciting dependence given that many of the findings from expert judgement processes for univariate quantities are not readily applicable.

A first remark for concluding this chapter is that a few areas still lack insight to a considerable extent. For instance, we have discussed that the biases and heuristics which influence dependence assessments might be mitigated by training and knowledge structuring. In particular, experts' potential misinterpretations of dependence parameters need to be corrected and ways to do so might be informed by the educational literature on teaching concepts such as conditional and joint probabilities. Nevertheless, we need to acknowledge that experiences here might not be directly transferable to designing experts' training due to a different understanding of that of students and therefore further research in training design is necessary.

Further, more insight is needed on the exact triggers of the potential biases and their relative influence on judgements. It would be desirable for behavioural researchers to take a similar interest in this field as they do with the more common (typically univariate probability) heuristics and biases. This would allow developing the various (undeveloped) steps in the pre-elicitation phase, e.g. format choices.

In the elicitation phase, in particular the topic of structuring knowledge is identified as a key area for which further research is necessary. For instance, the graphical representation of BNs offers a way to incorporate qualitative dependence information. However issues still remain such as eliciting the structure of highly complex BNs as well as eliciting tail dependencies graphically. Therefore, again, we need to obtain more experiences for this part of the elicitation process.

Lastly, we have discussed that when combining assessments mathematically, more research is necessary for addressing some common desiderata for this step, such as performance-based as well as mathematically coherent aggregation.

Acknowledgements We would like to thank the editors for the feedback on the previous version of this chapter.

References

Abou-Moustafa KT, De La Torre F, Ferrie FP (2010) Designing a metric for the difference between Gaussian densities. In: Angeles J, Boulet B, Clark J J, Kövecses J, Siddiqi K (eds) Brain, body and machine. Springer, Berlin, pp 57–70

Aczél J, Wagner C (1980) A characterisation of weighted arithmetic means. SIAM J Algebr Discrete Methods 1(3):259–260

Ajzen I (1977) Intuitive theories of events and the effects of base-rate information on prediction. J Pers Soc Psychol 35(5):303–314

Allan LG (1980) A note on measurement of contingency between two binary variables in judgment tasks. Bull Psychon Soc 15(3):147–149

Balakrishnan N, Nevzorov VB (2004) A primer on statistical distributions. Wiley, Hoboken

Bar-Hillel M (1980) The base-rate fallacy in probability judgments. Acta Psychol 44(3):211–233

Batanero C, Díaz C (2012) Training school teachers to teach probability: reflections and challenges. Chilean J Stat 3(1):3–13

Bechlivanidis C, Lagnado DA (2013) Does the "why" tell us the "when"? Psychol Sci 24(8):1563–1572

Bedford T, Cooke RM (2001) Probabilistic risk analysis: foundations and methods. Cambridge University Press, Cambridge

Bertsch McGrayne S (2011) The theory that would not die. Yale University Press, New Haven

Bes B, Sloman S, Lucas CG and Raufaste E (2012) Non-Bayesian inference: causal structure trumps correlation. Cogn Sci 36(7):1178–1203

Bolger F, Wright G (1994) Assessing the quality of expert judgment: issues and analysis. Decis Support Syst 11(1):1–24

Borovcnik M, Kapadia R (2014) From puzzles and paradoxes to concepts in probability. In: Chernoff EJ, Sriraman B (eds) Probabilistic thinking. Springer, Dordrecht, pp 35–73

Bradley R, Dietrich F, List C (2014) Aggregating causal judgments. Philos Sci 81(4):491–515

Browne GJ, Curley SP, Benson PG (1997) Evoking information in probability assessment: knowledge maps and reasoning-based directed questions. Manag Sci 43(1):1–14

Carranza P, Kuzniak A (2009) Duality of probability and statistics teaching in French education. In: Batanero C (ed) Joint ICMI/IASE study: teaching statistics in school mathematics, p 5

Chapman LJ, Chapman JP (1969) Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. J Abnorm Psychol 74(3):271–280

Clemen RT, Reilly T (1999) Correlations and copulas for decision and risk analysis. Manag Sci 45(2):208–224

Clemen RT, Reilly T (2014) Making hard decisions with decision tools. South-Western, Mason

Clemen RT, Fischer GW, Winkler RL (2000) Assessing dependence: some experimental results. Manag Sci 46(8):1100–1115

Cooke RM (1991) Experts in uncertainty: opinion and subjective probability in Science. Oxford University Press, New York

Cooke RM (2013) Uncertainty analysis comes to integrated assessment models for climate change and conversely. Climatic Change 117(3):467–479

Cooke RM, Goossens LHJ (1999) Procedures guide for structured expert judgment. Commission of the European Communities, Brussels

Cooke RM, Goossens LHJ (2004) Expert judgement elicitation for risk assessments of critical infrastructures. J Risk Res 7(6):643–656

Cooke RM, Goossens LHJ (2008) TU Delft expert judgment data base. Reliab Eng Syst Saf 93(5):657–674

Costello FJ (2009) How probability theory explains the conjunction fallacy. J Behav Decis Mak 22(3):213–234

Dawes RM (1988) Rational choice in an uncertain world. Harcourt, Brace, Jovanovich, New York

DeGroot MH (1974) Reaching a consensus. J Am Stat Assoc 69(345):118–121

Díaz C, Batanero C, Contreras JM (2010) Teaching independence and conditional probability. Boletín de Estadística e Investigación Operativa 26(2):149–162
Dietrich F, List C (2016) Probabilistic Opinion Pooling. In: Hajek A, Hitchcock C (eds) The oxford handbook of probability and philosophy. Oxford handbooks. Oxford University Press, Oxford, pp 179–207
DuCharme WM (1970) Response bias explanation of conservative human inference. J Exp Psychol 85(1):66–74
Eddy DM (1982) Probabilistic reasoning in clinical medicine: problems and opportunities. In: Kahneman D, Slovic P and Tversky A (eds) Judgment under uncertainty: heuristics and biases. Cambridge University Press, New York, pp 249–267
Eder AB, Fiedler K, Hamm-Eder S (2011) Illusory correlations revisited: the role of pseudocontingencies and working-memory capacity. Q J Exp Psychol 64(3):517–532
Edwards W (1965) Optimal strategies for seeking information: models for statistics, choice reaction times, and human information processing. J Math Psychol 2(2):312–329
EFSA=European Food and Safety Authority (Bolger F, Hanea AM, O'Hagan A, Mosbach-Schulz O, Oakley J, Rowe G, Wenholt M) (2014) Guidance on expert knowledge elicitation in food and feed safety risk assessment. EFSA J 12(6):3734
Eichler A, Vogel M (2014) Three Approaches for Modelling Situations with Randomness. In: Chernoff EJ, Sriraman B (eds) Probabilistic thinking. Springer, Dordrecht, pp 75–99
Einhorn HJ, Hogarth RM (1986) Judging probable cause. Psychol Bull 99(1):3–19
Falk R (1983) Conditional probabilities: insights and difficulties. In: Tall D (ed) Proceedings of the third international conference for the psychology of mathematics education, Warwick, 1983
Fenton NE, Neil M (2013) Risk Assessment and Decision Analysis with Bayesian Networks. Taylor and Francis Group, Boca Raton
Fenton NE, Neil M, Caballero JG (2007) Using Ranked Nodes to Model Qualitative Judgments in Bayesian Networks. IEEE Trans Knowl Data Eng 19(10):1420–1432
Fiedler K, Brinkmann B, Betsch T, Wild B (2000) A sampling approach to biases in conditional probability judgments: beyond base rate neglect and statistical format. J Exp Psychol Gen 129(3):399–418
Flores MJ, Nicholson AE, Brunskill A, Korb KB, Mascaro S (2011) Incorporating expert knowledge when learning Bayesian network structure: a medical case study. Artif Intell Med 53(3):181–204
Fountain J, Gunby P (2011) Ambiguity, the certainty illusion, and the natural frequency approach to reasoning with inverse probabilities. N Z Econ Pap 45(1-2):195–207
French S (2011) Aggregating expert judgement. Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Ser A Mate 105(1):181–206
Gal I (2002) Adults' statistical literacy: meanings, components, responsibilities. Int Stat Rev 70(1):1–25
Garthwaite PH, Kadane JB, O'Hagan A (2005) Statistical methods for eliciting probability distributions. J Am Stat Assoc 100(470):680–701
Gavanski I, Hui C (1992) Natural sample spaces and uncertain belief. J Pers Soc Psychol 63(5):766–780
Gilovich T, Griffin D, Kahneman D (2002) Heuristics and biases: the psychology of intuitive judgment. Cambridge University Press, New York
Gokhale DV, Press SJ (1982) Assessment of a prior distribution for the correlation coefficient in a bivariate normal distribution. J R Stat Soc Ser A (General) 145:237–249
Goldstein M, Wooff D (2007) Bayes linear statistics, theory and methods. Wiley, Chichester
Hamilton DL (2015) Cognitive processes in stereotyping and intergroup behavior. Psychology Press, New York
Hanea A, McBride MF, Burgman MA, Wintle BC, Fidler F, Flander L, Twardy CR, Manning B, Mascaro S (2017) Investigate discuss estimate aggregate for structured expert judgement. Int J Forecast 33:267–279
Hanea A, Morales Nápoles O, Ababei D (2015) Non-parametric Bayesian networks: improving theory and reviewing applications. Reliab Eng Syst Saf 144:265–284

Hanea D, Ale B (2009) Risk of human fatality in building fires: a decision tool using Bayesian networks. Fire Saf J 44(5):704–710
Hastie R (2016) Causal thinking in judgments. In: Keren G, Wu G (eds) The Wiley Blackwell handbook of judgment and decision making. Wiley, Chichester, pp 590–628
Hastie R, Dawes RM (2001) Rational choice in an uncertain world. Sage, Thousand Oaks
Henrion M (1989) Some practical issues in constructing belief networks. In: Kanal L, Levitt T, Lemmer J (eds) Uncertainty in artificial intelligence. Elsevier Science Publishing Company, New York, pp 132–139
Hora SC (2007) Eliciting Probabilities from Experts. In: Edwards W, Miles RF Jr, Von Winterfeldt D (eds) Advances in decision analysis - from foundations to applications. Cambridge University Press, New York, pp 129–163
Hora SC, Kardeş E (2015) Calibration, sharpness and the weighting of experts in a linear opinion pool. Ann Oper Res 229(1):429–450
Howard RA (1989) Knowledge maps. Manag Sci 35(8):903–922
Howard RA, Matheson JE (2005) Influence diagrams. Decis Anal 2(3):127–143
Hume D (1748/2000) An enquiry concerning human understanding. Beauchamp TL (ed) Oxford University Press, New York
Joe H (2014) Dependence modeling with copulas. Chapman and Hall, Boca Raton
Kadane J, Wolfson LJ (1998) Experiences in elicitation. J R Stat Soc Ser D (Stat) 47(1):3–19
Kahneman D, Frederick S (2002) Representativeness revisited: attribute substitution in intuitive judgment. In: Gilovich T, Griffin D, Kahneman D (eds) Heuristics of intuitive judgment: extensions and applications, 1st edn. Cambridge University Press, New York, pp 19–49
Kahneman D, Tversky A (1972) Subjective probability: a judgment of representativeness. Cogn Psychol 3(3):430–454
Kahneman D, Tversky A (1973) On the psychology of prediction. Psychol Rev 80(4):237–251
Kao SF, Wasserman EA (1993) Assessment of an information integration account of contingency judgment with examination of subjective cell importance and method of information presentation. J Exp Psychol Learn Mem Cogn 19(6):1363–1386
Keeney RL, Von Winterfeldt D (1991) Eliciting probabilities from experts in complex technical problems. IEEE Trans Eng Manag 38(3):191–201
Keren G, Teigen KH (2006) Yet another look at Heuristics and biases approach. In: Koehler DJ, Harvey N (eds) Blackwell handbook of judgment and decision making, 1st edn. Blackwell Publishing, Oxford, p 89–109
Kleinmuntz DN, Fennema MG, Peecher ME (1996) Conditioned assessment of subjective probabilities: Identifying the benefits of decomposition. Organ Behav Hum Decis Process 66(1):1–15
Koehler JJ (1996) The base rate fallacy reconsidered: descriptive, normative, and methodological challenges. Behav Brain Sci 19(1):1–17
Kraan BCP (2002) Probabilistic inversion in uncertainty analysis: and related topics. PhD Thesis, Delft University of Technology, The Netherlands
Kruskal WH (1958) Ordinal measures of association. J Am Stat Assoc 53(284):814–861
Kurowicka D, Cooke RM (2006) Uncertainty analysis with high dimensional dependence modelling. Wiley, Chichester
Kynn M (2008) The "heuristics and biases" bias in expert elicitation. J R Stat Soc Ser A (Stat Soc) 171(1):239–264
Lad F (1996) Operational subjective statistical methods: a mathematical, philosophical, and historical introduction. Wiley-Interscience, New York
Lagnado DA, Shanks DR (2002) Probability judgment in hierarchical learning: a conflict between predictiveness and coherence. Cognition 83(1):81–112
Lagnado DA, Sloman SA (2006) Inside and outside probability judgment. In: Koehler DJ, Harvey N (eds) Blackwell handbook of judgment and decision making, 1st edn. Blackwell Publishing, Oxford, pp 157–176

Liu TC, Lin YC (2010) The application of Simulation Assisted Learning Statistics (SALS) for correcting misconceptions and improving understanding of correlation. J Comput Assis Learn 26(2):143–158

Mandel DR, Lehman DR (1998) Integration of contingency information in judgments of cause, covariation, and probability. J Exp Psychol Gen 127(3):269–285

McConway KJ (1981) Marginalization and linear opinion pools. J Am Stat Assoc 76(374):410–414

McKenzie CR, Mikkelsen LA (2007) A Bayesian view of covariation assessment. Cogn Psychol 54(1):33–61

Meder B, Gigerenzer G (2014) Statistical thinking: no one left behind. In: Chernoff EJ, Sriraman B (eds) Probabilistic thinking. Springer, Dordrecht, pp 127–148

Medin DL, Coley JD, Storms G, Hayes BL (2003) A relevance theory of induction. Psychon Bull Rev 10(3):517–532

Meehl PE, Rosen A (1955) Antecedent probability and the efficiency of psychometric signs, patterns, and cutting scores. Psychol Bull 52:194–216

Merkhofer MW (1987) Quantifying judgmental uncertainty: Methodology, experiences, and insights. IEEE Trans Syst Man Cybern 17(5):741–752

Mitchell CJ, De Houwer J, Lovibond PF (2009) The propositional nature of human associative learning. Behav Brain Sci 32(2):183–198

Mkrtchyan L, Podofillini L, Dang VN (2015) Bayesian belief networks for human reliability analysis: a review of applications and gaps. Reliab Eng Syst Saf 139:1–16

Montibeller G, Von Winterfeldt D (2015) Cognitive and motivational biases in decision and risk analysis. Risk Anal 35(7):1230–1251

Morales-Nápoles O (2010) Bayesian belief nets and vines in aviation safety and other applications. PhD Thesis, Delft University of Technology, The Netherlands

Morales-Nápoles O, Worm DTH (2013) Hypothesis testing of Multidimensional probability distributions. WP4 GAMES2R TNO Report No. 0100003764, TNO, The Netherlands

Morales-Nápoles O, Hanea AM, Worm DTH (2013) Experimental results about the assessments of conditional rank correlations by experts: Example with air pollution estimates. In: Proceedings of the 22nd European safety and reliability conference safety, reliability and risk analysis. University of Amsterdam, Amsterdam, 2013

Morales-Nápoles O, Worm DTH, Abspoel LM, Huibregtse E, Courage W (2015) Trends and uncertainties regarding rain intensity in the Netherlands. TNO 2015 R10009, TNO

Morales-Nápoles O, Paprotny D, Worm DTH, Abspoel LM, Courage W (2016a) Characterization of precipitation through copulas and expert judgement for risk assessment of infrastructure. In: Proceedings of the 25th European safety and reliability conference safety, reliability and risk analysis. Eidgenössische Technische Hochschule Zürich, Zürich, 2016

Morales-Nápoles O, Paprotny D, Worm D, Abspoel-Bukman L, Courage W (2017) Characterization of precipitation through copulas and expert judgement for risk assessment of infrastructure. ASCE-ASME J Risk and Uncertain Eng Syst A Civ Eng 3(4), 04017012

Morgan MG, Henrion M (1990) Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis. Cambridge University Press, New York

Mosleh A, Apostolakis G (1986) The assessment of probability distributions from expert opinions with an application to seismic fragility curves. Risk Anal 6(4):447–461

Norrington L, Quigley J, Russell A, Van der Meer R (2008) Modelling the reliability of search and rescue operations with Bayesian belief networks. Reliab Eng Syst Saf 93(7):940–949

O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, Oakley JE, Rakow T (2006) Uncertain judgements: eliciting experts' probabilities. Wiley, Chichester

Over D (2004) Rationality and the normative/descriptive distinction. In: Koehler DJ, Harvey N (eds) Blackwell handbook of judgment and decision making, 1st edn. Blackwell Publishing, Oxford, pp 3–18

Pearl J (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, San Francisco

Pearl J (2009) Causality: models, reasoning, and inference. Cambridge University Press, Cambridge

Pollatsek A, Well AD, Konold C, Hardiman P, Cobb G (1987) Understanding conditional probabilities. Organ. Behav Hum Decis Process 40(2):255–269

Pollino CA, Woodberry O, Nicholson A, Korb K, Hart BT (2007) Parameterisation and evaluation of a Bayesian network for use in an ecological risk assessment. Environ Model Softw 22(8):1140–1152

Pólya G (1941) Heuristic reasoning and the theory of probability. Am Math Mon 48(7):450–465

Ramsey FP (1926) Truth and probability. In: Braithwaite RB (ed) Ramsey FP (1931) The foundations of mathematics and other logical essays. Kegan, Paul, Trench, Trubner and Co, London, pp 156–198

Revie M, Bedford T, Walls L (2010) Evaluation of elicitation methods to quantify Bayes linear models. J Risk Reliab 224(4):322–332

Rottman BM, Hastie R (2014) Reasoning about causal relationships: inferences on causal networks. Psychol Bull 140(1):109–139

Rowe G, Wright G (2001) Expert opinions in forecasting: the role of the Delphi technique. In: Armstrong JS (ed) Principles of forecasting. Springer, New York, pp 125–144

Russell B (1912) On the notion of cause. Proc Aristot Soc 13:1–26

Shachter RD (1988) Probabilistic inference and influence diagrams. Oper Res 36(4):589–604

Simon HA (1957) Models of man: social and rational. Wiley, New York

Sklar M (1959) Fonctions de répartitionà n dimensions et leurs marges. Publ Inst Statist Univ Paris 8:229–231

Sloman SA, Lagnado D (2005) The problem of induction. In: Holyoak KJ, Morrison RG (eds) The Cambridge handbook of thinking and reasoning. Cambridge University Press, New York, pp 95–116

Smedslund J (1963) The concept of correlation in adults. Scand J Psychol 4(1):165–173

Spetzler CS, Staël von Holstein CA (1975) Exceptional paper-probability encoding in decision analysis. Manag Sci 22(3):340–358

Spirtes P, Glymour CN, Scheines R (2000) Causation, prediction, and search. MIT Press, Cambridge

Staël von Holstein CA, Matheson JE (1979) A manual for encoding probability distributions. In: Defense advanced research projects agency, decisions and designs. SRI International, Menlo Park, CA

Stanovich KE, West RF (2000) Individual differences in reasoning: implications for the rationality debate? Behav Brain Sci 23:645–726

Stone M (1961) The opinion pool. Ann Math Stat 32(4):1339–1342

Suppes P (1970) A probabilistic theory of causality. North-Holland Publishing, Amsterdam

Swain AD, Guttman HE (1983) Handbook of human reliability analysis with emphasis on nuclear power plant applications. US Nuclear Regulatory Commission. NUREG/CR-1278, Washington, DC

Tentori K, Crupi V, Russo S (2013) On the determinants of the conjunction fallacy: probability versus inductive confirmation. J Exp Psychol Gen 142(1):235–255

Tversky A, Kahneman D (1973) Availability: a heuristic for judging frequency and probability. Cogn Psychol 5(2):207–232

Tversky A, Kahneman D (1980) Causal schemas in judgments under uncertainty. Progress Soc Psychol 1:49–72

Tversky A, Kahneman D (1983) Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. Psychol Rev 90(4):293–315

USNRC = US Nuclear Regulatory Commission (1975) Reactor safety study: An assessment of accident risks in U.S. commercial nuclear power plants. NUREG 75–014, Washington, DC

USNRC = US Nuclear Regulatory Commission (1987) Reactor risk reference document. NUREG-1150, Washington, DC

Utts J (2003) What educated citizens should know about statistics and probability. Am Stat 57(2):74–79

Villejoubert G, Mandel DR (2002) The inverse fallacy: an account of deviations from Bayes's theorem and the additivity principle. Mem Cogn 30(2):171–178

Walls L, Quigley J (2001) Building prior distributions to support Bayesian reliability growth modelling using expert judgement. Reliab Eng Syst Saf 74(2):117–128

Werner C, Bedford T, Cooke RM, Hanea AM, Morales Nápoles, O (2017) Expert judgement for dependence in probabilistic modelling: a systematic literature review and future research directions. Eur J Oper Res 258(3):801–819

Wilson KJ (2017) An investigation of dependence in expert judgement studies with multiple experts. Int J Forecast 33(1):325–336

Winkler RL, Clemen RT (2004) Multiple experts vs. multiple methods: combining correlation assessments. Decis Anal 1(3):167–176

Wright G, Goodwin P (2009) Decision making and planning under low levels of predictability: Enhancing the scenario method. Int J Forecast 25(4):813–825

Wunderlich K, Symmonds M, Bossaerts P, Dolan RJ (2011) Hedging your bets by learning reward correlations in the human brain. Neuron 71(6):1141–1152

Zwirglmaier K, Straub D (2016) Approaches to Bayesian network structure elicitation. In: Walls L, Revie M, Bedford T (eds) Risk, reliability and safety: innovating theory and practice. Proceedings of the 26th European safety and reliability conference, ESREL 2016, Glasgow, September 2016. Taylor and Francis Group, London, p 333

Chapter 9
Combining Judgements from Correlated Experts

Kevin J. Wilson and Malcolm Farrow

Abstract When combining the judgements of experts, there are potential correlations between the judgements. This could be as a result of individual experts being subject to the same biases consistently, different experts being subject to the same biases or experts sharing backgrounds and experience. In this chapter we consider the implications of these correlations for both mathematical and behavioural approaches to expert judgement aggregation. We introduce the ideas of mathematical and behavioural aggregation and identify the possible dependencies which may exist in expert judgement elicitation. We describe a number of mathematical methods for expert judgement aggregation, which fall into two broad categories; opinion pooling and Bayesian methods. We qualitatively evaluate which of these methods can incorporate correlations between experts. We also consider behavioural approaches to expert judgement aggregation and the potential effects of correlated experts in this context. We discuss the results of an investigation which evaluated the correlation present in 45 expert judgement studies and the effect of correlations on the resulting aggregated judgements from a subset of the mathematical methods. We see that, in general, Bayesian methods which incorporate correlations outperform mathematical methods which do not.

9.1 Introduction

In this chapter, we consider problems for which we consult multiple experts and elicit their judgements on quantities of interest. Earlier chapters of this book have considered different approaches to treating the judgements of the experts, in most cases attempting to coerce the judgements of the different experts into a single probability distribution on each unknown representing the uncertainty about its true value. We will characterise these approaches as either "mathematical aggregation" or "behavioural aggregation" in the next section. We call the probability statements resulting from an aggregation method the aggregated judgements.

K.J. Wilson (✉) • M. Farrow
Newcastle University, Newcastle upon Tyne, UK
e-mail: kevin.wilson@ncl.ac.uk; malcolm.farrow@ncl.ac.uk

L.C. Dias et al. (eds.), *Elicitation*, International Series in Operations Research & Management Science 261, https://doi.org/10.1007/978-3-319-65052-4_9

We could think of the judgements on a single unknown quantity from a group of experts as a sample of data on that quantity. A typical assumption made when fitting models to data is that the data are independent. This is also a common assumption made in methods to aggregate expert judgements. However, there are good reasons to believe that the judgements of different experts on a single quantity of interest, or the judgements of a single expert on multiple quantities, are not independent. One possible reason for this is that experts in a certain field may all come from similar backgrounds They may all have been educated on similar courses at similar universities or colleges. They may have worked together in the same organisation. They may have worked on the same part of a larger process.

Another possible reason for correlations between the judgements given by experts is that they may share common biases associated with the heuristic processes people often use to assess probabilistic information on unknowns. If a single expert is subject to the same biases consistently then the errors in their assessments of unknown quantities are likely to be correlated. Similarly, if two experts are subject to the same biases, then the errors in their assessments of a single unknown are also likely to be correlated. We consider this in the context of some common heuristics;

1. judgement by availability
2. judgement by representativeness
3. anchoring

Judgement by availability concerns the ability of an expert to call to mind examples of an event occurring. Suppose we are interested in the number of deaths per year as a result of a number of different causes. There are certain newsworthy events, such as natural disasters or plane crashes, which are easy to call to mind but occur relatively rarely. The number of deaths by such causes tend to be overestimated in peoples' judgement. Conversely, there are other deaths, for example as a result of common diseases such as cancer, which typically go unreported in the media and so examples of deaths by these causes are, relatively, more difficult to recall. The result is that they tend to be underestimated. Two experts who both use judgement by availability to assess unknowns are likely to overestimate and underestimate the same unknown quantities.

In judgement by representativeness, experts assess the likelihood of events by considering similarity to the description of that event, crucially ignoring base rates. For example, when asked to estimate the probabilities of death by a paragliding accident, cancer and in a house fire of a man who is described as "an adrenaline junkie, who enjoys travelling, outdoor sports and looks after his body through a healthy diet and no alcohol or fast food", somebody using judgement by representativeness may give the highest probability to a paragliding accident, ignoring the very small number of people who die each year as a result of paragliding accidents and the much larger number who die of cancer. Again, there are likely to be correlations between the errors in the judgements of two experts using judgement by representativeness.

Anchoring is an effect resulting from the order in which an expert considers numerical values for unknowns. It appears as a reluctance to move far enough

from a previously considered value, at which the expert's judgement seems to be "anchored". This could be not adjusting a prior judgement sufficiently when new information is received or it could be staying close to a value considered in a previous question. For example, if an expert were to be asked whether the number of deaths per year in the UK from road accidents were smaller or larger than 10,000 the expert might judge it to be smaller. However, when the expert is then asked for a judgement of the number of deaths from road accidents, the expert would take 10,000 as the starting point and there is a tendency then to provide an estimate too close to 10,000. If multiple experts are asked the same questions in the same order, there is a possibility of correlations between the errors in their answers induced by anchoring.

There are other issues which could lead to biases in expert judgements. Examples include pruning bias, which implies that model complexity has an influence on the judgements given by experts, partition bias, in which peoples' judgements of the likelihood of events change based on how options to a question are set out, and framing bias. Additionally, in behavioural approaches to aggregation, there is the possibility of social pressure, for example "group think", to bias the experts. While elicitation methods are typically set up to minimise biases, it would be naive to assume that the resulting judgements are always free from bias. Much work has been undertaken to consider heuristics and biases in expert judgement. For more information see Chap. 15 of this book (Montibeller and von Winterfeldt 2018): biases and pitfalls, Section 3.4 of O'Hagan et al. (2006) and Slovic (1972), Kahneman and Tversky (1971), Bar-Hillel and Neter (1993), Garthwaite et al. (2005).

In the next section we describe in general mathematical and behavioural methods for the aggregation of expert judgement, making specific reference to earlier chapters in this book. In Sect. 9.3 we identify all of the possible sources of correlation in expert judgement studies and focus our attention on those representing correlated experts. In Sect. 9.4 we describe some commonly used approaches to mathematical aggregation, and discuss their ability to capture correlations between experts. In Sect. 9.5 we return to the idea of behavioural aggregation methods, and consider them in the light of correlations between experts. Section 9.6 is concerned with an evaluation of a subset of the mathematical aggregation methods described in Sect. 9.4, specifically focussing on the effect of correlations between experts on the aggregated judgements resulting from the aggregation methods. Section 9.7 concludes the chapter with a summary and a consideration of future directions for the field.

9.2 Mathematical and Behavioural Aggregation

There are two main approaches to aggregating the judgements of multiple experts into a single probability distribution for each unknown; mathematical aggregation and behavioural aggregation. In mathematical aggregation methods, a subjective

probability distribution is elicited from each member of a group of experts and these distributions are then combined by a procedure outside the group. In behavioural methods the combination of judgements takes place within the group. The experts themselves come to a single consensus distribution by means of discussion and interaction.

Behavioural methods may be structured. That is the experts go through a prescribed sequence of stages, usually under the guidance of a facilitator, to help them to reach a consensus. The SHELF protocol (Oakley and O'Hagan 2016) is a structured method.

In a mixed approach to aggregation, behavioural and mathematical aggregation are combined. While the experts do interact, a final mathematical aggregation stage may be applied so that the experts are not forced to reach agreement. The well-known Delphi procedure and the recently developed IDEA protocol, (Hanea et al. 2016a,b) are mixed aggregation procedures.

In mathematical aggregation, a mathematical rule is used to aggregate the judgements. Two classes of techniques have been advocated for doing this. The first is opinion pooling, in which the aggregated probability distribution is a weighted average of the individual distributions of the experts. The average can be arithmetic or geometric. Opinion pooling aims to give weights to individual experts. The aggregated distribution for an unknown quantity θ is then the weighted average of all of the experts' judgements for that quantity.

Suppose that expert i, for $i = 1, \ldots, E$, gives elicited values which result in the distribution $f_i(\theta)$. Typically quantiles are elicited from each expert, a convenient parametric probability distribution is used and the parameters of this distribution are chosen to match the quantiles elicited from the expert. A simple example of parameter estimation based on elicited quantiles is given in Appendix 1 to illustrate this approach. Further suppose that the weight attached to expert i is w_i, $0 \leq w_i \leq 1$, where $\sum_{i=1}^{E} w_i = 1$. Define $\underline{D} = (f_1(\theta), \ldots, f_E(\theta))$ to be the set of expert distributions for θ. The aggregated distribution will take one of two forms, a linear pool

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

or a logarithmic pool,

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

where k is a normalising constant to ensure that the distribution integrates to 1. Note that, in the logarithmic pool, if any expert gives θ a probability of 0 then it will have a probability of 0 in the aggregated distribution. In Chap. 2 of this book (Quigley et al. 2018), we saw the most commonly used opinion pooling method, the Classical Method.

In the Classical Method, weights for the experts were chosen based on their performance on questions to which the answers were known to the facilitator, but not the expert. Other methods have investigated alternative performance weighting schemes. We will investigate three such methods in Sect. 9.4.2. Alternative methods to weighting the experts are to use equal weights, to allow the analyst or decision maker to choose the weights or to weight the experts on their views of their own or each other's expertise. Each approach has strengths and weaknesses. A recent review of different approaches to weighting experts is given in Bolger and Rowe (2015) and the accompanying discussions.

The alternative to opinion pooling is Bayesian aggregation. This uses the continuous version of Bayes Theorem to aggregate the individual distributions of the experts. It relies on the existence of a decision maker, sometimes called a "supra-Bayesian", who may be the analyst combining the judgements. Suppose we are interested in an event or unknown quantity which we shall call θ. Then the experts will give us, through an elicited probability or quantiles of θ, individual prior distributions, $f_{0,i}(\theta)$, for experts $i = 1, \ldots, E$. The set of these elicited distributions is $\underline{D} = (f_{0,1}(\theta), \ldots, f_{0,E}(\theta))$. A Bayesian aggregation method then works by applying Bayes Theorem,

$$f_{1,DM}(\theta \mid \underline{D}) \propto f_{0,DM}(\theta) L_{DM}(\underline{D} \mid \theta), \tag{9.1}$$

where $f_{0,DM}(\theta)$ represents the decision maker's prior probability distribution for unknown θ, $L_{DM}(\underline{D} \mid \theta)$ is the decision maker's likelihood of observing $\underline{D}$ given θ and $f_{1,DM}(\theta \mid \underline{D})$ is the decision maker's posterior distribution for θ.

The main challenge in this method is eliciting from the decision maker the likelihood function $L_{DM}(\underline{D} \mid \theta)$. It is in this likelihood function that the correlations in the expert judgement study can be captured. We see that the outcome of Bayesian aggregation methods is a subjective probability distribution which gives the updated beliefs of the decision maker in the tradition of a subjectivist Bayesian analysis. In this sense, the Bayesian method has an advantage over opinion pooling methods, in which the aggregated distribution does not represent the views of any single person. However, in opinion pooling the contribution of each of the experts to the final aggregated distribution is more transparent than in the Bayesian approach, through the expert weights, w_i.

The methods explored in Chap. 6 of this book (Hartley and French 2018) took the Bayesian approach to expert judgement aggregation. We consider four Bayesian aggregation methods in Sect. 9.4.1.

9.3 Sources of Correlation

There are several different areas within group expert judgement studies in which there are potential correlations. They were identified in Wilson (2016) and we will

review them here. Such correlations have the potential to affect the accuracy of the aggregated distribution resulting from expert judgement studies.

In order to identify the potential correlations present in expert judgement studies, it will be useful to consider two types of uncertainty which are relevant to experts making judgements. The first is *aleatory uncertainty*, which represents randomness in the state of the world. For example, if we were to roll a die, we are uncertain as to how many spots will end face up. It doesn't matter how many times we have rolled the die in the past, this will always be an uncertain event. The second type we shall consider is *epistemic uncertainty*, which represents our own lack of knowledge. For example, if someone were to hand us a loaded die then we would have additional uncertainty around how likely we are to see a six, for example. We could reduce our uncertainty about this event by rolling the die a large number of times and counting the number of sixes.

The possible correlations within an expert judgement study are:

1. Correlation between the experts for individual quantities: these could be a result of the similar past experience and common knowledge of the experts or because the experts are susceptible to the same biases through their use of heuristics.
2. Correlation within individual experts' assessments of different quantities: these could be as a result of a consistent susceptibility of an expert to the same biases.
3. Correlation between the experts for different quantities: these could be as a result of multiple experts being consistently susceptible to the same biases.
4. Underlying aleatory correlation between the values of the quantities to be assessed in the expert judgement study: plotting one against another there is a relationship.

The first three types of correlation are conditional on the true value of the underlying variable. In the Bayesian mathematical aggregation methods there are a further two possible correlation types. They are:

5. Underlying epistemic correlation between the quantities in the study: learning about one quantity will inform us as to the likely value of another.
6. Correlation between the experts' judgements and the decision maker's judgements (French 2011). These could again come about as a result of common knowledge or susceptibility to the same biases.

In this chapter, our focus is on combining judgements from correlated experts. This could be multiple experts whose judgements are correlated with each others', or a single expert whose judgements over several unknowns are correlated. Thus our focus is on correlations of types 1–3 from the list above.

The correlations between experts for individual quantities and within individual experts for multiple quantities can be assessed empirically for a given study using seed variables, questions which are related to the current problem but for which the answers are known to the analyst, as long as experts are being asked to give values for multiple quantities. This can then be built into aggregation methods.

The correlations between the judgements of the experts and the judgements of the decision maker can in theory also be estimated empirically if the decision maker

also gives their judgements for each of the unknown quantities in the study. This is rare in expert judgement studies.

Simply because these correlations exists within a study it does not necessarily mean that they are having an influence on the accuracy of the outputs of an aggregation method, however. There is of course a difference between statistical importance and practical importance.

9.4 Mathematical Aggregation Methods

In this section, we review some mathematical aggregation methods, both opinion pooling and Bayesian, from the literature. The methods have been chosen to provide a range of techniques, to assess the ability of the methods to incorporate correlated experts. We do so qualitatively in Sect. 9.4.3 and quantitatively in Sect. 9.6.

9.4.1 Bayesian Methods

Multivariate Normal model

Winkler (1981) proposed the use of a Bayesian model for unknown θ in which the variable of interest was the error of expert i, $u_i = \mu_i - \theta$, where μ_i is the mean of expert i's prior distribution $f_{0,i}(\theta)$. The posterior distribution then took the form

$$f_{1,DM}(\theta \mid \underline{D}) \propto f_{0,DM}(\theta) f_{0,E}(\mu_1 - \theta, \ldots, \mu_k - \theta),$$

and so $L_{DM}(\underline{D} \mid \theta) = f_{0,E}(\mu_1 - \theta, \ldots, \mu_k - \theta)$, where the likelihood incorporates the dependence between the errors in estimation from the experts. Winkler suggested using a flat prior for the decision maker and using data to calibrate individual expert densities. He proposed the use of the multivariate Normal distribution for $f_{0,E}(\cdot)$ of the form $N(\underline{0}, \Sigma)$, where

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \cdots & \cdots & \sigma_{1,k} \\ \vdots & \sigma_2^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ \sigma_{k,1} & \cdots & \cdots & \sigma_k^2 \end{pmatrix} \tag{9.2}$$

so that expert marginal distributions are univariate Normal with mean μ_i and variance σ_i^2. He proposed taking the experts' specifications for the marginal distributions and estimating the correlations either from data or if none were available from the judgements of the decision maker. An inverse Wishart distribution was suggested as a prior for Σ.

Copula Model

Jouini and Clemen (1996) proposed a Bayesian aggregation model which was based on the idea of copulas: joint probability distributions incorporating dependence on the unit square with given marginal distributions.

Their approach assumed a non-informative prior for the decision maker, $f_{0,DM}(\theta)$, and is sufficiently flexible to consider the bias in experts and the correlations between experts. Both were assessed by the decision maker. If the decision maker did have some substantive prior knowledge, they could be included as an extra expert in the likelihood. The model used Kendall's Tau as its measure of dependence between experts due to its exposition in terms of concordance.

In particular the authors advocated use of Archimedian copulas. The decision maker's posterior distribution becomes

$$f_{1,DM}(\theta \mid \underline{D}) \propto f_{0,1}(\theta) \times \ldots \times f_{0,E}(\theta) \times c_\tau(F_{0,1}(\theta), \ldots, F_{0,E}(\theta)),$$

where, for $u_i = F_{0,i}(\theta)$, the Archimedian copula takes the form

$$c_\tau(u_1, \ldots, u_E) = \psi(\phi(u_1) + \cdots + \phi(u_E)),$$

where $\psi(\cdot)$ is a completely monotonic function and $\phi(\cdot)$ a function which satisfies $\psi(u) = \phi^{-1}(u),\ u \in [0, \phi(0)]$ and is zero elsewhere. If the experts are assumed to be exchangeable a priori then the dependence reduces to a single Kendall's Tau between experts.

Empirical Bayes Model

Ganguly (2017) worked in the context of considering multiple assessments by multiple experts. She proposed a parametric model which assumed that the difference between the true values of the unknowns and the estimates from the experts, i.e. the errors u_i, followed a multivariate Normal distribution for the vector of expert assessments. The standard deviation for each was inferred from a self-assessment by each expert and free to change question by question.

The correlations between the experts were assumed constant for all questions. The correlations were estimated pairwise relative to the difference between an expert's estimates and the group averages for the unknowns as their true values were not known. An empirical prior was assumed across all question and an empirical Bayesian solution was implemented.

9.4.2 Opinion Pooling Methods

Cooke's Classical Method

Perhaps the most widely used opinion pooling method is Cooke's Classical Method (Cooke 1991). This is a linear pooling method which gives performance-based weights to experts. These weights are based on the performance of the experts on seed questions, for which the answers are known to the analyst, on two criteria; calibration and information.

Each expert is asked for quantiles, usually 5%, 50% and 95%, for each seed variable. Then, if an expert were well calibrated, the proportion of true values to fall into the m bins created by these quantiles would be $p = (p_1, \ldots, p_m)$, where usually $m = 4$ and $p = (0.05, 0.45, 0.45, 0.05)$. Let the actual proportions be $(s_1(e), \ldots, s_m(e))$ for expert e. Let $D(u \mid v)$ denote the Kullback-Leibler divergence of u from v, where u and v are probability distributions, and let

$$r = 2ND(s(e) \mid p) = 2N \sum_{i=1}^{m} s_i(e) \log \left(\frac{s_i(e)}{p_i} \right)$$

where N is the number of seed questions. Under the hypothesis H_e, that the interval containing the true value for each variable is drawn independently from p, r is a realisation of a random variable which has approximately a chi-squared distribution on $m - 1$ degrees of freedom. Then the expert's calibration score is

$$C(e) = \Pr(X \geq r \mid H_e),$$

where X has a chi-squared $(m - 1)$ distribution. The expert's information score is given by

$$I(e) = \frac{1}{N} \sum_{i=1}^{N} D(f_{e,i} \mid g_i),$$

where $f_{e,i}$ is the expert's density for seed quantity i and g_i is a background distribution which is typically chosen to be uniform or log-uniform. The weights are then calculated to maximise the joint calibration and information score of a global or item based decision maker. The weight given to expert e is proportional to $C(e)I(e)$. The Classical method is an asymptotically proper scoring rule. A full explanation of the Classical method is given in Chap. 2 of this book (Quigley et al. 2018).

The Moment Method

The moment method (Wisse et al. 2008) is also a linear opinion pool which uses performance based weighting as a result of questions on seed variables. Instead of probability judgements, however, its inputs are moments. In the case where 5%, 50% and 95% quantiles are elicited they can be approximately converted to a mean and variance using the Pearson-Tukey method (Smith 1993).

The method assigns weights on the basis of a penalty function. When experts assess means and variances then this penalty function takes the form

$$\phi(\underline{a}, \underline{b}) = \sum_{i=1}^{N} c_1(x_i - a_i)^2 + \sum_{i=1}^{N} c_2(x_i^2 - b_i)^2,$$

where the realised values of the variables are $x_1, \ldots, x_N$, the expert's assessment for the first two moments are $\underline{a} = (a_1, \ldots, a_N)$ and $\underline{b} = (b_1, \ldots, b_N)$ and c_1, c_2 are constants which dictate the relative importance of the two moment assessments.

To give the first two moment assessments equal total penalty (Wisse et al. 2008) set $c_1 = 1$ and c_2 such that $r = 0.5$, where

$$r = \frac{\sum_{i=1}^{N} c_2(x_i^2 - b_j)^2}{\phi(\underline{a}, \underline{b})}.$$

Both the Classical Method and the moment method make use of a threshold, α. If the weight of any expert falls below this threshold, then they are excluded from the aggregated distribution.

Babuscia and Cheung Approach

Babuscia and Cheung (2014) proposed a four part method to linear pooling aggregation incorporating probabilistic thinking, calibration, elicitation and aggregation. In the calibration step the calibration score for expert e is

$$S(e) = 100\left\{1 - \frac{1}{N}\sum_{i=1}^{N}\left(\frac{e_i - r_i}{r_i}\right)^2\right\},$$

where e_i is the expert's assessment and r_i is the true value of the seed variable for question i. The value for e_i is limited to $2r_i$. This score is then combined with the score from the probabilistic thinking part of the process and the weightings for the experts are found by intersection: the highest scoring expert receives weight 0.5, the second highest receives weight 0.25, etc. The two lowest scoring experts receive the same weight.

Non-parametric Approach

Ganguly (2017) proposed a non-parametric approach to assess the optimal weights in a linear opinion pool. In this approach the same covariance matrix was assumed as in the Empirical Bayes Model. In this case, however, rather than making an assumption of multivariate Normality, the correlations between experts were estimated using a squared loss function.

9.4.3 Correlations in Mathematical Approaches

In this section we consider the seven mathematical aggregation approaches detailed in Sects. 9.4.1 and 9.4.2 in reference to the three types of correlation between experts we identified as relevant in Sect. 9.3: correlation between experts for individual quantities, correlation within individual experts' assessments of different quantities and correlation between the experts for different quantities.

In Table 9.1 we provide details about the relevance of each of these types of correlation to each of the methods.

We see that the main thing missing from those methods considered here is the ability to incorporate correlations between errors in the judgements of the experts over different quantities of interest. In a Bayesian context, this could represent either an extension of the multivariate Normal model or the copula model to consider multiple unknowns $\theta_1, \ldots, \theta_k$.

9.5 Correlations in Behavioural Approaches

An argument which is used in favour of behavioural approaches is that the interaction between the experts themselves is beneficial. That is, by seeing or hearing each other's opinions and judgements, the experts are able to combine different knowledge and experience to produce an improved collective judgement.

In this section we consider specifically how behavioural approaches relate to the possible correlations listed in Sect. 9.3.

As a starting point, let us consider a Bayesian approach and consider the problem from the point of view of a "supra-Bayesian" decision maker. In a "mathematical" Bayesian approach, the decision maker applies Bayesian inference to the judgements of the individual experts using (9.1). This requires the decision maker to assess her likelihood function $L_{DM}(\underline{D} \mid \theta)$. For example, one possibility is to consider the distribution of the experts' prior means, given the true value of an unknown quantity, to be multivariate normal. In the case of a single unknown quantity, the covariance matrix of this multivariate normal distribution would take the form (9.2). This can, of course, be extended to the case of several unknowns in which case the covariances within Σ could reflect between-question correlations

Table 9.1 Descriptions of the relevance of each type of correlation to each of the mathematical aggregation methods

Method	Between experts correlations	Between questions correlations	Betweens experts and questions correlations
Multivariate Normal	Dependence is expressed in Σ through correlations. They can be assumed known, and estimated by the decision maker using past observations, or unknown and given an inverse-Wishart prior distribution.	Individual expert distributions can be calibrated based on past observations through their means and variances (μ_i, σ_i^2).	Both specification of Σ and calibration of (μ_i, σ_i^2) can be done for for any particular unknown θ. However, the model only considers a single unknown of interest, although it could be extended to include more than one.
Copula	Dependence is specified using the parameters of the multivariate Archimedian copula, through Kendall's Tau. This specification can be made completely independently of the expert marginal distributions.	Expert distributions could be calibrated in a similar manner to the multivariate Normal method, though this has not been explicitly carried out using the copula method.	This model also considers the aggregation of judgements for a single unknown θ.
Empirical Bayes	Dependence is expressed in Σ, based on empirically calculated covariances from previous questions or seed variables. Constant covariances are assumed across questions for pairs of experts.	Dependence is expressed in Σ, based on empirically calculated covariances from previous questions or seed variables. Constant covariances are assumed across experts for pairs of questions.	Not possible in this approach.
Classical	Independence of judgements is assumed between experts	For each expert, performance on seed variables, in terms of calibration and information, determines their weight in the linear pool. The accuracy of the method relies on similar performance from the experts on seed variables and quantities of interest.	Independence of judgements is assumed.

Moment	Independence of judgements is assumed between experts	For each expert, performance on seed variables, in terms of the penalty induced by the first two moments, determines their weight in the linear pool. The accuracy of the method relies on similar performance from the experts on seed variables and quantities of interest.	Independence of judgements is assumed.
Babuscia and Cheung	Independence between experts is assumed for the calibration score. The probabilistic thinking score assesses expert biases, but does not consider the correlations between the errors in the judgements resulting from these biases.	The weights are based on the calibration score and the probabilistic thinking score. If the aggregated distribution is to be accurate, both the experts' calibration and quality (probabilistic thinking score) need to be consistent between seed variables and quantities of interest.	This is not explicitly considered in this approach.
Nonparametric	Empirical covariances are calculated, assuming independence between questions, between pairs of experts based on observations or seed variables. Weights are proportional to the inverse of the covariance matrix.	Empirical covariances are calculated, assuming independence between experts, between pairs of questions based on observations or seed variables. Weights are proportional to the inverse of the covariance matrix.	Not possible in this approach.

within an expert and between experts as well as the between-expert correlations for a single unknown suggested by (9.2). The weights given to the judgements of the experts, in the decision maker's revised judgement $f_{1,DM}(\theta \mid \underline{D})$, will depend on the elements of Σ so that, for example, the weight given to two experts who are judged to be strongly positively correlated will be less than twice the weight which would be given to just one of this pair.

Suppose that the decision maker's initial view of Σ is that she might revise her assessment of Σ in the light of new information about the experts. Thus, in a mathematical approach, seed questions might be used and the assessment of Σ revised after seeing the results. We can represent this by saying that Σ depends on some unknown parameters ψ and data can be used to update the decision maker's beliefs about ψ. For example Winkler (1981) suggested using an inverse Wishart prior for Σ. However the decision maker may have a vague prior for Σ, the information available from the use of seed questions may be fairly limited in terms of learning about Σ, a model as outlined here may not be a good representation of the real situation and the decision maker may feel that a more effective means of producing an aggregated judgement, with appropriate weights given to the experts, would be to use the experts' own knowledge and insights about their own and each others' judgements to determine the aggregation. Behavioural methods might be supposed to work in this way. So we now need to consider the extent to which behavioural methods are likely to deal appropriately with correlations. These considerations apply even if there is no identifiable supra-Bayesian decision maker.

Typically a behavioural group elicitation process will involve a carefully designed structure and one or more facilitators. Both of these have roles to play in helping the process to be successful. See, for example, Reagan-Cirincione (1994). A recent example of a behavioural group elicitation process is given by Gosling et al. (2012). Although there is a considerable literature on behavioural approaches to group elicitation and efforts to reduce various biases and to avoid the domination of the group by a minority of participants, little is said about the problem of correlations. The participants themselves may be able to take account of correlations but this depends, among other things, on the information available to participants. Some methods, such as the Delphi method, conceal from participants the source of other opinions and so may restrict the ability of participants to use their judgements about correlations. In mixed procedures there is a combination of behavioural interaction within the group with a final mathematical aggregation by the facilitator, or someone else, for example when consensus is not reached. At this stage the facilitator or decision maker may be able to take account of correlations but it is not clear, for example, how correlations which might be measured using seed questions given to individual experts might relate to effects remaining after group interaction.

In the Delphi method (e.g. Dalkey and Helmer 1963; Linstone and Turoff 1975), typically opinions are communicated anonymously between participants. In each of a number or rounds, answers to questions and comments giving the reasoning for these answers are collected by the facilitator. The facilitator then summarises this material and passes the summary to all participants, without

identifying contributors. The idea is that, in a sequence of rounds, a consensus will be approached. A final aggregation may be used if complete consensus is not reached. The anonymity of the process may limit the ability of the participants to use their knowledge of possible associations and correlation between and within the other experts in judging how much they should adjust their own responses in the light of the responses of the other experts. On the other hand the anonymity may also reduce the "bandwagon effect" and thus reduce the effect of a between-expert correlation which might otherwise be induced by the group-interaction process itself.

The Sheffield Elicitation Framework (SHELF) (Oakley and O'Hagan 2016), detailed in Chap. 4 of this book (Gosling 2018), involves three stages. First a distribution is elicited from each expert independently. This is followed by a group discussion in a "workshop". This discussion is not anonymous and would typically take place with all of the experts together in a room. Finally the experts are asked to judge what would be the distribution of another, hypothetical, person called the Rational Impartial Observer, or "RIO". The RIO is supposed to have seen the individual judgements and heard the group discussion.

This procedure does give the participants some opportunity to use their judgement about correlations in the experts' judgements although it is difficult to assess how effectively they are able to use this opportunity. Advice is given to the facilitator on managing the group discussion phase, for example encouraging quieter participants to give their views and preventing outspoken participants from dominating the discussion. However the advice does not explicitly address the problem that the group may contain a number of experts whose similar backgrounds and experience might lead a rational observer who knows of this reasonably to suppose that their judgements might be correlated and to assess them as correlated for the purpose of her own aggregation. On the other hand, advice is given in the SHELF material on the recruitment of experts, including that the group should have "sufficient diversity of experience and opinion." The desirability of diversity is also mentioned by Wintle et al. (2012).

The IDEA ("Investigate Discuss Estimate Aggregate") protocol (Hanea et al. 2016a,b), described in Chap. 5 of this book (Hanea et al. 2018), is a Delphi-like method. Experts first give individual judgements. These are then passed anonymously to all of the other participants. The collection of initial judgements is then debated in a facilitated discussion involving all of the participants. After this, each expert then gives a separate revised judgement. Up to this point IDEA follows an "estimate–talk–estimate" pattern as described by Gustafson et al. (1973). However, finally the revised judgements are aggregated mathematically. The method for mathematical aggregation may involve the use of seed questions. The IDEA protocol is thus a mixed method.

The desirability of diversity among the participants in an IDEA group is emphasised by Hanea et al. (2016b). Such diversity may tend to avoid between-expert correlations. The anonymity of the individual judgements is justified by Hanea et al. (2016b) on the grounds of avoiding dominance and "halo" effects, a "halo" effect being where participants feel under pressure not to disagree with

an individual with a strong reputation. The halo effect could, of course, induce between-expert correlation and so avoiding it is desirable. However the anonymity also makes it more difficult for participants to make judgements about possible correlations between other participants which might inform their decisions about how far ro revise their own opinions. The revised judgements are also anonymous to help avoid participants feeling under pressure to conform to the opinions of more dominant experts. Hanea et al. (2016a) discuss dependence between experts, propose a method for measuring this and discuss some results from an experiment. According to these results, the between-expert correlation appears to change little between the judgements before and after the discussion phase. However this does not address the question of whether proper account is taken of these correlations either in revising the individual judgements or in the final aggregation.

While much has been published comparing different behavioural methods with each other and with mathematical aggregation, relatively little attention has been paid to the problem of correlations between judgements in behavioural methods. One exception is Hanea et al. (2016a). So far, the most constructive relevant advice seems to be to select a diverse collection of experts in order to try to avoid between-expert correlation.

9.6 Evaluation of Mathematical Approaches

In this section we consider empirical evaluation of the presence of dependence between experts in expert judgement elicitation, and the effect of incorporating this dependence into the modelling on the accuracy of aggregation techniques. The analysis will take two stages, initially considering the ability of aggregation techniques to predict the value of an unknown and then considering the ability of the aggregation techniques to capture the uncertainty around the unknown.

To do so, we require data from expert judgement elicitations in which multiple unknowns have been elicited from multiple experts. As part of Cooke and Goossens (2007), the data from 45 expert judgement elicitations conducted by TU Delft and analysed using the Classical Method, were released. The findings which follow are based on an analysis of the data from all 45 elicitation. For each elicitation:

- there were multiple experts (between 3 and 77),
- each was asked each of the seed questions (between 5 and 48),
- quantiles were provided for each unknown by each expert (typically 5%, 50% and 95%),
- the true value of the seed variable is known.

The name of each elicitation in Cooke and Goossens (2007), the number of seed variables and the number of experts are given in the table in Appendix 2. Full details of each of these elicitations are given in Cooke and Goossens (2007).

9.6.1 Prediction

Wilson (2016) analysed the TU Delft elicitations in terms of their ability to make predictions. Their analysis produced a number of findings; general findings on elicitation, findings on the extent of dependencies in expert judgement elicitation with multiple experts and findings on the effect of the incorporation of these dependencies on the accuracy of predictions resulting from mathematical aggregation models. We now summarise each of these groups of findings, beginning with the general findings.

F1 There is a strong positive relationship between the true values of the seed variables and the medians given by the experts.
F2 A large amount of overconfidence is observed in the experts. The true values of the seed variables fall between the 5% and 95% quantiles assessed by the experts just 52% of the time.

The findings on the extent of dependencies in expert judgement elicitation with multiple experts were:

F3 A large proportion of the elicitations contain pairs of experts whose errors in their judgements, with respect to their medians, are strongly positively correlated using both Pearson and Kendall correlation (for Pearson, 93% above 0.67 and 73% above 0.95).
F4 A small proportion of the elicitations contain pairs of experts whose errors in their judgements are strongly negatively correlated (for Pearson, 16% below −0.67 and 4% below −0.95).
F5 There are very few elicitations in which the errors on the judgements of individual experts for multiple unknown quantities are highly correlated.

Wilson (2016) compared the prediction from four mathematical approaches, two Bayesian and two opinion pooling. The Bayesian methods used were the multivariate normal approach of Winkler (1981) and the copula approach of Jouini and Clemen (1996) and the opinion pooling approaches were the Classical Method of Cooke (1991) and the method of Babuscia and Cheung (2014). As we have seen, the two Bayesian approaches incorporate correlations between experts and the two opinion pooling methods do not. Both in-sample validation and leave-one-out cross validation were used to compare methods. The numbers reported here are those from the in-sample validation. The main findings were:

F6 When compared to equal weighting of experts, all of the mathematical aggregation methods provide better prediction in over 50% of the elicitations using Mean Absolute Percentage Error (MAPE), Residual Mean Square Percentage Error (RMSPE) and Maximum Absolute Percentage Error (MAXPE).

In Fig. 9.1 we see the MAXPE for each elicitation for both equal weights and each of the four mathematical aggregation methods. In each case, we see the majority of the points lying above the line of $x = y$ indicating that the method outperforms equal weights.

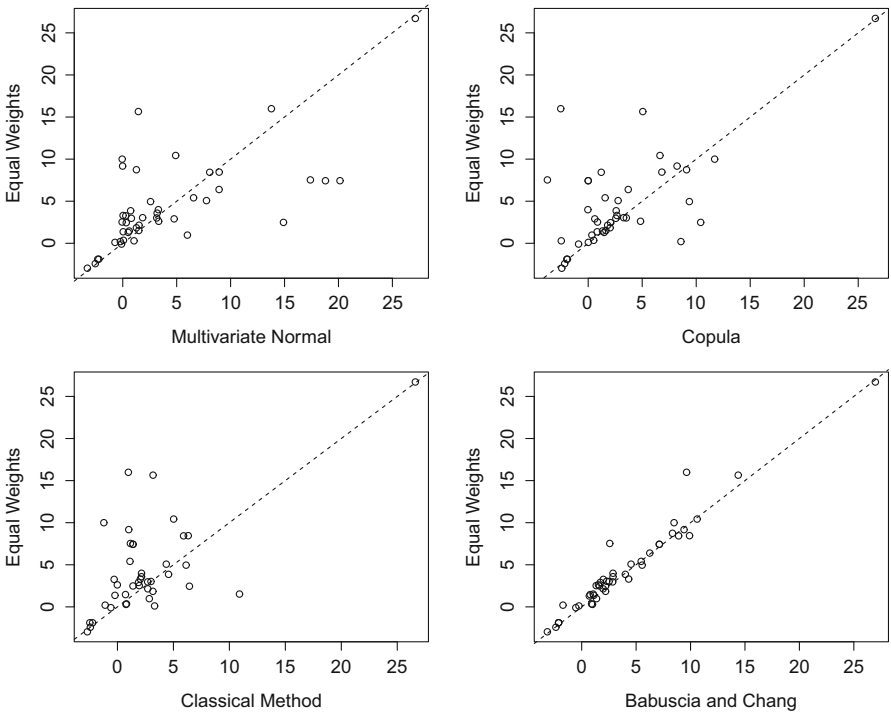


Fig. 9.1 Scatter plots comparing the MAXPE from Multivariate Normal, Copula, Classical and Babuscia and Cheung methods to equal weights

F7 The two Bayesian approaches to aggregation provide the best prediction in 60% (MAPE), 64% (RMSPE) and 75% (MAXPE) of the elicitations.

F8 If we consider only the elicitations in which at least one pair of experts is highly correlated (Kendall correlation above 0.75), all of the methods still provide better prediction that equal weighting in more than half of the elicitations.

In Fig. 9.2 we see the MAXPE for each elicitation for both equal weights and each of the four mathematical aggregation methods for only the elicitations in which there was at least one pair of highly correlated experts. We see, as in Fig. 9.1, the majority of the points lying above the line of $x = y$ indicating that the method outperforms equal weights.

F9 The two Bayesian aggregation approaches provide superior predictions more often when we consider only the elicitations with highly correlated experts than when we consider all of the elicitations. The Bayesian aggregation approaches offer the best predictions in 60% (MAPE), 68% (RMSPE) and 88% (MAXPE) of the elicitations with highly correlated experts.

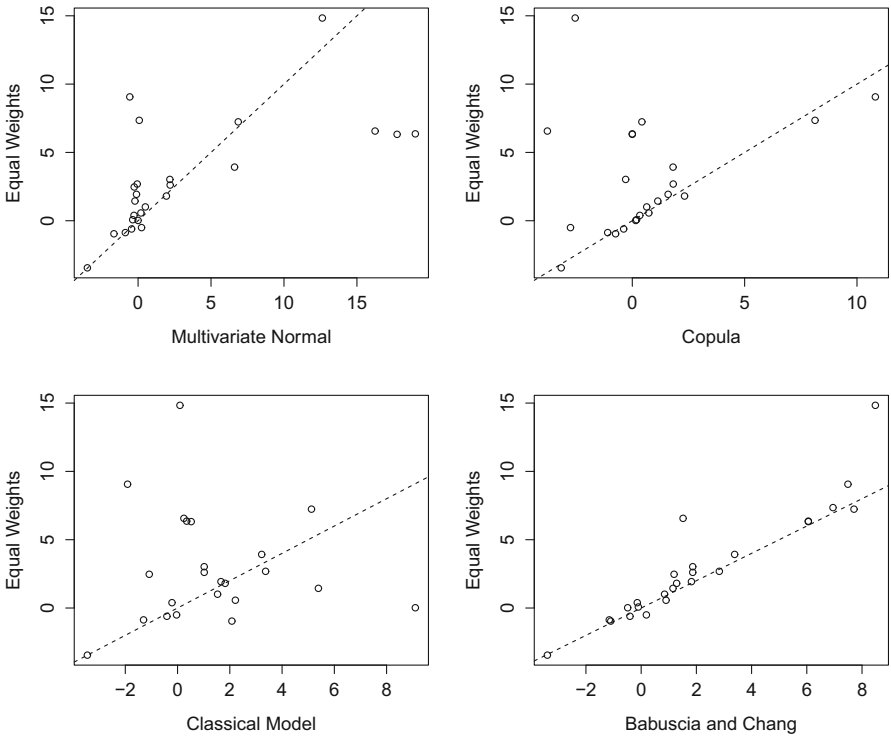


Fig. 9.2 Scatter plots comparing the MAXPE from Multivariate Normal, Copula, Classical and Babuscia and Change methods to equal weights for the elicitations with highly correlated experts

9.6.2 Uncertainty

To assess the uncertainty of the aggregated distributions we consider the best performing Bayesian method and the best performing opinion pooling method from above: the multivariate Normal approach of Winkler (1981) and the Classical Method of Cooke (1991). Correlations between experts were estimated empirically using the seed questions for the Bayesian aggregation method. From the 45 TU Delft studies, we investigate three: the flange leak study (number 1), which has 8 seed variables and 10 experts, the space debris study (number 4) which has 18 seed variables and 7 experts and the return1 study which has 15 seed variables and 5 experts. The highest Kendall correlations between experts in the three studies are 0.84, 0.98 and 0.38 respectively and so we are considering one study (space debris) with at least one pair of very highly correlated experts, one study with at least one pair of reasonably strongly correlated experts (flange leak) and one study with only moderately correlated experts (return1).

In each case we assess the 5%, 50% and 95% quantiles of the aggregated probability distribution for each seed variable. If the aggregated expert is well-

calibrated, then the true value of the seed variable should fall between the 5% and 95% quantiles with a probability of approximately 0.9. The three quantiles for the aggregated distributions using the two approaches and the realisation of the seed variable for all of the seed variables in the return1 study are given if Fig. 9.3.

We see that both of the methods include the seed variable within the upper and lower 5% quantiles for all 15 of the seed variables. The biggest differences between the methods in this case is that the aggregated distribution resulting from the multivariate Normal method typically has larger uncertainty than that from the Classical method. If we consider the four probability bins made up by these four quantiles (below 5%, between 5% and 50%, between 50% and 95% and above 95%), then we might hope to see 5%, 45%, 45% and 5% of the seed variables lying in these four bins respectively. In the case of the Classical Method, these proportions are (0,0.33,0.67,0) for this study and for the multivariate Normal method they are (0,0.27,0.73,0).

We can perform the same analysis for one of the studies with at least one pair of strongly correlated experts. Let us consider the flange leak study. The equivalent plot to Fig. 9.3, showing the three quantiles for the aggregated distributions using the two approaches and the realisations of the seed variables for each of the seed variables in the flange leak study are given in Fig. 9.4.

We see a very different picture in this case. The Classical Method is still giving 5% and 95% quantiles between which the true realisation of the seed variable lies the majority of the time. However, the multivariate Normal method is now producing quantiles which display far too little uncertainty. The result is that the true value of the seed variable lies between the 5% and 95% quantiles rarely in this case. The proportion of seed variables lying in the four bins for this study using the Classical Method are $(0.125, 0.5, 0.375, 0)$ and for the multivariate Normal method they are (0.25,0.125,0.125,0.5).

A similar story presents itself when we analyse the results of the two aggregation approaches for the space debris study. The relevant plots are given in Fig. 9.7 in Appendix 3. Again, the Classical Method is giving reasonable estimates of the uncertainty on the seed variables whereas the multivariate Normal method is typically underestimating the uncertainty. The proportions of observations falling into the four bins from the two methods (0.27,0.23,0.44,0.06) and (0.44,0,0.06,0.5) respectively.

From these three studies, it would appear that highly correlated experts have the effect of reducing the uncertainty in the aggregated distribution resulting from the multivariate Normal model to the extent that the true observations often fall outside the 5% and 95% quantiles. By contrast, the Classical Method appears to estimate the uncertainty on the seed variable reasonably well for both uncorrelated and correlated experts.

In fact, we can use a simple example to show that the multivariate Normal method does indeed have particular problems assessing uncertainty when experts are highly positively correlated. Consider an elicitation with two experts. Suppose that, from elicited quantiles, expert 1's mean and variance for unknown θ are (μ_1, σ_1^2) and expert 2's mean and variance are (μ_2, σ_2^2). Further suppose that the correlation

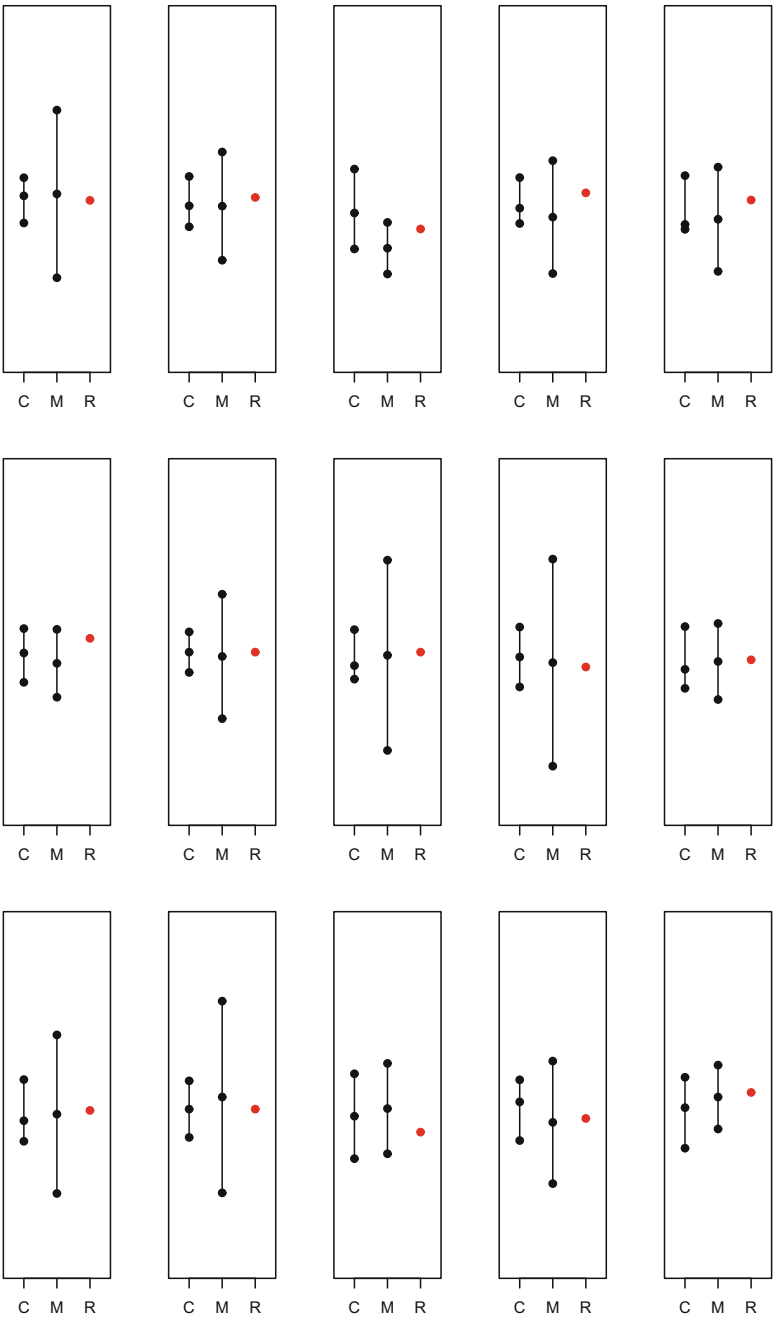


Fig. 9.3 The three quantiles for the aggregated distributions using the Multivariate Normal (M) and Classical methods (C) and the realisation of the corresponding seed variable (R) for all of the seed variables in the return1 study

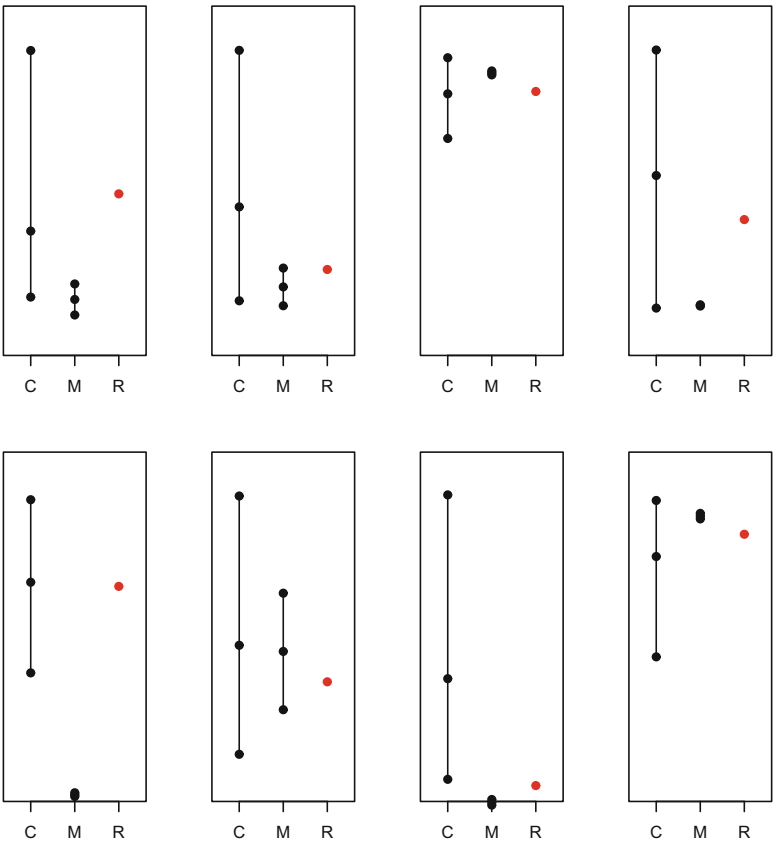


Fig. 9.4 The three quantiles for the aggregated distributions using the Multivariate Normal (M) and Classical methods (C) and the realisation of the corresponding seed variable (R) for all of the seed variables in the flange leak study

between experts 1 and 2 is ρ. Then Winkler (1981) shows that the mean and variance of the aggregated distribution are

$$\mu^* = \frac{(\sigma_2^2 - \rho\sigma_1\sigma_2)\mu_1 + (\sigma_1^2 - \rho\sigma_1\sigma_2)\mu_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2},$$

$$\sigma^{*2} = \frac{(1-\rho^2)\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}.$$

Suppose that $\sigma_1^2 = 1$. We can see the effect on the variance of the aggregated distribution for θ of varying σ_2^2 between 0 and 10 and specifying ρ to represent virtually uncorrelated and highly correlated experts. In Fig. 9.5, the black line represents $\rho = 0.05$ and the red line represents $\rho = 0.95$.

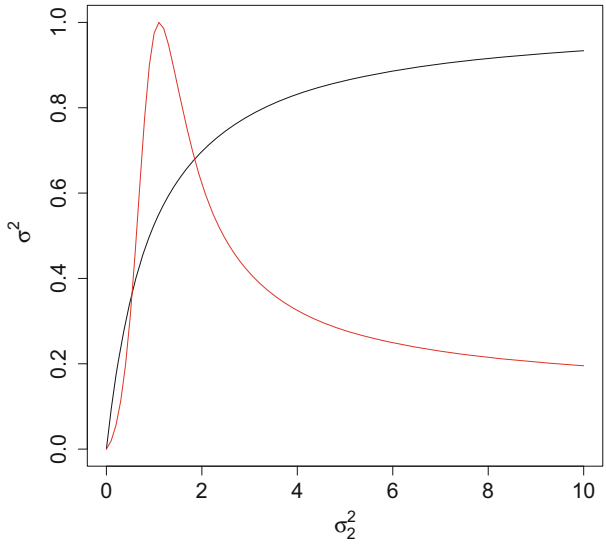


Fig. 9.5 The aggregated variance using the multivariate Normal method for varying σ_2^2 with $\rho = 0.05$ (*black*) and $\rho = 0.95$ (*red*)

We see from the plot that when $\rho = 0.05$ as the uncertainty expressed by expert 2, in the form of σ_2^2, increases, this leads to increased uncertainty, in the form of σ^2, in the final aggregated distribution for θ. However, when $\rho = 0.95$, then there is a maximum in the plot, beyond which, as expert 2's uncertainty about the true value of θ increases, the uncertainty about its value in the aggregated distribution decreases.

It can be shown that the turning point in general is $\sigma_2^2 = \sigma_1^2/\rho^2$. We see that, as the correlation between the experts, ρ, increases, the variance for which this strange behaviour will begin to happen decreases.

We can see the effect of this behaviour in terms of the aggregated quantiles resulting from the specifications of the two experts. Suppose that the means for experts 1 and 2 are $\mu_1 = 1, \mu_2 = 1$ respectively. Then, for the variances used previously and $\rho = 0.05$, the median and 5% and 95% quantiles of the aggregated distribution for θ are given in the left hand side of Fig. 9.6.

We see that, as σ_2^2 increases, the 90% uncertainty limits for θ become wider, as we would expect. The same plot, with $\rho = 0.95$, is given in the right hand side of Fig. 9.6. In this case we see the effect of the increasing aggregated variance up until the turning point at $\sigma_1^2/\rho^2 = 1.1$. Beyond this value, the uncertainty in θ in the aggregated distribution is reducing as expert 2's uncertainty increases.

Thus we see that the multivariate Normal method is not suitable for assessing the uncertainty in the aggregated distribution when there are highly correlated experts in the elicitation.

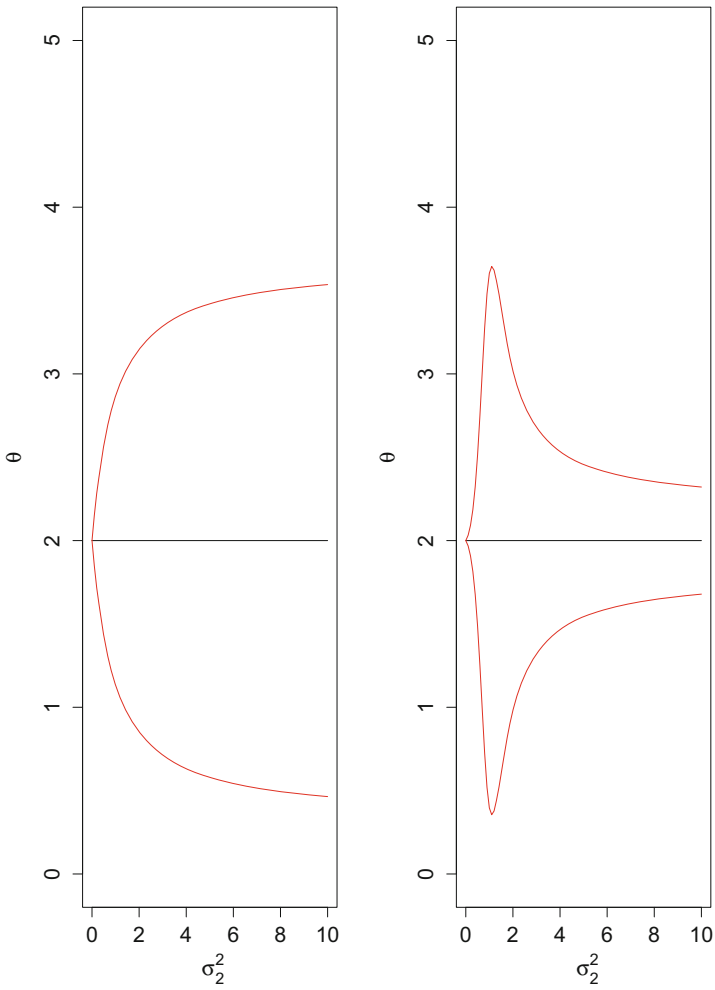


Fig. 9.6 The median (*black*), 5% and 95% quantiles (*red*) of the aggregated distribution using the multivariate Normal method for varying σ_2^2 with $\rho = 0.05$ (*left*) and $\rho = 0.95$ (*right*)

9.7 Summary and Future Directions

In this chapter we have considered the problem of combining judgements from correlated experts. We have seen that there are several sources of correlation which are relevant to any expert judgement study. In particular, correlations between experts and between experts' judgements for multiple questions could have the effect of introducing biases into aggregated judgements, whether judgements are aggregated using behavioural or mathematical methods. We made the distinction between correlations resulting from aleatory and epistemic uncertainty and it is

important to separate as much as possible these two types of uncertainty when eliciting unknowns from experts.

We considered some specific mathematical aggregation methods and identified which of the types of correlation identified were relevant to each. In general, opinion pooling approaches could not incorporate the correlations between experts, and typically assumed that different experts give independent information. In contrast, the Bayesian methods considered could all incorporate correlations between the judgements of different experts and also correlations between the judgements of individual experts for different quantities.

The issue of correlations between the judgements of experts is not usually considered in commonly used behavioural methods. However, we found that behavioural methods are typically designed to try to minimise the biases of the experts via the use of training and the order in which questions are asked. Also, the advice given in many behavioural methods to select a diverse collection of experts was identified as good practice to try to reduce the correlations between the judgements of the experts.

We saw in our empirical investigation that mathematical aggregation methods which incorporate correlations between experts typically produce slightly better point predictions for quantities of interest than those which do not, particularly for studies in which there are pairs of highly correlated experts. However, the best of these methods for point prediction which we considered, the Winkler method, produced estimates of uncertainty for the quantities of interest which were often far too tight for studies in which experts were highly correlated. In contrast, the Classical Method, which does not incorporate correlations between experts, still produced reasonable uncertainty estimates for studies in which there were highly correlated experts.

There is clearly much scope for future work in this area. In terms of mathematical methods, Bayesian methods are promising as they can explicitly account for correlations in judgements between and within experts. What is needed, is a method which can provide both good point prediction and good uncertainty estimation when there are highly correlated experts in a study. There have been some recent efforts in this direction, and an assessment of these in an analysis such as the one reported in this chapter would be a good first step towards this.

In a purely behavioural approach, attempts to deal with the issue of correlation depend on the selection of a diverse group of experts, training and facilitation. However there is scope for more experimental study to try to measure the effects. In mixed approaches, where there is a final mathematical aggregation phase after the behavioural phase, there is the possibility of using the results of seed questions to measure correlations and allowing for these correlations in the aggregation phase. This could be done, for example, by an extension of the IDEA protocol.

Expert judgements play a crucial role in scientific theory, for example in the quantification of inputs for climate models. If we do not assess the correlations present in the judgements of experts and model them appropriately, then this could have very serious impacts on the decisions being informed by the outputs from these models.

Acknowledgements The authors would like to thank Roger Cooke for discussions about the empirical study and John Quigley for helpful suggestions on an earlier version of the chapter.

Appendix 1

Suppose we have elicited from an expert the quantiles $q_1, \ldots, q_k$ corresponding to probabilities $p_1, \ldots, p_k$ for unknown θ. For example, if $p_1 = 0.5$ then q_1 would be the expert's median for θ. Now suppose that we will use an exponential distribution to represent the beliefs of the expert as expressed in the quantiles. For the exponential distribution, quantiles are given by

$$Q(p_i, \lambda) = \frac{-\log(1-p_i)}{\lambda},$$

for rate parameter λ which is estimated based on the elicited quantiles. One way to achieve this is to choose λ to minimise the sum of squared differences between the expert's judgements and the quantiles of the exponential distribution, i.e.,

$$\hat{\lambda} = \min_{\lambda \in [0,\infty)} \left\{ \sum_{i=1}^{k} (q_i - Q(p_i, \lambda))^2 \right\}.$$

We can find this value analytically by differentiating once and setting the differential equal to zero. Doing so gives

$$\sum_{i=1}^{k} \left(q_i + \frac{\log(1-p_i)}{\hat{\lambda}} \right) \left(\frac{-\log(1-p_i)}{\hat{\lambda}^2} \right) = 0,$$

and so

$$\hat{\lambda} = \frac{-\sum_{i=1}^{k} [\log(1-p_i)]^2}{\sum_{i=1}^{k} q_i \log(1-p_i)}.$$

For example, suppose that three quantiles are elicited from an expert, the lower and upper quartiles and the median. Then $p_1 = 0.25, p_2 = 0.5, p_3 = 0.75$. Suppose that the elicited values are $q_1 = 0.3, q_2 = 0.7, q_3 = 1.5$. In each case, there is an exact value of λ which satisfies this individual quantile. They are $\lambda_1 = 0.96, \lambda_2 = 0.99, \lambda_3 = 0.92$. Using the method above, we can find our estimate of λ which approximately satisfies all three quantiles. This is $\hat{\lambda} = 0.94$. Thus, we would say that this expert's distribution for unknown quantity θ is

$$\theta \sim \text{Exp}(0.94).$$

Appendix 2

Number	Study	Seed variables
1	Flange leak	8
2	Crane risk	11
3	Propulsion	13
4	Space debris	18
5	Composite materials	12
6	Option trading	38
7	Risk management	11
8	Groundwater transport	10
9	Acrylo-nitrile	10
10	Dispersion panel TUD	36
11	Dispersion panel TNO	36
12	Dry deposition	24
13	Ammonia Panel	10
14	Sulphur trioxide	10
15	Water pollution	11
16	Environm. panel	28
17	Montserrat volcano	8
18	Campylobacter NL	10
19	Campy Greece	10
20	Oper. risk	16
21	Infosec	10
22	PM25	12
23	Falls ladders	10
24	Dams	11
25	MVOseeds Monserrat follup	5
26	Pilots	10
27	Sete cidades	10
28	TeideMay 05	10
29	VesuvioPisa21Mar05	10
30	Volcrisk	10
31	Sars	10
32	A seed	8
33	Atcep	10
34	Bswaal	8
35	Dcpwwlwl	48
36	Guadeloupe	5
37	Greece NL Carma	10
38	Infoseces	10
39	Oninx	47
40	Pbearlyh	15
41	Return1	15
42	ReturnAfter	31
43	S seed	31
44	Dww exp	15
45	Exp dd	14

Appendix 3

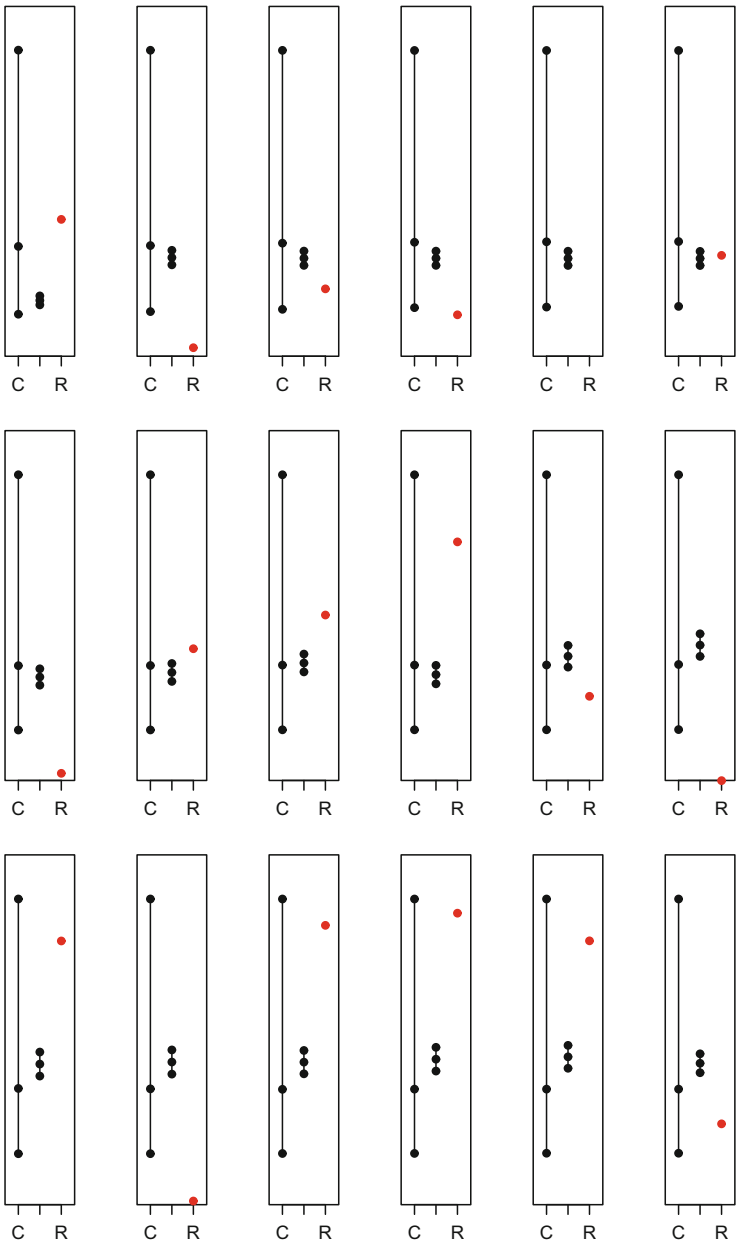


Fig. 9.7 The three quantiles for the aggregated distributions using the Multivariate Normal (M) and Classical methods (C) and the realisation of the corresponding seed variable (R) for all of the seed variables in the space debris study

References

Babuscia A, Cheung KM (2014) An approach to perform expert elicitation for engineering design risk analysis: methodology and experimental results. J R Stat Soc Ser A 177:475–497

Bar-Hillel M, Neter E (1993) How alike it is versus how likely it is: a disjunction fallacy in probability judgements. J Pers Soc Psychol 65:1119–1131

Bolger F, Rowe G (2015) The aggregation of expert judgement: do good things come to those who weight? Risk Anal 35:5–26

Cooke RM (1991) Experts in uncertainty. Oxford University Press, Oxford

Cooke RM, Goossens LHJ (2007) TU Delft expert judgement database. Reliab Eng Syst Saf 93:657–674

Dalkey N, Helmer O (1963) An experimental application of the Delphi method to the use of experts. Manag Sci 9:458–467

French S (2011) Aggregating expert judgement. Revista de la Real Academia de Ciencias Exactas 105:181–206

Ganguly T (2017) Mathematical aggregation of probabilistic expert judgements. PhD thesis, University of Strathclyde

Garthwaite PH, Kadane JB, O'Hagan A (2005) Statistical methods for eliciting probability distributions. J Am Stat Assoc 100:680–700

Gosling JP (2018) SHELF: the Sheffield elicitation framework. In: Dias LC, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgment. Springer, New York

Gosling JP, Hart A, Mouat DC, Sabirovic M, Scanlan S, Simmons A (2012) Quantifying experts' uncertainty about the future cost of exotic diseases. Risk Anal 32:881–893

Gustafson DH, Shukla RK, Delbecq A, Walster GW (1973) A comparative study of differences in subjective likelihood estimates made by individuals, interacting groups, Delphi groups, and nominal groups. Organ Beh Hum Perform 9:280–291

Hanea AM, McBride MF, Burgman MA, Wintle BC (2016a) Classical meets modern in the IDEA protocol for structured expert judgement. J Risk Res doi:10.1080/13669877.2016.1215346

Hanea AM, McBride MF, Burgman MA, Wintle BC, Fidler F, Flander L, Twardy CR, Manning B, Mascaro S (2016b) Investigate discuss estimate aggregate for structured expert judgement. Int J Forecast doi:10.1016/j.ijforecast.2016.02.008

Hanea A, Burgman M, Hemming V (2018) IDEA for uncertainty quantification. In: Dias LC, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgment. Springer, New York

Hartley D, French, S (2018) Elicitation and calibration: A Bayesian perspective. In: Dias LC, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgment. Springer, New York

Jouini MN, Clemen RT (1996) Copula models for aggregating expert opinions. Oper Res 44: 444–457

Kahneman D, Tversky A (1971) Subjective probability: a judgement of repetitiveness. Cogn Psychol 3:430–454

Linstone HA, Turoff M (eds) (1975) The Delphi method: techniques and applications. Addison-Wesley, Reading, MA

Montibeller G, von Winterfeldt D (2018) Individual and group biases in value and uncertainty judgments systems. In: Dias LC, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgment. Springer, New York

Oakley JE, O'Hagan A (2016) SHELF: the Sheffield elicitation framework (version 3.0). School of Mathematics and Statistics, University of Sheffield, UK. http://tonyohagan.co.uk/shelf

O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, Oakley JE, Rakow T (2006) Uncertain judgements: eliciting experts' probabilities. Wiley, New YorK

Quigley J, Colson A, Aspinall W, Cooke RM (2018) Elicitation in the classical method. In: Dias LC, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgment. Springer, New York

Reagan-Cirincione P (1994) Improving the accuracy of group judgment: A process intervention combining group facilitation, social judgment analysis, and information technology. Organ Beh Hum Decis Process 58:246–270

Slovic P (1972) From Shakespeare to simon: speculation - and some evidence - about man's ability to process information. Oregon Res Bull 12:1–19

Smith J (1993) Moment methods for decision analysis. Manag Sci 39:340–358

Wilson KJ (2016) An investigation of dependence in expert judgement studies with multiple experts. Int J Forecast 33:325–336

Winkler RL (1981) Combining probability distributions from dependent information sources. Manag Sci 27:479–488

Wintle B, Mascaro M, Fidler F, McBride M, Burgman M, Flander L, Saw G, Twardy C, Lyon A, Manning B The intelligence game: assessing delphi groups and structured question formats. In: Proceedings of the 5th Australian security and intelligence conference, Perth, Western Australia, Dec (2012)

Wisse B, Bedford T, Quigley J (2008) Expert judgement combination using moment methods. Reliab Eng Syst Saf 93:675–686

Chapter 10
Utility Elicitation

Jorge González-Ortega, Vesela Radovic, and David Ríos Insua

Abstract This chapter introduces key concepts in modelling preferences under uncertainty, focusing on utility elicitation, both in single and multiple attribute problems. We also discuss issues in relation with adversarial preference assessment. We illustrate all concepts with a case combining aspects of energy and homeland security.

10.1 Introduction

This chapter targets eliciting decision maker (DM) preferences under uncertainty. Complementing the many other contributions in this volume referring to modelling beliefs, it provides the other ingredient required to support decision making under uncertainty, if we opt for a Subjective Expected Utility (SEU) model. We briefly introduce basic concepts such as utility function, risk aversion and independence preference conditions. We also refer to issues concerning adversarial preferences. Our focus will be on modelling aspects and practical issues in relation with eliciting utility functions.

Before proceeding, we note the distinction between *value* and *utility* functions: the first one encodes preferences under conditions of certainty and may include a notion of strength of preference, see Chap. 12 in this volume (Morton 2017); the second one encodes both preferences and also an attitude towards risk, so that they may be used meaningfully in SEU models. The key difference is thus that values do not reflect the risks present in decisions under uncertainty. Therefore, expected values are not of use in such contexts due to preferences being typically non-linear in payoffs: decisions based on expected utilities are better suited when taking risks into account. Thus a utility function is a value function, but not vice versa. Note, though,

J. González-Ortega • D. Ríos Insua (✉)
Instituto de Ciencias Matemáticas, CSIC-UAM-UC3M-UCM, Madrid, Spain
e-mail: jorge.gonzalez@icmat.es; david.rios@icmat.es

V. Radovic
Institute for Multidisciplinary Research, Belgrade University, Belgrade, Serbia
e-mail: vesela.radovic@imsi.rs

L.C. Dias et al. (eds.), *Elicitation*, International Series in Operations Research & Management Science 261, https://doi.org/10.1007/978-3-319-65052-4_10

that it is possible to assess a value function using strength of preference comparisons and, then, transform this to a utility function, as we sketch in Sect. 10.2.

We initially refer to problems with just one criterion. We then deal with problems with multiple criteria and, finally, with adversarial problems. To support the discussion, we illustrate the main concepts with a case referring to energy security, which we briefly introduce now.

Case: Energy Security in Serbia. Energy is a fundamental input for the global community and every country's normal functioning. Thus, energy security has become a priority issue, particularly within the European Union. Recall, as an example, the recent crisis in Ukraine. The International Energy Agency (IEA) defines energy security as the "uninterrupted availability of energy sources at an affordable price". Its achievement requires "to reduce risks to energy systems, both internal and external, and build resilience in order to manage the risks that remain".

Serbian energy supply is highly influenced by international relations, consider e.g. the gas supply crisis in 2009. In practical terms, such risks can be mitigated and, consequently, energy security improved, by diminishing dependence from a single energy origin and diversifying to other energy sources. However, difficulties arise as Serbia legislated against the use of nuclear energy in 1989 and numerous obstacles, mostly in transmission systems, affect the use of renewables. Due to these factors, electricity from thermal and hydropower plants is the only one in use. A stable and secure energy infrastructure is an imperative for Serbia, being an important part of its national security. In order to follow the EU's Programme for European Critical Infrastructure Protection, Serbia has to protect its energy infrastructure against numerous threats, both anthropogenic and natural.

Internal and external energy dependence prevents the country from providing an adequate level of energy security. Thus, the Serbian government is trying to decrease its energy dependence from Russia and diversify its sources of natural gas. Western governments are trying to help it by promoting two alternative ways to import energy: the Trans-Adriatic Pipeline (TAP) and a pipeline from the Croatian gas storage facility. The EU has also devoted a specific budget for the creation of the Trans-Balkan Power Corridor interconnecting the electricity transmission systems from Serbia, Montenegro and Bosnia and Herzegovina to those of Croatia, Hungary, Romania and Italy.

Some other issues, such as ethnic and internal political divisions, create also threats to energy security in the country. For example, the Serbian-Kosovo conflict results in unsolved energy issues in the region. Experts believe that the Balkans constitute a unique and complex territory where terrorism poses a major threat to future development, being the most likely threats lone wolf attacks, related to Al-Nusra and other parts of the Islamic State, and small terrorist cells, for example in Kosovo and Sandzak. Last, but not least, there are also numerous cases in which wild fires impact the Serbian electrical transmission system, as well as criminal acts in which transformation stations and other facilities are broken and robbed.

Threats to energy infrastructure in Serbia are, therefore, numerous. Future tasks for policy makers include planning and designing a more resilient energy infras-

tructure, evaluating the work of the emergency management sector and creating Preventive Action and Emergency Plans, reducing dependency and deciding about the optimal protection level, having in mind the increasingly dynamic regional and international relations.

10.2 (Single Attribute) Utility Elicitation

We introduce key concepts in utility elicitation focusing on the single attribute case. We assume that a DM needs to make decisions under risk. This corresponds to choosing among lotteries $\mathbf{p}$, also referred to as gambles, which will be simple distributions over a consequence set $\mathcal{C}$. The set of such lotteries is designated $\mathcal{P}$. We first recall the basic mathematical structure leading to the expected utility model, then describe a protocol for utility elicitation and remind key concepts in risk aversion, temporal preferences and behavioral issues affecting preference modelling.

Under this framework, utility measurement helps the prescription of decisions under risk. By eliciting the DM's preferences for simple lotteries, which may be introspected with confidence, choice in more complex situations can be extrapolated. However, consistency in the evaluation of lotteries must be ensured. Basic utility elicitation is therefore performed by repeatedly asking the DM to assign expected utilities to different gambles, through various methods.

10.2.1 Basic Utility Concepts

In expected utility theory under risk, a preference relation $\precsim$ is assumed over $\mathcal{P}$, which satisfies three axioms due to Von Neumann and Morgestern, see French and Ríos Insua (2000):

$\mathcal{A}1$. **Weak-order:** $\precsim$ on $\mathcal{P}$ is *complete* (that is, for all $\mathbf{p}, \mathbf{q} \in \mathcal{P}$, either $\mathbf{p} \precsim \mathbf{q}$ or $\mathbf{q} \precsim \mathbf{p}$) and *transitive* (that is, for all $\mathbf{p}, \mathbf{q}, \mathbf{r} \in \mathcal{P}$, $\mathbf{p} \precsim \mathbf{q}$ and $\mathbf{q} \precsim \mathbf{r}$ imply $\mathbf{p} \precsim \mathbf{r}$).

$\mathcal{A}2$. **Archimedean:** For all $\mathbf{p}, \mathbf{q}, \mathbf{r} \in \mathcal{P}$, if $\mathbf{p} \prec \mathbf{q} \prec \mathbf{r}$, there exist $\alpha, \beta \in (0, 1)$ such that $\alpha\,\mathbf{p} + (1-\alpha)\,\mathbf{r} \prec \mathbf{q} \prec \beta\,\mathbf{p} + (1-\beta)\,\mathbf{r}$.

$\mathcal{A}3$. **Independence:** For all $\mathbf{p}, \mathbf{q}, \mathbf{r} \in \mathcal{P}$ and $\alpha \in (0, 1]$, $\alpha\,\mathbf{p} + (1-\alpha)\,\mathbf{r} \precsim \alpha\,\mathbf{q} + (1-\alpha)\,\mathbf{r}$ if and only if $\mathbf{p} \precsim \mathbf{q}$.

Under these conditions, there is a function u, called *utility function*, such that for all $\mathbf{p}, \mathbf{q} \in \mathcal{P}$:

i. $\mathbf{p} \precsim \mathbf{q} \;\Leftrightarrow\; u(\mathbf{p}) \leq u(\mathbf{q})$.
ii. $u(\alpha\,\mathbf{p} + (1-\alpha)\,\mathbf{q}) = \alpha\,u(\mathbf{p}) + (1-\alpha)\,u(\mathbf{q})$.

From this, we easily deduce that

$$\mathbf{p} \precsim \mathbf{q} \quad \Longleftrightarrow \quad E_{\mathbf{p}}[u] \leq E_{\mathbf{q}}[u],$$

where $E_{\mathbf{p}}[u]$ represents the expected utility of lottery $\mathbf{p}$.

Note that representation theorems show that utility functions are *unique up to a positive affine transformation*: two utility functions $u(\cdot)$ and $w(\cdot)$ represent the same preferences if and only if for some $\alpha > 0$ and $-\infty < \beta < \infty$,

$$w(\cdot) = \alpha\, u(\cdot) + \beta,$$

The results are extended to more general consequence and lottery sets, see French and Ríos Insua (2000).

10.2.2 An Elicitation Protocol

We suggest now how a utility function over a general consequence space $\mathcal{C}$ may be elicited. To start with, the DM is asked for her preferences between simple gambles in which the randomisations are based upon a reference experiment with outcomes drawn from $\mathcal{C}$. In the simplest case, two consequences $c_* \prec c^*$ are fixed and for each $c \in \mathcal{C}$ with $c_* \precsim c \precsim c^*$, the DM is asked to determine a value of p for which she is indifferent between:

$$\text{Gamble } \mathbf{A}: \quad c \quad \text{for certain;}$$

$$\text{Gamble } \mathbf{B}: \begin{cases} c^* & \text{with probability } p, \\ c_* & \text{with probability } 1-p. \end{cases}$$

To support the elicitation, we may consider, for example, a *probability wheel* and adjust p by varying the size of one sector of the wheel. Moreover, we may design a protocol which iteratively bounds p above and below until sufficiently specifying it. As mentioned, $u(\cdot)$ is unique up to a positive affine transformation, so we may set $u(c^*) = 1$ and $u(c_*) = 0$, without loss of generality, and deduce from her indifference:

$$u(c) = p\, u(c^*) + (1-p)\, u(c_*) = p.$$

For finite or bounded $\mathcal{C}$ it is usual to choose c^* and c_* as the most and least preferred consequences, respectively.

Once with a procedure to assign the utilities of specific consequences, we may introduce an elicitation protocol to assess the utility function, which may run as follows:

1. Determine the range of interest for the attribute.
2. Assign utility 0 to the worst value c_* and utility 1 to the best one c^*.
3. Assign utilities to a few intermediate values $c_1, \ldots, c_n$; say $u_1, \ldots, u_n$, respectively.
4. Fit a utility function to the data $((c_*, 0), (c_1, u_1), \ldots, (c_n, u_n), (c^*, 1))$, e.g. through non-linear least squares.
5. Check for consistency, by asking a few verification questions.

Note that when the utility function has been fitted, step 4, the elicitation process is not over yet. The DM's preferences should be rational, i.e. transitive and complete, so consistency must be checked, as step 5 suggests. Also, sensitivity analysis methods, see Ríos Insua (1990), may provide relevant information about the stability of the result and how critical are the analysis conclusions with respect to the elicited preferences. In particular, it might be the case that utilities are elicited imprecisely within the intervals. It is also important to take into account and counter the potential DM's inaccuracies and biases, some of them described in Sect. 10.2.4.

The key point thus is to assess the utility values of a few points, with the required amount determined by the nature of the selected utility function and the desired accuracy as well as the time available. The classic paper by Farquhar (1984) surveys a wide variety of forms of indifferences which may be sought in elicitation, being *probability equivalence* and *certainty equivalence* methods the most frequently used, which we briefly outline.

Probability equivalence methods include our above motivating example. In them, the DM is required to specify a *probability equivalent* p such that a gamble with values $x \prec y$ with respective probabilities p and $1 - p$ is equally preferred to a certain value w ($[x, y; p] \sim w$). We begin by selecting two reference points c_* and c^* in $\mathcal{C}$, where $c_* \prec c^*$. The task is to assess the utilities of the points $c_* \prec c_1 \prec \ldots \prec c_n \prec c^*$, possibly using one of the following methods:

1. *Extreme gambles*: $[c_*, c^*; p_i] \sim c_i$. The reference points in $\mathcal{C}$ are used as extremes in every gamble. If the utility of a value c not lying between c_* and c^* is needed, one can ask additional questions of the form $[c_*, c; p] \sim c^*$ for $c^* \prec c$ or $[c, c^*; p] \sim c_*$ for $c \prec c_*$. The method is easy to use, but is susceptible to serial dependence in the responses and biases from range effects, if c_* and c^* are too extreme.
2. *Adjacent gambles*: $[c_{i-1}, c_{i+1}; p_i] \sim c_i$. Instead of using extreme values as reference points, this method uses gambles over the "locally best and worst" values for each c_i. Points outside the range are easily determined by additional comparisons of the form $[c_n, c; p] \sim c^*$ for $c^* \prec c$ or $[c, c_1; p] \sim c_*$ for $c \prec c_*$. One advantage over the previous method is the attenuation of biases from range effects, though further comparisons of the form $[c_{i-k}, c_{i+k}; p] \sim c_i$ are recommended to provide consistency checks on the assessed utilities.
3. *Assorted gambles*: $[c_{j_i}, c_{k_i}; p_i] \sim c_i$, where $j_i < i < k_i$. This method generalizes the previous ones, yet requires further structure on the gamble comparisons. In any case, it is appropriate for making consistency checks.

Certainty equivalence methods ask the DM to specify a sure outcome w, called the *certainty equivalent*, for which $[x, y; p] \sim w$. Although weaker assumptions are possible, we consider a continuum of values in $\mathcal{C}$ so that w exists and strictly increasing preferences on $\mathcal{C}$ so that w is unique. We begin by fixing two reference points c_* and c^* in $\mathcal{C}$, where $c_* \prec c^*$, $u(c_*) = 0$ and $u(c^*) = 1$, and a set of probabilities $0 < p_1 < \ldots < p_n < 1$. The task is to identify the points $c_1 \prec \ldots \prec c_n$ corresponding to each of the probabilities, possibly using one of the following methods:

1. *Fractile*: $[c_*, c^*; p_i] \sim c_i$. This method is similar to the extreme gambles one, with analogous advantages and disadvantages. Simple implementation faces biases from range effects when the reference points are far apart and potential distortions in risk behaviour and other biases when probabilities are too close to 0 and 1.
2. *Chaining methods*: $[c_{i'}, c_{i''}; p_i] \sim c_i$, where $c_{i'}, c_{i''} \in S_{i-1} = S_{i-2} \cup c_{i-1}$ with $S_0 = \{c_*, c^*\}$. This method makes use of previously elicited values in subsequent gamble comparisons, obtaining *chained* responses. Well-known methods in this category are the *fractionation* or the *midpoint methods*. These allow us to assess additional values one at a time until sufficient points are available to estimate the utility function satisfactorily. Its drawbacks include serial dependence, range effects and certainty effects, among other biases.

A crucial issue is choosing the functional form of the utility function. Monotonicity is often a guide to sketching the function: e.g. more profit is invariably preferred to less. Concavity is also specially important, and we refer to it in the next section.

10.2.3 Risk Attitudes and Utility Functional Forms

Obviously, since expectation is a linear operator, linear transformations of utilities do not affect the ordering given by expected utilities. Two utility functions which represent the same preferences are said to be *strategically equivalent*. This is an equivalence relation over utility functions and those in the same equivalence class share the same risk attitude.

In supporting this claim, we limit the discussion to the case in which consequences are monetary, and $\mathcal{C}$ is an interval of the real line. Then, for any lottery $\mathbf{p}$ we may consider two expectations: the *expected utility*, $E_{\mathcal{C}}[u(c) \mid \mathbf{p}]$, and the *expected monetary value*, $E_{\mathcal{C}}[c \mid \mathbf{p}]$, which gives the average sum that the DM would receive if she could take the gamble repeatedly. In line with Sect. 10.2.2, the *certainty equivalent*, $c_{\mathbf{p}}$, of a gamble is the monetary value that the DM would place on taking it a single time:

$$u(c_{\mathbf{p}}) = E_{\mathcal{C}}[u(c) \mid \mathbf{p}],$$

or, equivalently,

$$c_{\mathbf{p}} = u^{-1}(E_{\mathcal{C}}\,[u(c) \mid \mathbf{p}]).$$

The *risk premium*, $\pi_{\mathbf{p}}$, of a gamble is the difference between its average monetary value if it is taken infinitely often and its monetary value if taken just once:

$$\pi_{\mathbf{p}} = E_{\mathcal{C}}\,[c \mid \mathbf{p}] - c_{\mathbf{p}}.$$

It provides a financial evaluation of the benefit gained by the DM by being able to "play the odds" in repetitions of the lottery.

The sign of $\pi_{\mathbf{p}}$ is determined by the shape of $u(\cdot)$. If it is concave, the risk premium of a gamble is necessarily non-negative. The DM would, therefore, value a single opportunity to take a gamble less in monetary terms than its average payoff. She is averse to the risk inherent in a single play. Thus, concave utility functions represent *risk averse* preferences. A risk-averse person prefers a small guaranteed payoff to a random payoff that has larger expected value but some chance of being very small. Concave utility is the foundation of the insurance industry. Similarly, if $u(\cdot)$ is convex, the risk premium of a gamble is necessarily non-positive. The DM values a single play of the gamble more in monetary terms than its average payoff. Convex utility functions represent *risk prone* preferences. When $u(\cdot)$ is linear, the risk premium is identically zero and the preferences are said to be *risk neutral*.

Pratt (1964) defined the *local risk aversion* of an increasing utility function as

$$r\,(c) = -\frac{u''(c)}{u'(c)} = -\frac{\mathrm{d}}{\mathrm{d}c}(\ln(u'(c))), \tag{10.1}$$

assuming that the consequences c are represented by a continuous variable and preferences increase with its value, e.g. money. If $r(\cdot)$ is everywhere non-positive, $u(\cdot)$ is convex and models risk prone preferences. Similarly, if $r(\cdot)$ is everywhere non-negative, $u(\cdot)$ is concave and models risk averse preferences.

A simple but very useful form of utility function arises when the local risk aversion is set to a constant, in which case we have Constant Absolute Risk Aversion (CARA). Integrating (10.1) gives:

$$\begin{aligned}
u(c) &= 1 - \exp(-\rho\, c), && \text{if } \rho > 0, \text{ i.e. constant positive risk aversion}\\
u(c) &= c, && \text{if } \rho = 0, \text{ i.e. positive risk neutrality}\\
u(c) &= -1 + \exp(\rho\, c), && \text{if } \rho > 0, \text{ i.e. constant positive risk proneness}
\end{aligned}$$

where strategic equivalence has been used to set arbitrary constants of integration to conventional values of ± 1. Another important case is when $u(c) = (c-\alpha)^{1-\beta}/(1-\beta)$, which corresponds to Hyperbolic Absolute Risk Aversion (HARA).

The use of exponential utility functions is specially convenient as it limits the number of utility values to be elicited, say certainty equivalents, to one of such values to assess the parameter ρ, called the *risk tolerance coefficient*, and

characterises the function. Assume, for example, that the DM is risk averse and her utility function is $u(c) = 1 - \exp(-\rho\, c)$. One way of eliciting it is to ask the DM to determine the largest stake c_{max} for which she would accept the 50-50 gamble:

$$\text{50-50 Gamble}: \begin{cases} 2c_{max} & \text{with probability } \frac{1}{2}, \\ -c_{max} & \text{with probability } \frac{1}{2}. \end{cases}$$

Then, the DM would be indifferent between 0 (the current fortune) and the 50-50 gamble, leading to the expression:

$$u(0) = \tfrac{1}{2}\, u(2c_{max}) + \tfrac{1}{2}\, u(-c_{max})$$
$$\Longleftrightarrow$$
$$1 = \tfrac{1}{2}\, (\exp(-2\rho\, c_{max}) + \exp(\rho\, c_{max})),$$

whose approximate solution is $\rho \approx \frac{1}{2c_{max}}$. Of course, consistency checks would lead us to elicit additional values.

Note that utility functions may exhibit many shapes other than concave, convex or linear. Many empirical studies have suggested that individuals' utility for money passes through regions of convexity and concavity as the sums involved increase, with risk proneness changing to risk aversion. Furthermore, an individual's utility for money and her risk attitude is undoubtedly related to her total assets. Thus, in assessing a DM's utility, it is usual to "integrate" monetary outcomes into her final level of wealth. For example, the gamble $\langle p_1, c_1; p_2, c_2; \ldots; p_r, c_r\rangle$, which offers potential changes in assets of c_i, would be framed for her as $\langle p_1, w + c_1; p_2, w + c_2; \ldots; p_r, w + c_r\rangle$, where w is her total wealth before the gamble.

Case: Utility Function. We assess now a utility function for the introduced case. In the past, Serbia has had to spend enormous financial means to mitigate the consequences of natural disasters in the energy sector due to lack of prevention. In fact, the Global Climate Risk Index 2016 (Kreft et al. 2015) considered Serbia as the country most affected by the impact of weather-related loss events in 2014. Thus, one of the key objectives considered is the minimisation of natural risks measured through the total amount of money spent on repairs of the energy infrastructure, see Fig. 10.2.

The best value for this criterion would be 0M€ if consequences of natural disasters were negligible and no extra funds were needed. On the other hand, we shall base the worst value on the effect of floods in Serbia in 2014 which accounted to almost 500M€. We use a probability equivalence method to assess the utilities of three intermediate values, as expressed in Table 10.1. Observe that the utility function is decreasing, since it refers to costs.

Table 10.1 Utilities for five values of consequence repair costs

Cost (M€)	0	100	250	400	500
Utility	1.00	0.89	0.67	0.48	0.00

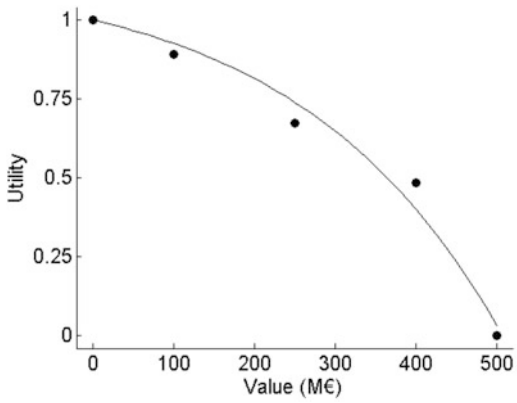


Fig. 10.1 Utility values and fitted utility function $u(\cdot)$

As an example, for a cost of 250M€, the associated utility is 0.67. This means that our expert found equally desirable a sure loss of 250M€ to a lottery which gives her a cost of 0M€ with probability 0.67 and 500M€ with probability 0.33. To come out with that value, we first offered the DM the reference lottery with probability $1/2 = 0.5$ and she responded that she preferred the certain amount of 250M€, therefore suggesting that $u(250) > 0.5$. Then, we offered her the reference lottery with probability $1/2 + 1/4 = 0.75$ and she said that she preferred the lottery, therefore suggesting that $u(250) < 0.75$ and, consequently, $0.5 < u(250) < 0.75$. Next, we offered her the reference lottery with probability $1/2 + 1/4 - 1/8 = 0.625$ and she again preferred the certain amount. We iterated the procedure until equivalence was found in the 6th iteration, rounding the value to 0.67.

Figure 10.1 shows the fitting of the utility function based on the data in Table 10.1. The data suggest fitting a curve of the form $u(c) = 1 + \lambda\,(1 - \exp(\rho\, c))$ where $\lambda, \rho \geq 0$. Through least squares we identify that the parameters are $\lambda \approx 0.15319$ and $\rho \approx 0.00398$. This corresponds to a risk averse behaviour.

We finally performed some consistency checks. As an example, we asked the DM to choose between these two lotteries:

$$\mathbf{A}: \begin{cases} 0\text{M€ with probability } \frac{1}{2}, \\ 500\text{M€ with probability } \frac{1}{2}. \end{cases} \qquad \mathbf{B}: \begin{cases} 100\text{M€ with probability } \frac{1}{2}, \\ 400\text{M€ with probability } \frac{1}{2}. \end{cases}$$

She found lottery **B** more preferred, which coincides with the ranking based on her expected utilities, since the expected utility of **A** is 0.5 and the expected utility of **B** is 0.685.

10.2.4 Behavioural Issues

Chapter 15 by Montibeller and von Winterfeldt (2017) in this volume discusses behavioural aspects related with the elicitation of subjective probabilities, in connection with psychological biases. Similar issues apply to the elicitation of preferences and value judgements.

As an example, we must pay attention to framing issues to ensure that the DM understands the questions asked to her. Gambles **A** and **B** stated above are framed as briefly with as many assumptions as in *Allais Paradox* (French and Ríos Insua 2000) and so are susceptible to similar "misunderstandings". Thus, in interacting with the DM, the analyst must discuss the description of the hypothetical choice to ensure that she understands the judgement asked of her. Also, the entire elicitation process should be enhanced with consistency checks to ensure that her judgements cohere, as illustrated above. Typical biases include certainty effects, anchoring based on the initial values assessed, serial dependence or risk distortions when dealing with probabilities close to 0 or 1. Moreover, we must take into account the imprecision in the DM's judgements, see Ríos Insua (1990). French et al. (2009) offer a description of an interview between an analyst and a DM, which illustrates some of these points.

10.3 (Multi-Attribute) Utility Elicitation

We describe now issues in relation with utility elicitation when the consequence space has a *multi-attributed* structure: a DM's preferences for the possible consequences are usually complex, formed by balancing conflicting objectives, so they must trade-off a variety of factors. We use the term *attribute* to name a factor which the DM wishes to take into account when making a decision. The term *(sub-)objective* is used to specify a factor which one wishes to maximise or minimise: i.e. an objective is "an attribute plus a direction of preference".

10.3.1 Multi-Attribute Hierarchies

As Brownlow and Watson (1987) point out, structuring attribute trees can help a DM overcome the cognitive overload brought by the volume of information which needs to be integrated into the solution of large, complex issues. There are cognitive advantages in arranging attributes in an *attribute hierarchy* or *tree*.

There are a number of ways that an analyst can work with DMs to build an attribute hierarchy. He may ask "top down" questions such as "What issues are you thinking about when you talk of *energy security*?". To which she may reply: "Oh, I guess natural and anthropogenic risks would be key factors." "What anthropogenic risks?" "Well, diversity of supply sources, legislative limitations,

workforce risks and terrorist threats, obviously." And so on. Such a discussion explores the meaning of the DM's overall objective analysing its components in a logical fashion. Alternatively the analyst may get the DM to brainstorm factors which will affect her preferences and then draw these together into a hierarchy. Moreover, the process may be a mixture of the two. Keeney (1992) is a key text describing how hierarchies can be built to meet a DM's needs.

Case: Attribute Hierarchy. Figure 10.2 provides a hierarchy for our case. The overall objective of maximising energy security—a nebulous concept—is broken down into two groups of attributes relating to the impacts of natural and anthropogenic risks. This last one is analysed further and divided into diversity of supply sources, legislative limitations, workforce risks and terrorist threats.

In such a manner, issues that matter in a decision are identified and grouped in a cognitively sensible way. In reality, the tree might be more highly structured than in Fig. 10.2 with, say, impacts of natural risks broken down into impacts of different types of natural disasters.

There are several requirements that objectives must meet if they are to be useful, see Keeney and Gregory (2005). Some of them correspond to being:

i *Comprehensive*: Covering the whole range of relevant consequences for the corresponding alternatives.
ii *Measurable*: Either objectively or subjectively, for each consequence.
iii *Non-overlapping*: Since two attributes should not measure similar aspects of consequences.
iv *Relevant*: In the sense of being capable of distinguishing between the alternatives.
v *Unambiguous*: Having a clear relationship between consequences and their description using the attribute.
vi *Understandable*: With consequences and value trade-offs made using the attribute readily understandable and clearly communicated.

The lowest nodes in the tree provide a series of dimensions, say q of them, which may be used to describe the consequences of alternatives, energy security policies in our case, and uncertain scenarios. The intention is that each of these attribute scales may be quantified, allowing each consequence to be represented as a vector of attribute levels: $c = (c_1, c_2, \ldots, c_q)$. We distinguish three types of scales.

Natural A *natural attribute* gives a direct measure of the objective involved and the attribute is universally understood. For example, repair costs in € is a natural attribute to evaluate the impacts of natural hazards, which we aim at minimising.

Constructed Some objectives are clearly subjective, lacking a clear, agreed measure: e.g. the external image impact of physical terrorist attacks. We used the number of attacks as a means of evaluating this attribute, but an alternative would have been to build an artificial ordinal scale, say from 1 to 10. Level 1 would have been associated with a situation of minimal impact. On the other hand, level 10 would have been associated with a maximum impact accident with total destruction of the Serbian energy infrastructure and numerous

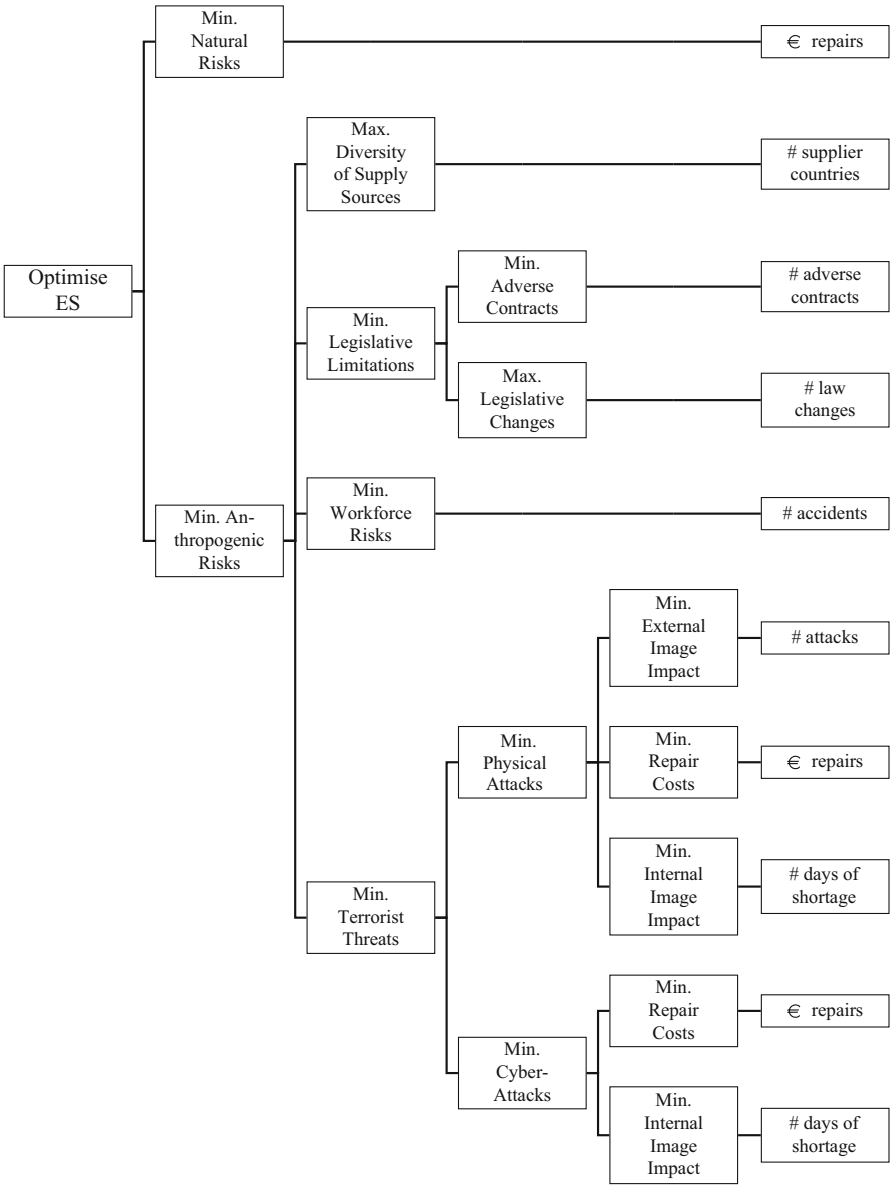


Fig. 10.2 Objectives of ES management at state level. The Serbian case

fatalities resulting in a tremendous image loss for Serbia. Henceforth, we would associate each of the levels with a qualitative description of severity with respect to image. *Constructed (or subjective) attributes* are created for a specific decision context and, therefore, are not universally understood.

Proxy *Proxy attributes* are used because of its perceived relationship to the objective, when no natural attributes are available and constructed scales are deemed ambiguous. The DM believes that variations in a proxy attribute correlate well with the issue of concern to her: e.g. legislative limitations could be measured through the amount of legislative changes, although it should be recognised that different laws may have different repercussion.

10.3.2 Multi-Attribute Utilities

Throughout the following, we assume that $\mathcal{C} = \mathcal{C}_1 \times \mathcal{C}_2 \times \ldots \times \mathcal{C}_q \subset \Re^q$ and write $c = (c_1, c_2, \ldots, c_q)$. We discuss here under what conditions can $u(\cdot)$ be written in a simplified form, e.g. an additive one. These are known as *independence conditions*.

We note first the importance of the simplification brought by such conditions. Without these, the DM could find the task cognitively complicated. Consider the choice between gambles **A** and **B**, similar to that in Sect. 10.2.2:

$$\text{Gamble } \mathbf{A}: \quad (c_1, c_2, \ldots, c_q) \qquad \text{for certain;}$$

$$\text{Gamble } \mathbf{B}: \begin{cases} (c_1^*, c_2^*, \ldots, c_q^*) & \text{with probability } p, \\ (c_{1*}, c_{2*}, \ldots, c_{q*}) & \text{with probability } 1-p. \end{cases}$$

Here the DM is being asked simultaneously to trade-off potential differences in q attributes and account for her attitude to risk. Since she is likely to find such tasks very difficult, the structured support of decision analysis may facilitate them.

Let $I \subset \{1, 2, \ldots, q\}$ be an index set, which we use to designate a subset of the components of c, and let $J = I^{\complement}$. Then, we write $c = (c_I, c_J)$ to represent a consequence in which we re-order the attributes to list first the attribute levels on $\mathcal{C}_I = \prod_{i \in I} \mathcal{C}_i$, and then those on $\mathcal{C}_J = \prod_{i \notin I} \mathcal{C}_i$. We say that attributes $\mathcal{C}_I$ are *preferentially independent* of $\mathcal{C}_J$ for the DM if

$$(c_I, \alpha_J) \precsim (c_I', \alpha_J), \text{ for some } \alpha_J \quad \Longrightarrow \quad (c_I, \beta_J) \precsim (c_I', \beta_J), \ \forall \beta_J.$$

Preferential independence seeks to capture the judgements behind statements of the form: "All other things being equal, I prefer…". When this holds whatever the index set I is taken, the attributes are said to be *mutually preferentially independent* for the DM. In decision making problems under certainty in which preferences may be modelled with a value function, we may decompose this into an additive form, see Chap. 12 in this volume (Morton 2017).

Turning to the case of modelling preference under uncertainty, we discuss two conditions: *utility* and *additive independence*. For an index set I, the DM's preferences for $\mathcal{C}_I$ are *utility independent* of $\mathcal{C}_J$ if her preferences for gambles in which for all possible consequences the levels of $\mathcal{C}_J$ are fixed at a common value c_J do not depend upon such fixed value. Then, her marginal utility function on $\mathcal{C}_I$ can be assessed independently of the attributes $\mathcal{C}_J$; and her marginal attitude to risk on

$\mathcal{C}_I$ does not depend upon $\mathcal{C}_J$. As an example, suppose that $\mathcal{C} = \mathcal{C}_1 \times \mathcal{C}_2$ and the utility function is

$$\begin{aligned} u(c_1, c_2) &= u_1(c_1) + u_2(c_2) + k\, u_1(c_1)\, u_2(c_2) \\ &= (1 + k\, u_2(c_2))\, u_1(c_1) + u_2(c_2) = \alpha(c_2)\, u_1(c_1) + \beta(c_2). \end{aligned}$$

It is clear that if $\alpha(c_2) > 0$ for any c_2, $u(c_1, c_2)$ considered as a utility function for c_1, for fixed c_2, is strategically equivalent to $u_1(c_1)$; i.e. $\mathcal{C}_1$ is utility independent of $\mathcal{C}_2$. A similar rearrangement shows that $\mathcal{C}_2$ is utility independent of $\mathcal{C}_1$; and so $\mathcal{C}_1$ and $\mathcal{C}_2$ are mutually utility independent. The converse is also true in the case of two attributes.

We consider now the main decompositions in q dimensions: attributes $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_q$. In all three cases considered, $u(\cdot)$ will be normalised so that $u(c_1^*, c_2^*, \ldots, c_q^*) = 1$ and $u(c_{1*}, c_{2*}, \ldots, c_{q*}) = 0$ for some $c_* \precsim c^*$; and $u_i(\cdot)$ will be a marginal utility function on $\mathcal{C}_i$, normalised so that $u_i(c_i^*) = 1$, $u_i(c_{i*}) = 0$, for $i \in \{1, 2, \ldots, q\}$. The involved constants are often called *weights* or *scaling constants*.

When the attributes are mutually utility independent, then

$$\begin{aligned} u(c) = \sum_{i=1}^{q} k_i\, u_i(c_i) + k \sum_{i=1, j>i}^{q} k_i\, k_j\, u_i(c_i)\, u_j(c_j) \\ + k^2 \sum_{i=1, j>i, \ell>j}^{q} k_i\, k_j\, k_\ell\, u_i(c_i)\, u_j(c_j)\, u_\ell(c_\ell) \\ + \cdots \\ + k^{q-1}\, k_1\, k_2 \ldots k_q\, u_1(c_1)\, u_2(c_2) \cdots u_q(c_q), \end{aligned}$$

where $k_i = u(c_i^*, c_{\{i\}^{\mathsf{C}}*})$ and k satisfies $1 + k = \prod (1 + k\, k_i)$. Note that if $k = 0$, we have an *additive utility function*; whereas if $k \neq 0$ we have a form called a *multiplicative utility function* which, after simple rearrangements, is $1 + k\, u(c) = \prod (1 + k\, k_i\, u_i(c_i))$.

When the DM only holds each $\mathcal{C}_i$ to be utility independent of the other $(q-1)$ attributes, then

$$\begin{aligned} u(c) = \sum_{i=1}^{q} k_i\, u_i(c_i) + \sum_{i=1, j>i}^{q} k_{ij}\, u_i(c_i)\, u_j(c_j) \\ + \sum_{i=1, >i, \ell>j}^{q} k_{ij\ell}\, u_i(c_i)\, u_j(c_j)\, u_\ell(c_\ell) \\ + \cdots \\ + k_{12\ldots q}\, u_1(c_1)\, u_2(c_2) \cdots u_q(c_q), \end{aligned}$$

where the constants are defined by

$$k_i = u(c_i^*, c_{\{i\}^{\complement}*}),$$

$$k_{ij} = u(c_{\{i,j\}}^*, c_{\{i,j\}^{\complement}*}) - k_i - k_j,$$

$$k_{ij\ell} = u(c_{\{i,j,\ell\}}^*, c_{\{i,j,\ell\}^{\complement}*}) - k_i - k_j - k_\ell - k_{ij} - k_{i\ell} - k_{j\ell},$$

etc. This form is known as a *multi-linear utility function*. When $q = 2$, the multiplicative and multi-linear forms coincide.

The attributes are *additively independent* if the DM's preferences between gambles on $\mathcal{C} = \mathcal{C}_1 \times \ldots \times \mathcal{C}_q$ depend only on her marginal probability distributions over the $\mathcal{C}_i$ (and not the full joint distribution over $\mathcal{C}$). Additive independence implies that her attitude to risk on each of the attributes does not depend on the other $(q-1)$ attributes. Therefore, if the attributes are additively independent, then

$$u(c) = \sum_{i=1}^{q} k_i \, u_i(c_i),$$

where $k_i = u(c_i^*, c_{\{i\}^{\complement}*})$.

We noted in Sect. 10.1 that a utility function u may be constructed by transforming a value function: $u(\cdot) = \xi(v(\cdot))$. Suppose that $v(c) = \sum v_i(c_i)$ is additive and we take $\xi(x) = 1 - \exp(-\rho\, x)$ to be an exponential unidimensional utility. Then

$$u(c) = 1 - \exp(-\rho \sum v_i(c_i)) = 1 - \prod \exp(-\rho\, v_i(c_i)).$$

It follows that the attributes must be mutually utility independent. Thus, the previous expression must provide an alternative form of the multiplicative utility function. We can show that if the attributes $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_q$ $(q \geq 3)$ are compatible with an additive value function $v(c) = \sum v_i(c_i)$ and $\mathcal{C}_i$ is utility independent of the remaining $(q-1)$ attributes, then the utility function $u(\cdot)$ must have one of the following forms:

1. $u(c) = 1 - \exp(-\rho \sum v_i(c_i)), \quad \rho > 0.$
2. $u(c) = \sum v_i(c_i).$
3. $u(c) = -1 + \exp(\rho \sum v_i(c_i)), \quad \rho > 0.$

10.3.3 Time Dependent Utilities

We consider now the form of utility functions when each consequence results in a timestream of outcomes. We focus on monetary outcomes and assume that a particular consequence comprises the receipt of a sum c_i at time t_i. Assume that the times t_i are equally spaced for $i = 1, 2, \ldots, q$ (although this is not strictly

necessary). The problem is structured so that the consequences are multi-attributed: $c = (c_1, c_2, \ldots, c_q)$. When q is finite, we may use all the above developments to structure and assess $u(\cdot)$; but doing so misses the extra structure brought in by the temporal context. Here we describe a number of key issues.

Firstly, discounting models are commonly used to develop $u(\cdot)$, having the form:

$$u(c) = \sum_{i=1}^{q} \rho^{i-1}\, \omega(c_i), \tag{10.2}$$

where ρ is a discount factor and $\omega(\cdot)$ is a uni-dimensional utility function common to all times. In case that the c_i are monetary and $\omega(\cdot)$ is the identity function, (10.2) gives the *net present value* (NPV). This model can be justified in a number of ways, but the most telling involves a principle of *stationarity*.

> *Stationarity*. Let $(c)^t$ be a timestream in which c is received at time t and nothing is received at any other time. Then, the DM holds $(c)^t \precsim (d)^s$ if and only if $(c)^{t+\ell} \precsim (d)^{s+\ell}$ for any $\ell > 0$.

Stationarity demands that the DM's intertemporal trade-off between two periods depends only on the relative time between the periods and not on the absolute time they occur.

Economists and others have long found such arguments persuasive: see, e.g. Strotz (1955–1956). Therefore, discounting models are extremely common. For instance, NPV, in which the $\omega(c_i)$ in (10.2) are taken as monetary values, is used throughout much of industry, commerce and government to evaluate projects. It has the property that as $q \to \infty$, the later terms in (10.2) tend rapidly to zero. This can be a vital property since it guarantees convergence of many of the summations that need be evaluated.

It is unfortunate that this rapid convergence of terms to zero is precisely the property that brings some applications of discounting into question. For instance, its implications in decisions on disposal of nuclear waste are that financial and other costs on future generations are effectively neglected. For this reason several authors have suggested models which decay more slowly; e.g. Ahlbrecht and Weber (1995) investigate models of the form:

$$u(c) = \sum_{i=1}^{q} \rho^{\alpha(i)} \omega(c_i).$$

This is the standard discounting model when $\alpha(\cdot)$ is a linear function; but it can accommodate slower (or faster) decay for other choices of $\alpha(\cdot)$. The difficulty is that unless $\alpha(\cdot)$ is linear, the preferences modelled cannot be stationary. For many, this is not a problem because for any real DM her information and circumstances do change over time and the opportunity to make a decision may only occur at one point in time and be irrevocable later.

Atherton and French (1997) provide a survey of behavioural studies on intertemporal preferences and discounting, while Atherton and French (1998) offer an alternative structuring of long term consequences which avoids many of the discounting issues.

10.3.4 An Elicitation Protocol

We discuss now how multi-attribute utility functions may be elicited. Consider first the additive or multiplicative case. A protocol would run like this:

1. Elicit the q single attribute utility functions $u_i(\cdot)$ separately with the methods in Sect. 10.2.2.
2. Identify the scaling constants k_i. This can be accomplished, for example, by *swing-weighting* (von Winterfeldt and Edwards 1986) in which trade-offs under certainty are required between a pair of attributes. First the DM is asked to order the attributes such that for $i < j$:

$$(c_i^*, c_{\{i\}\complement *}) \succeq (c_j^*, c_{\{j\}\complement *}).$$

   This might be done by asking her if she has the worst possible consequence c_* and could increase just one attribute level to its best value, which would it be. Then, which is the second attribute that she would choose to raise, and so on. Next she is asked to consider pairs of attributes, keeping the other $(q-2)$ fixed at their worst values and to identify c_i such that:

$$(c_i, c_{\{i\}\complement *}) \sim (c_j^*, c_{\{j\}\complement *}).$$

   Because of her ordering of the attributes, c_i will lie below c_i^* in her marginal preference order on $\mathcal{C}_i$. Such indifferences give simple linear equations from which the k_i may be determined.

   In the multiplicative case, a further indifference will be required in which one consequence has two attributes different from the worst values in order to determine the single k.
3. Finally $u(\cdot)$ may be formed by the appropriate additions and multiplications of the marginal utilities.

Keeney and Raiffa (1993) contain further discussion of these elicitation processes and also of the multi-linear case, which with its greater number of scaling constants needs a more subtle series of elicitations. Besides, the *swing-weighting method* mentioned above, there are other methods for direct elicitation of the various attribute weights in decision analysis including the *ratio method* (Edwards 1977) and the *trade-off and pricing-out methods* (Keeney and Raiffa 1993).

An alternative and possibly easier approach to eliciting multi-attribute utilities is to use $u(\cdot) = \varphi(v(\cdot))$, where $v(\cdot)$ is a multi-attribute value function. This process clearly separates the task of trading off attributes from that of considering attitude to risk and thus again helps the DM by not confounding two cognitively difficult tasks. Assuming that an additive value function is appropriate, we have the following assessment process:

1. Assess a measurable value function $\varphi_i(\cdot)$ on each attribute $\mathcal{C}_i$.
2. Noting that each $\varphi_i(\cdot)$ is unique up to a positive affine transformation, it is next necessary to bring each to a common scale. This may be done, for example, by *swing weighting*, as discussed above. Take $v_i(\cdot) = k_i\,\varphi_i(\cdot)$ as the consistently scaled marginal value function on $\mathcal{C}_i$.
3. Form the overall measurable additive value function by addition:

$$v(c) = \sum k_i\,\varphi_i(c_i).$$

4. The interval $[v(c_*), v(c^*)]$ now defines a domain for a uni-dimensional utility function. This may be assessed by the methods described previously with the complication that the consequences involved in the elicitation, say the c in Gamble **A**, need be chosen so that:
   a. They identify sufficient, well spaced points along $[v(c_*), v(c^*)]$ for $u(\cdot)$ to be sketched in.
   b. As many as possible of the attribute levels in c are set at their extreme values so that the DM can focus her judgements on changes in one or two attributes only.

For an example in aviation safety, see Ríos Insua et al. (2016).

Case: Multi-Attribute Utility Function. We describe now the assessment of the multi-attribute utility function in our case. For space reasons, we just limit the assessment to two of the attributes in Fig. 10.2:

1. Mitigation of the risks associated with natural hazards (total amount of € spent in repairs).
2. Attenuation of the internal image impact for physical terrorist attacks (days of shortage in supply per year).

The utility $u_1(\cdot)$ for the first attribute was assessed in Sect. 10.2.3. Using similar methods, we assess the utility function for the second attribute, which we represent in Fig. 10.3, with the form $u_2(c_2) = 1 + \gamma\,c_2$ where $\gamma \approx -0.00848$. This corresponds to a risk neutral component utility.

We determine that the utility function is additive. We used a swing-weighting method to determine the corresponding weights at $k_1 \approx 0.36$ and $k_2 \approx 0.64$. For example, for a combined cost of 500M€ (worst case) and shortage of 0 days (best case), the associated utility is 0.64. This means that our expert found equally desirable a sure loss of 500M€ with no shortage to a lottery which gives her a combined cost of 500M€ and shortage of 120 days with probability 0.64 and

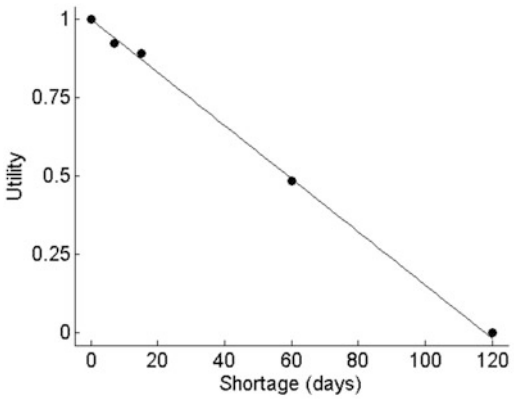


Fig. 10.3 Utility values and fitted utility function $u_2(\cdot)$

no cost and no shortage with probability 0.36. To come out with that value, we first offered the DM the reference lottery with probability $1/2 = 0.5$ and she responded that she preferred the certain loss of 500M€ with no shortage, therefore suggesting that $k_2 > 0.5$. Then, we offered her the reference lottery with probability $1/2 + 1/4 = 0.75$ and she said that she preferred the lottery, therefore suggesting that $k_2 < 0.75$ and, consequently, $0.5 < k_2 < 0.75$. Next, we offered her the reference lottery with probability $1/2 + 1/4 - 1/8 = 0.625$ and she again preferred the certain amount. We iterated the procedure until equivalence was found in the 6th iteration, rounding the value to 0.64. Then, the utility function used is

$$\begin{aligned} u(c_1, c_2) &= 0.36\,(1 + 0.153\,(1 - \exp(0.004\,c_1))) + 0.64\,(1 - 0.008\,c_2) \\ &= 1.055 - 0.055\,\exp(0.004\,c_1) - 0.005\,c_2. \end{aligned}$$

10.4 Eliciting Adversarial Preferences

The previous sections outlined how to assess the preferences of a DM we aim to support. Recently, see Banks et al. (2015) for a review, there has been an interest in modelling the preferences of adversaries whose decisions affect the performance of a system of interest to the DM we support. Typical applications include security, cybersecurity, competitive marketing or social robotics. The problem is also of interest in non-cooperative game theory, see e.g. Gibbons (1992), although its literature remains silent about this problem.

To start with, we shall usually have information about the multiple interests of the attackers. For example, Keeney (2007) and Keeney and von Winterfeldt (2010) provide what may be viewed as catalogues from which we can choose appropriate criteria in the domain of terrorism. Keeney (2007) suggests that methods similar to the ones described in Sects. 10.2 and 10.3 may be used by interviewing experts in the problem at hand, therefore producing utility functions modelling the preferences of the adversaries.

However, note that we are not directly eliciting preferences from the adversary, but rather from a surrogate of the adversary. Thus, intrinsically we have uncertainty about the adversarial preferences. One possible approach, illustrated in Banks et al. (2015), would aggregate the objectives with a weighted measurable value function, as in Dyer and Sarin (1979). Using the relative risk aversion concept (Dyer and Sarin 1982), we could assume risk proneness when modelling the attacker's utility function, see Sect. 10.3.2. Finally, the uncertainty associated with the attacker's utility would be reflected through distributions over the weights and risk proneness coefficients. For this, we may ask experts to elaborate such distributions, or ask several experts to provide point estimates of the weights and coefficients and build a distribution from them.

An alternative approach for obtaining a distribution over the adversaries preferences is described in Wang and Bier (2013). As before, we assume that the adversarial preferences are represented by a multi-attribute utility function, which may include unobserved attributes that are important to the adversary but have not been identified by the defender. For simplicity, we consider the adversary's utility to be linear in each of the attributes and these attributes to be additively independent of each other. The task is then to derive probability distributions that can match the rank orderings of target valuations provided by several experts. To do this, we use as input such rankings and as output a distribution over the adversaries preferences. Two methods are suggested by Wang and Bier (2013). One is an adaptation of *probabilistic inversion* due to Neslo et al. (2008); basically, it identifies a probability distribution over the attribute weights that can reproduce the stated (theoretical or empirical) marginal distributions over the experts' rank orderings of target attractiveness, based on a closeness criteria according to Kullback-Leibler distance. The other one uses a *Bayesian density estimation* approach, see Müller et al. (2015).

Case: Adversarial Preferences. We assess some of the adversarial preferences for the case study. We focus on the adversary Al-Nusra, whose attack strategy is based on lone wolves. Their objectives will be different to more organised groups like Al-Qaeda or adversarial states. In the case of lone wolves, their main objective would be to kill as many people as possible.

Our expert provides her preferences over such attribute based on her knowledge of that organisation. The minimum number of dead people over the planning period would be 0 (no victims). As the maximum we adopt 55, associated with the bloodiest suicide bombing event carried out by an Al-Nusra terrorist. We use again a probability equivalent method to assess the utilities of three intermediate values, as displayed in Table 10.2.

Table 10.2 Utilities for five values of killings

Deaths	0	10	25	40	55
Utility	0.00	0.27	0.33	0.58	1.00

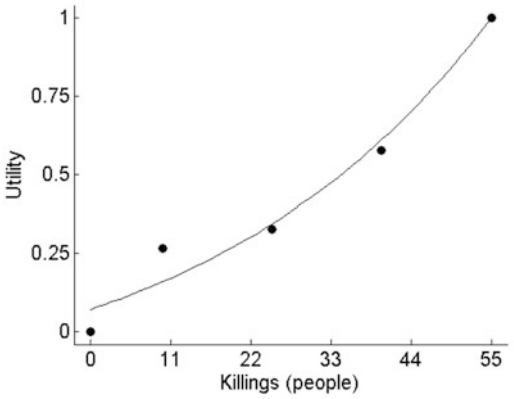


Fig. 10.4 Utility values and fitted utility function $u_A(\cdot)$

Table 10.3 Utility intervals for five values of killings

Deaths	0	10	25	40	55
Utility	0.00	[0.22, 0.29]	[0.30, 0.36]	[0.52, 0.65]	1.00

Figure 10.4 shows the fitting of the utility function based on the data in Table 10.2. The data suggest fitting a curve of the form $u_A(c) = 1 + \lambda\,(\exp(\rho\,(c - 55)) - 1)$ where $\lambda, \rho \geq 0$. Through least squares we identify that the parameters are $\lambda \approx 1.24359$ and $\rho \approx 0.02501$, corresponding to a risk prone utility.

To come out with a random utility model, we may proceed in several ways. For example, rather than assuming precise utilities as in Table 10.2, intervals based on the answers of the expert prior to fixing the probability equivalents may be considered, as in Table 10.3.

We then fit the utilities to the upper and lower probability equivalents with results:

- $\lambda^* \approx 1.72835$ and $\rho^* \approx 0.01423$;
- $\lambda_* \approx 1.10233$ and $\rho_* \approx 0.03416$.

Following that the random utility model is defined by $u(c) = 1 + \Lambda\,(\exp(\mathrm{P}\,(c - 55)) - 1)$, with $\Lambda \sim \mathcal{U}(1.10233, 1.72835)$ and $\mathrm{P} \sim \mathcal{U}(0.01423, 0.03416)$.

10.5 Discussion

We have provided an introduction to preference modelling and utility elicitation. The cases of single and multiple criteria have been covered as well as issues in relation with adversarial preferences. We have illustrated the methods with a case in energy security.

With reference to preference modelling, we note the seminal works of Keeney and Raiffa (1993) and Keeney (1992). Discussions of multi-attribute modelling may

be found in French et al. (1998) and Wright and Goodwin (1999). Other relevant references include Bell et al. (1977), Belton (1990), Edwards (2013), French (1986), French and Smith (1997) and von Winterfeldt and Edwards (1986). For further discussion of attitudes to risk and their relation to utility function shapes, see Eeckhoudt et al. (1995), Gelles and Mitchell (1999), Keeney and Raiffa (1993), Pratt (1964) and Prelec and Loewenstein (1991). The elicitation and assessment of utilities are discussed in Farquhar (1984), Keeney and Raiffa (1993) and Wright and Goodwin (1999).

In light of experiments as in Allais' or Ellsberg's paradoxes, see French and Ríos Insua (2000), non-expected utility theories seek to weaken the assumptions of SEU in the hope that the weaker axioms will be more acceptable to DMs. Such generalisations of SEU are known as: *non-linear preference theories*, because they often reject or modify independence conditions which are responsible for SEU's linearity with respect to the probabilities; *non-transitive utility theories*, because the generalisation allows preferences to be intransitive in certain circumstances; or, simply, *non-expected utility theories*, for obvious reasons. For a review see Wakker (2004).

Drawing on recent developments in technology and social media, there has been an upsurge in *preference analytics*, that is the elicitation of preferences based on information obtained from social networks such as Twitter or search engines such as Google, see Daniell et al. (2016). Related big data issues call for the research of new techniques. Applications are wide, but normally focus on the identification of customer preferences as a means to deliver personalized services. As an example, online games capable of analysing the preferences of players have been developed recently with the aim of providing more attractive products. Areas of interest include combining preference elicitation methods as here described with preference analytics methods and using preference analytics in adversarial problems.

Acknowledgements The work of DRI is supported by the Spanish Ministry of Economy and Innovation program MTM2014-56949-C3-1-R and the AXA-ICMAT Chair on Adversarial Risk Analysis. Besides, 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 ESF-COST Action IS1304 on Expert Judgement and ICMAT Severo Ochoa project SEV-2015-0554 (MINECO).

References

Ahlbrecht M, Weber M (1995) Hyperbolic discounting models in prescriptive theory of intertemporal choice. Z Wirtsch Sozialwissen 115:813–826

Atherton E, French S (1997) Issues in supporting intertemporal choice. Essays in decision making: a volume in honour of Stanley Zionts. Springer, Berlin, pp 135–156

Atherton E, French S (1998) Valuing the future: a MADA example involving nuclear waste storage. J Multi-Criteria Decis Anal 7(6):304–321

Banks D, Ríos J, Ríos Insua D (2015) Adversarial risk analysis. CRC Press, Boca Raton

Bell D, Keeney R, Raiffa H (1977) Conflicting objectives in decisions. Wiley, New York

Belton V (1990) Multiple criteria decision analysis: practically the only way to choose. Strathclyde Business School, Glasgow
Brownlow S, Watson S (1987) Structuring multi-attribute value hierarchies. J Oper Res Soc 38(4):309–317
Daniell K, Morton A, Ríos Insua D (2016) Policy analysis and policy analytics. Ann Oper Res 236(1):1–13
Dyer J, Sarin R (1979) Group preference aggregation rules based on strength of preference. Manag Sci 25(9):822–832
Dyer J, Sarin R (1982) Relative risk aversion. Manag Sci 28(8):875–886
Edwards W (1977) How to use multiattribute utility measurement for social decisionmaking. IEEE Trans Syst Man Cybern 7(5):326–340
Edwards W (2013) Utility theories: measurements and applications. Springer, New York
Eeckhoudt L, Gollier C, Schlesinger H (1995) The risk-averse (and prudent) newsboy. Manag Sci 41(5):786–794
Farquhar P (1984) State of the art—utility assessment methods. Manag Sci 30(11):1283–1300
French S (1986) Decision theory: an introduction to the mathematics of rationality. Halsted Press, New York
French S, Ríos Insua D (2000) Statistical decision theory. Edward Arnold, London
French S, Smith J (1997) The practice of Bayesian analysis. Hodder Education Publishers, London
French S et al (1998) Problem formulation for multi- criteria decision analysis: report of a workshop. J Multi-Criteria Decis Anal 7(5):242–262
French S, Maule J, Papamichail N (2009) Decision behaviour, analysis and support. Cambridge University Press, Cambridge
Gelles G, Mitchell D (1999) Broadly decreasing risk aversion. Manag Sci 45(10):1432–1439
Gibbons R (1992) Game theory for applied economists. Princeton University Press, Princeton
Keeney R (1992) On the foundations of prescriptive decision analysis. Utility theories: measurements and applications. Springer, Dordrecht, pp 57–72
Keeney R (2007) Modeling values for anti-terrorism analysis. Risk Anal 27(3):585–596
Keeney R, Gregory R (2005) Selecting attributes to measure the achievement of objectives. Oper Res 53(1):1–11
Keeney R, Raiffa H (1993) Decisions with multiple objectives: preferences and value tradeoffs. Cambridge University Press, Cambridge
Keeney G, von Winterfeldt D (2010) Identifying and structuring the objectives of terrorists. Risk Anal 30(12):1803–1816
Kreft S, Eckstein D, Drosch L, Fischer L (2015) Global climate risk index 2016: who suffers most from extreme weather events? Weather-related loss events in 2014 and 1995 to 2014. Germanwatch e.V, Bonn
Montibeller G, von Winterfeldt D (2017) Individual and group biases in value and uncertainty judgments. In: Dias LC, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgment. Springer, New York. doi:10.1007/978-3-319-65052-4_15
Morton A (2017) Multiattribute value elicitation. In: Dias LC, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgment. Springer, New York. doi:10.1007/978-3-319-65052-4_12
Müller P, Quintana FA, Jara A, Hanson T (2015) Bayesian nonparametric data analysis. Springer series in statistics. Springer, Cham
Neslo R et al (2008) Modeling stakeholder preferences with probabilistic inversion: application to prioritizing marine ecosystem vulnerabilities. Real-time and deliberative decision making. Springer, Dordrecht, pp 265–284
Pratt J (1964) Risk aversion in the small and in the large. Econometrica 32(1–2):122–136
Prelec D, Loewenstein G (1991) Decision making over time and under uncertainty: a common approach. Manag Sci 37(7):770–786
Ríos Insua D (1990) Sensitivity analysis in multi-objective decision making. Lecture notes in economics and mathematical systems, vol 347. Springer, Berlin, pp 74–126

Ríos Insua D, Alfaro C, Gómez J, Hernández-Coronado P, Bernal F (2016) A framework for aviation safety risk management at state level. Reliab Eng Syst Saf doi:10.1016/j.ress.2016.12.002

Strotz R (1955–1956) Myopia and inconsistency in dynamic utility maximization. Rev Econ Stud 23(3):165–180

von Winterfeldt D, Edwards W (1986) Decision analysis and behavioral research. Cambridge University Press, Cambridge

Wakker P (2004) Preference axiomatizations for decision under uncertainty. In: Uncertainty in economic theory: essays in honor of David Schmeidler's 65th birthday. Routledge, Abingdon, pp 20–35

Wang C, Bier V (2013) Expert elicitation of adversary preferences using ordinal judgments. Oper Res 61(2):372–385

Wright G, Goodwin P (1999) Rethinking value elicitation for personal consequential decisions. J Multi-Criteria Decis Anal 8(1):3–10

Chapter 11
Elicitation in Target-Oriented Utility

Robert F. Bordley

Abstract Target-oriented utility theory interprets the utility of a consequence as the probability of the consequence exceeding some benchmark random variable. This shifts the focus of utility assessment to the identification of the benchmark and the sources of uncertainty in that benchmark. Identification of the benchmark is often easy when the benchmark is based on a status quo outcome, a preferred outcome or an undesirable outcome. Benchmarks are generally easy to communicate and easy to track. Once identified, data and models can then be used to describe the uncertainty in the benchmark. This approach can be useful in those applications where the utility function needs to be justified to others.

11.1 Introduction

Target-oriented utility theory interprets the utility of a consequence as the probability of the consequence exceeding some benchmark random variable. In many problems, the benchmarks are easy to interpret and communicate. As a result, the problem of utility assessment simplifies to the problem of assessing probabilities over this uncertain benchmark. This is useful since many clients of decision analysis, while familiar with probability theory, are unfamiliar with the concept of utility (as understood by decision analysts.)

In addition defining the utility function in terms of probabilities allows the utility function to be informed by data and models. A data-based utility function can be useful when the decision maker must justify that utility function to others. For example,

1. An executive may want colleagues (both current colleagues and their future successors) to accept the utility function they use. This can be critical in long-term decisions where that executive may be replaced by new management before the decision is fully implemented. The executive's successors cannot always be

R.F. Bordley (✉)
Systems Engineering and Design, University of Michigan, Ann Arbor, MI, USA
e-mail: rbordley@umich.edu

L.C. Dias et al. (eds.), *Elicitation*, International Series in Operations Research & Management Science 261, https://doi.org/10.1007/978-3-319-65052-4_11

expected to continue implementing a decision if they do not understand, or agree with, the rationale for the previous decision.
2. An executive's decision can often be contested by disappointed third parties. For example, government procurement officers are often concerned with being sued by the vendor to whom a procurement contract was not awarded. If sued, they must prove to a judge that the utility function was not arbitrary, not driven by inappropriate personal considerations and consistent with what a 'reasonable man' would use.

Of course, data is sometimes used in the conventional approach to utility assessment. For example, General Motors executives were once asked to specify their willingness to pay for fuel economy. But before making this assessment, they were told the average amount by which a 1% change in fuel economy reduce a household's life-time gasoline costs. The executives then adjust this estimate upwards to reflect the strategic value associated with improving fuel efficiency. But as will be shown later, target-oriented utility assessment allows for an even more intensive use of data in informing the utility function.

Before discussing these aspects of target-oriented utility, we review its theoretical rationale. First suppose that an individual's preference for an outcome is solely determined by whether it exceeds some possibly uncertain benchmark. For example, suppose the individual is a firm trying to increase its sales. Then expected sales is the sum of the probability of each buyer choosing the firm's products over the competitor's products. The buyer's probability of choosing the firm's products is the probability of the buyer's perceived value for the firm's products and services exceeding the perceived value of competing products and services. As a result, the sales-maximizing firm will use a target-oriented utility with the benchmark being the perceived value of competing products.

This example showed that there are many important applications in which the utility function is target-oriented. But we now show that this, at least in theory, is true in all cases, i.e., that ANY utility function can ALWAYS be written as the probability of exceeding some benchmark random variable:

1. The axioms of utility theory presume that, in the absence of uncertainty, all consequences of interest can be ranked based on their preferability. This ranking (or any monotonic transformation of this ranking) is a value function. The utility function is a special kind of value function, i.e., a special kind of transformation of the ranking function, which adjusts for risks in an especially useful way.
2. Since utility functions are bounded, utility can be rescaled to lie between zero and one. Utility is non-decreasing in the value function (and typically right-continuous). This implies the existence of a benchmark random variable (Billingsley 1995) whose cumulative distribution function is the utility function. So the utility of any consequence is the probability of that consequence's value exceeding this benchmark random variable.
3. Since the utility of a gamble is the expected value of the utility of its outcomes, the utility of a gamble will likewise correspond to the probability of the gamble

having outcomes exceeding the benchmark random variable. (Exceeding this benchmark will be referred to as 'achieving success.')

As a result, any individual satisfying the decision theory axioms of rationality will make decisions as if they maximized

(a) Either the expected value of some appropriately defined utility function.
(b) Or the probability of achieving success

The axioms of rationality make no assertion about how a rational individual actually makes decisions. What they do imply is that if we could elicit an individual's subjective probabilities, their utility function and their 'success probabilities', then we could determine what decisions the individual should make if they were rational.

This suggests two different approaches toward prescriptive decision analysis, i.e., toward helping individuals make rational decisions. Both approaches require the elicitation of an individual's subjective probabilities. One approach elicits the utility function directly using 'conventional' techniques. The other 'target-oriented' approach elicits the random benchmark. So which approach is more useful in practice?

Eliciting the random benchmark involves specifying what it means to exceed the benchmark and thus involves specifying the success event in sufficient clarity so as to satisfy Howard's clairvoyant test (Howard 1988). As a result, the event must be defined so that it is hypothetically possible, at some point in the future, to unambiguously determine whether or not an outcome has exceeded this random benchmark. The clairvoyant test can never be satisfied for certain quantum mechanical events because of Heisenberg's Uncertainty Principle. And likewise, it may not always be possible to define a random benchmark which satisfies the clairvoyant test. Of course, there may be surrogate measures that do satisfy the clairvoyant test which can be used in place of the actual benchmark. (Surrogate measures, e.g., normalized scores or monetary values, are also commonly used in conventional utility assessment). But for an example of where target-oriented utility might not be useful, consider a consumer choosing among desserts on the basis of which is most pleasurable. The utility function would be target-oriented if the consumer was maximizing the probability of the dessert's pleasurability equaling or exceeding the pleasurability of some past desert whose pleasurability was measurable. But the utility function would not be target-oriented if the consumer simply wanted to maximize the pleasure they obtained from the desert.

Hence this paper does not claim that the target-oriented approach is always more useful than the conventional approach.

So a key challenge in using the target-oriented approach is identifying contexts where the benchmark random variable can be meaningfully defined. To specify some contexts in which the benchmark is meaningful and observable, note that applying the mathematical techniques of decision analysis presumes

- A decision maker
- A choice set of alternatives from which the decision maker will choose,
- Some preferences over the possible consequences of each alternative.

But before these three presumptions are satisfied, there will typically be:

1. A default decision: There is some default decision, e.g., to continue with current plans, which the decision maker would have implemented if decision analysis were not used. In some cases, the decision maker temporarily suspends implementing this default decision in order to explore decision analysis.
2. Goals: There are higher-level goals which shape the decision maker's preferences over alternatives. Problem framing may identify higher-level goals.
3. Screening criteria: The actual choice set was created by starting with a larger list of choice alternatives and then using screening criteria to eliminate all but the few considered in the choice set.
4. Innate expectations: Most decision makers have expectations which shape their preferences over alternatives.

Target-oriented utility elicitation differs from conventional utility elicitation in explicitly using these factors to structure the utility function in terms of random benchmarks. Once the benchmarks are specified, probability elicitation (Spetzler and Stael von Holstein 1975) is required to define a probability of outperforming the benchmark. This probability then defines the utility function. Since probability elicitation methods are discussed elsewhere in this volume, this section focuses on the identification of the benchmark.

The next four sections focus on benchmarks based on default decisions, goals, screening criteria and expectations respectively.

11.2 Default Decisions

In the absence of any explicit decision, a decision maker will continue on some default course of action (e.g., make no change to current plans). Hence, before any decision can be made, the decision maker has to 'declare a decision', (Parnell et al. 2013) i.e., declare that they are willing to invest time and resources in contemplating a change from the default course of action. Suppose that the decision maker, upon declaring a decision, chooses to make that decision using decision analysis.

The outcomes of the decision recommended by decision analysis will typically be uncertain. But the outcomes of the default course of action will also typically be uncertain. There has been growing interest in encouraging businesses to run controlled experiments to evaluate the merits of different methodologies (Davenport 2009; Anderson and Simester 2011; Thomke and Manzi 2014). Such an approach, applied to decision analysis, would

1. Develop a process for identifying decision problems requiring resolution
2. Randomly decide which decision problems would undergo decision analysis and which would not. (The company might use stratified sampling to ensure that there was no difference in impact between the problems assigned to decision analysis.)

3. Define the control group as the set of problems where the default course of action (and not decision analysis) was implemented
4. Define an experimental group of problems where decision analysis was applied
5. Use statistics to compare the size of the difference in the goodness of the outcomes in the experiment group with the control group.

The most general statistical measure of the size of a difference is the common language effect size measure[1] which includes many commonly used effect size measures as special cases. The common language effect size measure is simply the probability of the payoffs in the experimental group exceeding the payoffs in the control group. Thus it is the probability of a randomly chosen decision recommended by decision analysis outperforming a randomly chosen default decision. This statistical effect size measure corresponds to a target-oriented utility if the random benchmark is defined as the uncertain outcomes of default decisions. Hence the target-oriented utility can measure the incremental benefit of using decision analysis.

While there is growing interest in having companies experimentally test different methodologies, companies typically do not randomly assign problems in this way. An alternate approach records the default course of action, and then compares the outcomes of the decision analysis with what one thinks might have happened with the default course of action. A more sophisticated approach to identifying what might have happened with the default course of action uses Bayesian structural time-series and state-space diffusion-regression models (Brodersen et al. 2015) to infer the counterfactual response had no intervention taken place. A less rigorous approach is to simply make within industry comparisons of companies who use decision analysis with companies that do not. In this case, the uncertain benchmark becomes the performance of the companies not using decision analysis. Thus it is common to compare the market capitalization of Chevron, an avid user of decision analysis, with the market capitalization of other large oil companies who do not use decision analysis.

If the uncertain outcomes of the default decision are used to specify the random benchmark, then maximizing the target-oriented utility will identify that alternative with the highest probability of outperforming the default decision. This can be an appealing criterion to champions of decision analysis who want to minimize the probability of users regretting following the recommendation of a decision analysis. Since good decisions can lead to bad outcomes, this probability will usually be non-zero. As the author learned at General Motors, organizational enthusiasm for decision analysis usually wanes when the decisions recommended by decision analysis do not improve upon the default decision.

[1] Effect size measures (Grissom and Kim 2005), while originally introduced by Fisher and Pearson, as a complement to their statistical significance measure, is now sometimes used in place of it. Arguably one of the most widely used effect size measures, Cohen's d, is simply the mean difference between experimental and control outcome divided by the standard deviation. It is a special case of the common language effect size measure when all uncertainties are Gaussian.

If there is a single decision maker, there will generally be a single default decision. But in many cases, there are multiple stakeholders which each have their other default decisions. For example, in negotiation theory, each partner to the negotiation has a best alternative to a negotiated agreement (BATNA). This represents the minimum that each partner should be willing to accept from the agreement. If a designer is attempting to design a system that all N parties to the negotiation find acceptable, then the designer will want to make design decisions that have the highest probability of exceeding everyone's BATNA. So if X_k is the uncertain payoff from the kth stakeholder's BATNA, the designer wants to make decisions with an uncertain payoff, X, exceeding the maximum of $X_1 \ldots X_N$. If these uncertain payoffs are independent and distributed with a Gumbel distribution:

$$\Pr\left(X_k \le x\right) = \exp\left(-\exp\left(-\left(x - m_k\right)/s\right)\right)$$

Then the probability of the maximum of $X_1 \ldots X_N$ being less than x will also have a Gumbel distribution:

$$\Pr\left(\max\left(X_1 \ldots X_n\right) \le x\right) = \exp\left(-n \ \exp\left(-x/s\right) + \sum \exp\left(m_k/s\right)\right)$$
$$= \exp\left(-\exp\left(-\left(x - m^*\right)/s\right)\right.$$

where $m^* = s \ln\left((1/n)\sum \exp(m_k/s)\right)$. See Fig. 11.1.

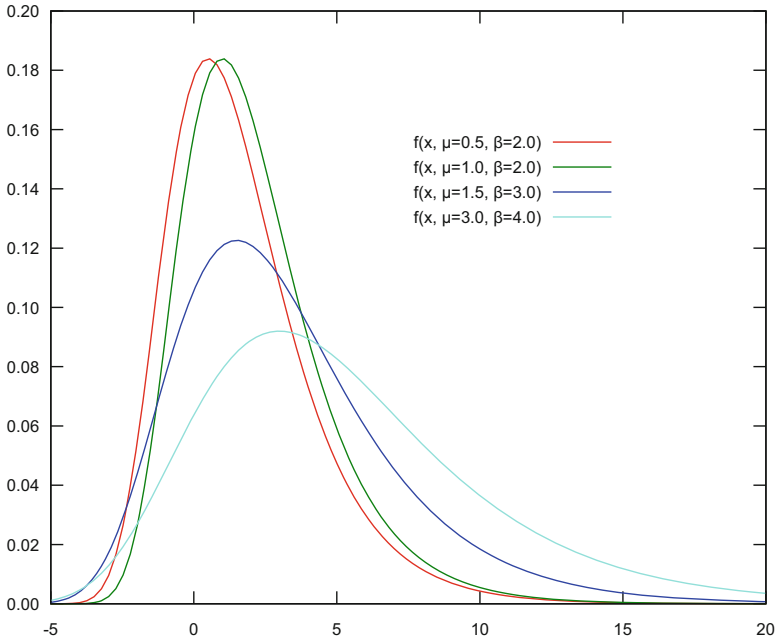


Fig. 11.1 Gumbel distribution

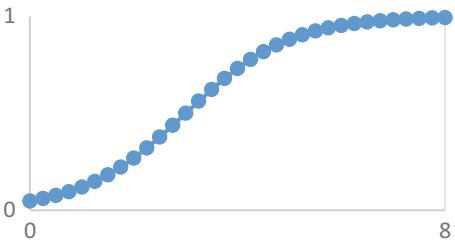


Fig. 11.2 Logistics utility distribution

If the uncertain outcomes of the decision analysis are also described by a Gumbel distribution, then the probability the decision analysis leads to outcomes equaling or exceeding all the stakeholder Gumbel distributed BATNA's has a logistics distribution (See Fig. 11.2).

Note that this target-oriented utility, the logistics distribution, is S-shaped indicating the individual is risk-prone for gambles involving lower-valued consequences and risk-averse for gambles involving higher-valued consequences.

To highlight the ease with which target-oriented utility maximization—based on the default alternative—can be implemented, it is useful to review the Pugh (1991) controlled convergence rule for product design. In applying the Pugh rule,

(a) A matrix is constructed with each row corresponding to each of the relevant design criteria and each column corresponds to a different alternative.
(b) A single reference alternative is specified but not scored on the criteria.
(c) Other alternatives get a score of one on a criteria if they are superior to the reference alternative on that criteria, score of minus one if they are inferior and a score of zero if they are comparable.
(d) Weights are assigned to each of the criteria.
(e) The score of each alternative is a weighted average.
(f) Before selecting the highest scoring alternative, the Pugh method focuses on developing hybrid alternatives with fewer minuses and more positives on the more important criteria. (Development of hybrid alternatives was a critical factor in the success of decision analysis at General Motors in the 1980's).

The Pugh rule has been criticized because it presumes additivity between problem criteria. Also its assignment of pluses and minus to each attribute ignores the difference between doing well on a criteria and doing extremely well.

But there is a situation in which the Pugh rule's conventions are theoretically defensible. Suppose each of the different criteria correspond to the performance of the alternative in a different scenario. Also suppose the list of possible scenarios are mutually exclusive and collectively exhaustive. Let the weight assigned to each criteria be the probability of the associated scenario occurring. Then the alternative's aggregate Pugh score is proportional to the probability of the alternative outperforming competing alternatives. Hence the Pugh Rule is just effect size utility maximization, i.e., maximization of a possible target-oriented utility function. This example highlights how utility maximization could be implemented as a version of an already widely used engineering rule.

So the default decision (or control treatment) can be used, like a yardstick, in evaluating the other alternatives. Changing which treatment is considered the control is equivalent to changing the yardstick (and thus the utility function). As a result, it is important to use the same default decision in evaluating all interventions. It can be argued, however, that using the default decision as the reference overly focuses on what is possible. Since Keeney (1992) criticized alternative-focused thinking in decision analysis because of its focus on what is possible, the next section considers goal-oriented approaches to specifying the benchmark random variable.

11.3 Goals

At the individual level, the nineteenth century economist Menger argued that an individual's utility function was determined by how well various biological needs were satisfied (Menger 1985). Greater utility was attached to satisfying more important needs (survival) than less important needs (shelter). The ranking of these needs was lexicographic, i.e., there was no value in satisfying a lower-ranked need until a higher-ranked need was satisfied. While Menger's theory was qualitative like Maslow's related hierarchy of needs, this paper considers a variant on that theory which is quantitative. If there are n needs, then for each $k \leq n$, define

1. P(k) as the probability that the k most important needs were met but the (k+1)st most important was not. (Note that it doesn't matter whether any needs less important than the (k+1)st are satisfied given that the (k+1)st need is not satisfied.)
2. P(S|k) as the probability the individual was considered successful at meeting their needs given that the k most important needs (but not the (k+1)st most important need) were met

Menger did not introduce the concept of P(S|k) in his theory which made his formalism qualitative. But with P(S|k), the probability of the individual's needs being met at an acceptable level becomes

$$P(S) = \sum\nolimits_{k=1} P(S|k)\, P(k)$$

This quantification presents Menger's theory in the form of an additive utility model.

We can further extend Menger's model to allow for non-lexicographic needs by defining X(k) to be the performance on need k with X(k) being the weighted sum of the value obtained from achieving a set of micro-needs associated with need k. Hence if k represents the need for food, the micro-needs might representing the need for different kinds of food with P(k) being the probability of X(k) exceeding the random benchmark. In this case, increasing the consumption of one kind of food decreases the need for consumption of a second kind of food. Micro-needs are not ordered lexicographically although needs are lexicographic.

Menger's theory focused on needs, i.e., on desires which are ultimately driven by an individual's biology. But professionals and professional codes of ethics provide another example. Most codes of ethics require that a professional act to meet their client's goals:

1. A project manager's task is to complete the project satisfying their client's cost, schedule and scope requirements.
2. A doctor's task is to restore the patient's health
3. An engineer's task is to design a product which, when manufactured, will be accepted and used by the intended customers for that project
4. An employee's task is to complete what their employer orders them to do.

In certain cases, these responsibilities are formalized as a contract with the professional's goal being to satisfy the terms of this contract. In this case, the goal of being a good professional drives satisfying the goals set by one's clients. Typically the contract provides incentives which make it in the professional's interest to serve the client. However many professional organizations do ask their members to take oaths to place duty to client and society over self-interest.

The conventional contract typically obligates the signatories either to deliver (or to accept delivery of) specified levels of products or services at specified levels of performance. This presumes that the individuals signing the contract can and will assume responsibility for controlling all of the uncertainties that might interfere with successful completion of the contract. Thus in tactical decision making (as opposed to strategic decision making), the decision maker is often viewed as responsible for managing uncertainties to achieve some goal.

But in professional (and even in non-professional work), there are many uncertainties about commodity prices, competitor behavior, etc. which are outside the control of both parties to the contract. This increases the riskiness of the contract. This has motivated the development of contingent contracts (Bazerman and Gillespie 1999; Brett 2007) which explicitly

(a) Identify certain future uncertainties which will only be observable when the contract is near completion
(b) Make the terms of the contract contingent upon the outcomes of these future uncertainties

Hence the goal of each signatory to the contract is contingent upon future uncertainties. As a result, the goal is a random variable and each action by a signatory is evaluated based on its probability of achieving this uncertain goal. In this case, it is natural to choose the random benchmark to be the contingent goal.

Note that even though target-oriented utility seems appropriate whenever there are contingent goals, there are important decisions that must be made before the individual agrees to a contingent contract. In particular, the individual must make a strategic decision about which clients to choose, which markets to enter, and which profession to join. If these strategic decisions are made with conventionally elicited utility functions, then conventional utility elicitation will have been made at one stage of decision making with target-oriented elicitation being made at a later stage.

Some goals are not contingent but are still uncertain. For example, an individual raised in poverty might have a strong drive to accumulate enough wealth to maximize the probability that neither they nor their children will ever know poverty again. In this case, each increment in wealth decreases the probability that sudden misfortune could eventually reduce the person or their children to poverty. The random benchmark is the uncertain amount by which misfortune might decrease the individual's wealth. This goal is an absolute goal, i.e., not defined relative to the performance of anyone else.

But some goals are relative, e.g., an individual who maximizes the probability of being able to afford whatever their role models (e.g., family, neighbors, colleagues, etc.) were able to perform. In this case, the random benchmark is whatever those role models would be able to afford.) Certain goals are inherently relative, e.g., an individual seeks the top prize in some competition and hence seeks to outperform all other contenders. In this case, the random benchmark is the quality of the uncertain performance of other contenders. Target-oriented utility is well-equipped to handle either absolute or relative goals.

Menger's formulation, while plausible for individuals, is especially plausible for an organization. Thus in the 1970's General Motors initially focused on research projects that looked at the long-run future of transportation. But as unexpected competition began to threaten the company's market share, research priorities shifted toward improving the company's marketing, manufacturing and design processes. And once the prospect of bankruptcy began to loom, priorities further shifted toward those projects most directly connected with short-term cash generation.

In most cases, the organization's highest priority need is the avoidance of bankruptcy. Borch (1968) considered an insurance company whose objective is to survive (avoid financial ruin) as long as possible. Avoiding ruin for an insurance company requires that its ability to pay claims exceed the uncertain amount of claims made against the company. Thus the insurance company needs to balance its investment in illiquid assets with higher expected return and more liquid assets with which to pay claims. Thus if T_t is the uncertain amount of claims filed at time t, then the insurance company needs to have liquid assets v_t at time t such that $v_t > T_t$ for all times t. This defines the utility function as one minus the probability of ruin. When claims arrive according to a Poisson process, the probability of ruin (Huzak et al. 2004) is described by the Pollaczek-Khinchine formula. In the case where claim sizes are exponential distribution, the probability of ruin has the form a exp.(-bx) if x is the starting wealth of the firm. Thus the firm's utility as a function of starting wealth x is 1-a exp.(-bx) which, with suitably rescaling, gives the widely used exponential utility. Note that managers who optimize an organizational utility—based on a goal of long-term organizational survival—may also optimize the probability of their having a personal legacy that others will remember.

'Mere' survival, i.e., the avoidance of bankruptcy may not fully encompass all the needs of the organization. For example, suppose the organization has the higher-level need to be recognized as the top organization in its segment, e.g.,

1. A firm might want to be recognized by Forbes as the most admired provider of information technology services
2. A university might wish to be listed by US News and World Report as the public university with the best business school in the world.

The articulation of such needs, which is related to the organizational vision, is typically set at the highest levels of the organization. In some companies, this vision has been specified using decision analysis with conventional utility functions.

This makes no specific statement about how the firm will design or deliver its products or services. However typically achievement of these higher-level goals is linked to how well its products or services meet the goals of stakeholders (i.e., customers, stockholders, the public, etc.) In public sector projects, considerable effort is often required to determine whether a stakeholder is a key stakeholder, i.e., is both

1. Impacted by the decisions the organization makes
2. Able to impact the organization based on its reaction to those decisions

Once the key stakeholders (e.g., customers who can afford a firm's product) have been identified, the goals of those stakeholders must be identified. This involves identifying the functional objectives, i.e., the functions which the stakeholders want accomplished and the performance objectives which qualify how well these functions must be accomplished. For example, the functional objective of a car is to transfer people and luggage from a starting point to a destination point. It also involves identifying performance objectives which focus on how much people and luggage the car transfers, the passenger comfort during the journey, and how long the journey requires. Quantitative marketing techniques can typically be applied to understand performance objectives once the functional objectives are identified. This provides understanding of what the organization must do to meet its goal. The next step is to translate this into organizational action.

Many hierarchical organizations employ management by objectives to set customer-tailored goals for each employee. Starting at the top of the organization, each level of the organization specifies goals for their subordinates in the next lower level. Goal-setting is based on a dialogue between manager—whose goals have been set by their leadership—and the manager's subordinates—who is more aware of what can realistically be delivered. Goals must be consistent, specific (unambiguous), measurable, time-related and focused on a result without specifying how the employee must achieve that result. As a result, attainment of the goal is a success event satisfying Howard's clairvoyant test.

Organizations use management by objectives because setting and tracking goals has been shown to be among the most valid and practical theories of employee motivation in organizational psychology (Miner 1980; Pinder 1984). They direct attention toward goal-relevant activities and away from goal-irrelevant activities (Rothkopf and Billington 1979) and can lead to the discovery and use of task-relevant knowledge and strategies (Wood and Locke 1990). In setting a goal, a manager must balance the fact that more difficult goals can motivate more effort

(Bandura and Cervone 1986) but can also increase the riskiness of strategies individuals use. Compared goals of equal difficulty and found that performance increases the more the individual believes the goal is achievable. Commitment to the goal (Seijts and Latham 2001) is also important in inducing performance when goals are difficult (Latham et al. 1994).

The effectiveness of goals in driving behavior is a major reason why they will continue to be so widespread. They provide an invaluable information source for target-oriented utility elicitation.

11.4 Screening

Value-focused thinking (Keeney 1992) highlights the importance of searching for new alternatives to make the choice set as rich as possible. In the absence of such an aggressive expansion of the choice set, conventional decision analysis can sometimes be reduced to 'choosing the best of a potentially mediocre lot.' (Gregory et al. 2012).

But in other problems, there is the opposite problem of having a choice set which is much too large. For example in purchase decisions, consumers often review possibly hundreds of products with possibly more than 50 product attributes and make screening decisions rapidly, sometimes in seconds (Payne et al. 1988, 1993). Because this screening process is designed to shortcut the more lengthy analysis used in the selection phase, it uses less information than selection. This can lead to certain paradoxes which have substantial importance in practice.

For example, in the automotive industry with more than 350 brands, the typical consumer only has two to four brands in their consideration set. It was observed that the Buick brand in 2008 had low sales relative to its competition even though it was.

1. Tied in 2008 with Lexus as the top-ranked automobile on a JD Power dependability study,
2. The top-ranked American car by Consumer Reports and.
3. Produced from the top-ranked US factory for quality.

Since Buick was comparable on other attributes, Buick's manufacturer did not understand how Buick sales could be so low when it did so well against such key competitors as Lexus on all the key attributes. The paradox was not resolved until the manufacturer collected survey data indicating that two thirds of California buyers wouldn't even consider GM cars in their choice. National studies showed that roughly half of all potential customers did not even spend the time to learn about Buick's superior performance on dependability and quality. This inability to even be considered reduced the value of automotive investments in reliability, quality, safety, ride and handling, comfort etc. and played an important role in the subsequent bankruptcy of two of the three major automotive manufacturers.

In the automotive industry, this understanding of the customer choice process was formalized in the so-called industry purchase funnel (Lancaster and Withey 2006) which postulated four stages in the consumer choice process:

1. Awareness: Customers become aware of a product
2. Consideration: Customers screen out products which are unacceptable on certain 'screening criteria'
3. Evaluation: Customers identify a smaller set of products which they will compare by gathering further information. They will also schedule test drives of vehicles in the consideration set.
4. Closing: Customers negotiate with the dealer on price and other details before determining the product they will ultimately buy.

The purchase funnel can also be viewed as a two-stage process of screening and selection. In screening, the customer becomes aware of products and screens out those that seem 'unworthy' of future consideration. In selection, the customer focuses on the remaining items in the choice set, gathers further information, and engages in a more systematic process to choose one of those options.

Screening is important with complex infrequently purchased durable products like automobiles. But it is also important with frequently purchased products, such as deodorants, where customers only consider a small fraction of the products available. Hauser and Wernerfelt (1990) found that only 10% of the products were in the final choice set. Evidence from other industries (Paulssen and Bagozzi 2005) similarly documents that customers initially screen the list of products to form a smaller list of products (the consideration set) and then focus attention on choosing products from the consideration set. Hauser (1978) found that 80% of the uncertainty about what a customer will choose can be explained from simply knowing their consideration set.

However instead of inferring the consideration set from choices, reliable data can often be collected by simply asking individuals to specify their consideration set. The reliability of responses to the question can be enhanced if the individual believes they will win a prize with the prize depending upon the quality of their answers (Ding 2007). In these cases, truthful responses are dominant. Such directly elicited considerations have enhanced a wide variety of new product forecasts (Hauser 1978).

There is also considerable research on the heuristics individual uses in screening their choice set. Heuristics, while less cognitively demanding, are often effective in real world choice environments because consumers can rely on market regularities, e.g., the fact that many market features are correlated. For example, automobiles with large engines tend to have good leg room, good trunk room and luxurious but also tend to be expensive and get lower gas mileage. As a result, these heuristics can be effective and reliable given conventional market regularities. But it is often relatively easy for experimental studies to create counter-examples—based on artificial environments without these correlations—in which these heuristics lead to absurd outcomes. Unfortunately when market circumstances change, these

heuristics, consistent with experimental studies, can fail. This is especially the case in rapidly changing competitive environments as well as in military environments.

Nonetheless using heuristics that worked well under current conditions can be risky when current conditions change. Target-oriented utility assessment applied to screening involves

1. Identifying and quantifying the heuristic a customer is using
2. Determine whether the heuristic is still appropriate for the problem at hand by
   (a) Identifying the logical limitations in that heuristic and discuss that with the decision maker
   (b) Gathering the individual's feedback either on possible rationales for this behavior and communicate the relevant insights from the behavioral decision theory literature to the individual
   (c) Identifying other ways in which the heuristic is exhibiting behavior which the decision maker does not consider reflective of their values
3. Discussing whether this decision rule applies to the existing application of interest
4. Producing an altered decision rule based on the heuristic which is consistent with expected utility theory and reflects how the decision maker would like the decision rule to make decisions for the applications in question.

For example, reliability-based design optimization is a commonly used approach for designing physical structures (bridges, planes, automobiles, etc.) which formalizes a heuristic where

1. Alternatives are first screened based on risk criteria
2. An alternative is selected from the screened criteria based on cost or other attributes.

Reliability-based design optimization formalizes this heuristic as a mathematical programming problem which maximizes some payoff function subject to a constraint on expected risk. So if there are two solutions

1. One which exactly meets the expected risk constraint and
2. The other which is negligibly more expensive and has no risk

reliability-based design optimization always chooses the first. But reliability-based design optimization is often applied in the design of structures, vehicles, planes etc. where the consequences of a risk can lead to an individual being injured or killed. American product liability law indicates that a product design is defective if an individual is injured and there is a solution (like the second solution) that could have avoided this risk at negligible cost. Hence the heuristic, in this case, could lead the firm to lose an expensive lawsuit.

The expected utility formulation of this problem is equivalent to maximizing the probability of

1. Cost being less than some uncertain threshold
2. The design not leading to injury or death

which is a target-oriented utility. Because of the continuity axiom of decision theory (von Neumann and Morgenstern 1944), this formulation will choose the second solution if its nonzero cost is small enough and will choose the first solution otherwise.

These undesirable features of screening heuristics should be explained to the decision maker to determine whether these considerations persuade them to adopt the more rigorous formulation made possible by target-oriented utility. As we now show, the structure of the heuristics people use in practice often makes this transition from heuristics to the rigorous formulation especially easy with target oriented utility. Commonly studied heuristics customers use in making decisions are

1. Disjunctive—a product is considered if it has at least one feature above a threshold.
2. Conjunctive—a product is used if all features are above their threshold. The conjunctive rule favors a 'balanced' product which is acceptable on all attributes to an 'unbalanced' product which is outstanding on most attributes but inferior on a few attributes.
3. Subset conjunctive—a profile must have at least some of its features above a threshold
4. Disjunction of conjunctions: a product must have at least one conjunction to be considered
5. Elimination by aspects

The marketing literature illustrates how the heuristics used by an individual can be estimated from the individual's observed behavior. For example, Gillbride and Allenby (2004a, b) specified a model in which individuals first screen products and then use an additive choice rule to select among those products. They then used a Bayesian approach to estimate the thresholds individuals used. (They found that 92% of individuals used a screening rule even when the choice set was modest.) Andrews and Srinivasan (1994), Chinag et al. (1999), Erdem and Swait (2004), Swait and Ben-Akiva (1987) used choice set explosion with maximum likelihood techniques to estimate thresholds. They assumed that individuals first made a choice among all possible choice sets (for n attributes, there will be 2^n -1 possible consideration sets.) Given a choice set, the individuals then pick an alternative using logit.

These particular heuristics, once identified, can be represented by a fault tree (see Fig. 11.3) or alternatively a reliability block diagram (also called a dependence diagram.) The fault tree representation begins with a highest-order success event. This success event is then logically related to the occurrence of some collection of higher-order success events. The occurrence of each of these higher-order success events is then related to the occurrence of some lower-order success events, etc.

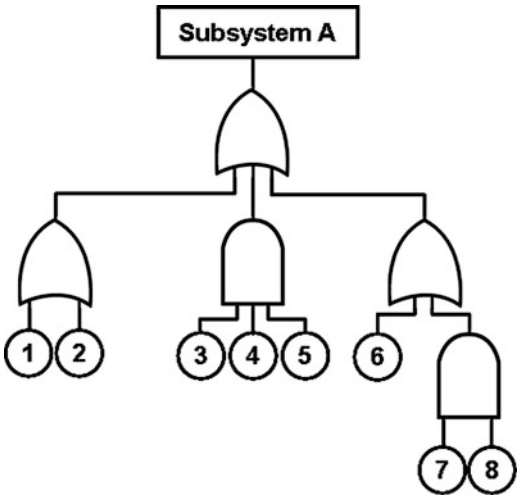


Fig. 11.3 A fault tree diagram

The fault tree continues decomposing success events into sub-events until it reaches a level of granularity (e.g., the part level) where assessing the probability of part failure is straightforward.

Bordley and Kirkwood (2004) showed that the normative multilinear multiattribute utility model is, in fact, isomorphic to a fault tree representation—when all sub-events are independent and each sub-event corresponds to an attribute. Thus in Fig. 11.1, a conjunction node indicates that both nodes 7 and 8 must be false for the conjunction node to be false. This is then followed by a disjunction node which indicates that either node 6 or the conjunction node must be false for the disjunction node must be false. This is then followed by a disjunction node which indicates the subsystem has failed if either the disjunction node associated with nodes 1 and 2, the disjunction node associated with nodes 3, 4 and 5, or the disjunction node associated with nodes 6, 7 and 8 are false. The fault tree (or reliability block diagram) allows the representation of all possible combinations of conjunctions and disjunctions in order to determine the state of the subsystem.

Tsetlin and Winkler (2006) showed how this sub-event independence assumption could be relaxed using copulas. Because of this isomorphism with the multiattribute utility formulation, heuristics expressed as fault trees are readily transformed into utility models. Once we have found the choice model the individual appears to be implicitly using, we can then use the previous process to modify the empirical decision model to reflect a model which the individual would wish to be applied in making future decisions.

For example, some of the heuristics assume that an alternative is discarded if its score exceeds some threshold, regardless of whether it slightly exceeds that threshold or substantially exceeds that threshold. Since measurement error could easily cause an outcome which is above the threshold to appear below the threshold (or vice versa), we need to address this with the individual and considering making the probability of discarding an alternative depend on the degree to which it violates the threshold.

11.5 Expectations

One of the most accepted findings from behavioral decision theory is the reference point effect. Individuals are risk-averse for prospects offering improvements above the reference point and risk-averse to prospects offering losses below the reference point. A value function can be developed whose curvature describes differences in individual attitudes toward risk above and below the reference point. This, of course, is similar to the behavior of an individual maximizing the probability of trying to exceed an uncertain benchmark where the reference point is the modal value of the benchmark.

Samuelson and Zeckhauser (1988) proposed interpreting the reference point as the status quo. But Locke and Lackham (2006)'s review of their own seminal work on goal setting theory highlighted that the goal in their theory was empirically very similar to the reference point in prospect theory. Heath et al. (1999) likewise found empirical evidence establishing a correspondence between prospect theory's reference point and goals for individuals with explicit goals. Noted the correspondence between falling short of a goal and dissatisfaction and exceeding a goal and satisfaction. More recently Koszegi and Rabin (2006, 2007, 2009) proposed interpreting the reference point as expectations. Since individual expectations are often easy to manipulate, this interpretation of the reference point as expectations provides an explanation for why the reference point can be manipulated by changing how experimental questions are worded.

To develop a normative model explaining the role of expectations, Viscusi (1989) considered a model in which individuals, instead of taking a gamble at face value, developed their own beliefs about what the gamble would pay off based on their prior experience with the payoffs of gambles. This is especially plausible in marketing settings where customers typically view the claimed (or advertised) performance of different brands with some suspicion (Bordley and Hazen 1991). Bordley (1992) expanded Viscusi's model to the context of multiple alternatives and showed that it allowed for:

1. Context-dependence, i.e., the goodness of a gamble depended upon the value of the other alternatives in the choice set. This kind of effect is often seen in commercial settings where many retailers want to be able to display 'halo' vehicles (e.g., sporty vehicles) in their showrooms. Retailers know that most customers have no interest in buying halo vehicles but are nonetheless attracted to brands that offer a halo vehicle. To explain the impact of halo vehicles, note that a company which offers a successful sporty vehicle is advertising its capability to create such a vehicle which, in turn, reflects positively on the quality of the other vehicles it produces.
2. Preference reversals: where adding an alternative to the choice set reversed how an individual rated the relative desirability of two different alternatives in the choice set

In this model, apparently non-normative behavior reflected individuals using the existence of certain alternatives in the choice set as information about the desirability of other alternatives. However these models are only normative because they presume certain market regularities, e.g., the fact that many market features are correlated, to make inferences about other alternatives in the choice set. Since these presumptions are often implicit, it is valuable to articulate them to determine whether they are still appropriate. Once this behavior and its apparent rationale is described to the decision maker, it can potentially be modified.

Interpreting the benchmark as individual expectations is consistent with customer satisfaction research (Bordley 2001) which defines customer satisfaction with a service as the gap or difference between what the service provides and customer expectations. When there are many customers, this probability will approximate the fraction of customers for whom the product exceeds expectations. Many groups within the firm often focus on spending their budgets to maximize the number of satisfied customers at a given price point. (The decision of how much money to allocate to various customer groups has sometimes been made previously at the strategic level.) These groups will therefore maximize a target-oriented utility defined as the probability that the service outperforms expectation (which is also an effect size measure of the gap.)

Reference points which are interpretable as goals (and not expectations) can also be unstable. Suppose the goals are SMART goals and time-bound, i.e., there is time in the future when the goal will either have been met or unmet. At this point, the probability of achieving the goal (and the utility function goes to zero or one.) Upon achieving or failing to achieve a goal, Simon's theory of bounded rationality then advised the individual to

1. Become satisfied and make no further changes to their life plans until external events disrupt their equilibrium
2. Set new goals

Organizations generally view an employee's successful (or unsuccessful) completion of a goal as the stimulus for the setting of another goal. As a result, organizations treats goal setting as a cyclical discrepancy-creating process (Bandura 1997). If people attain the goal they have been pursuing, they generally set a harder goal for themselves. The high-performance cycle explains how high goals lead to high performance, which in turn leads to rewards, such as recognition and promotion. Rewards result in high satisfaction as well as high self-efficacy regarding perceived ability to meet future challenges through the setting of even higher goals. Of course, the manager must set these higher-level goals to be intertemporally consistent. If intertemporal inconsistency is observed, then the goal-setter needs to be challenged in order to discover whether there is an underlying goal which resolves the apparent inconsistency. (As a trivial example, if an individual flies from Boston to Los Angeles on Sunday and then flies from Los Angeles to Boston on Tuesday, this apparent intertemporal inconsistency is resolved by noting that the individual's goal was be in Los Angeles on Monday while being in Boston prior to Sunday and after Tuesday.)

So goals (as well as expectations) vary systematically over time. To model dynamic expectations, define the expectations about what is required to be happy at time t+1 as a function of

- A random component, T*
- A time-dependent component, f_t.

For example, suppose that expectations are the sum of T* and f_t. with f_t. equalling the value achieved in the previous time period, v_{t-1}. Then the probability of the value achieved in the current period, v_t, exceeding the uncertain threshold, T, is the probability of the increase in value from the past period exceeding the expectations-independent component, i.e., the probability that v_t - v_{t-1} exceeds T*. In this case, the success event is achieved at time t if the level of improvement, v_t - v_{t-1}, is acceptable. But if it is achieved, the individual now has a new goal to achieve an acceptable level of improvement in v_{t+1} - v_t at time t+1. Thus expectations are dynamically updated at each point in time. This guarantees the familiar economic property of non-satiation where the economic agent constantly seeks more.

Since f is essentially a forecast based on previously experienced values, standard time-series methods could be used to provide a more sophisticated forecast. For example, exponential smoothing could be used to write the time-dependent component as a discounted weighted average of value in past periods, e.g.,

$$f_t = \mathrm{w}\, v_{t-1} + \mathrm{w}^2\, v_{t-2} + \mathrm{w}^3\, v_{t-3} + \mathrm{w}^4\, v_{t-4} + \ldots.$$

Simon had argued that if high value outcomes were easy to achieve, then the individual would set more aggressive goals. In contrast, if they were harder to achieve, individuals would set less aggressive goals. Simon's property will automatically be satisfied by dynamic updating of expectations. More complex specifications of f are possible (Bordley 1986). Thus the rate at which value increases, dv/dt, might be proportional to the amount of value left unachieved, (1-v). This implies that $v_t = 1$- exp.(-kt) for some constant k.

Note that these dynamic expectations do lead to a stable utility function defined over the rate at which value changes with time.

11.6 Conclusions

Target-oriented utility assessment is an approach for eliciting utility functions based on identifying random benchmarks that are often important in formulating the decision problem. This chapter has highlighted several advantages:

1. In many real problems, there are often easily identifiable benchmarks. For example, firms strive to outperform their closest competitors while also satisfying certain cash flow needs. Individuals strive to achieve goals and satisfy needs. Decision analysis is often expected to lead to decisions that outperform decisions

that would have been made without decision analysis. The existence of these benchmarks often makes the process of quantifying the utility function more straightforward and 'objective.'
2. Certain benchmarks are often familiar to the decision makers. This can make them more comfortable with a target-oriented utility function based on these benchmarks.
3. It is often easy to get other individuals to agree on the appropriateness of these benchmarks. This can help the decision maker justify the target-oriented utility to their peers and decrease the chances their decision will be overturned should they be replaced with another person.

There are added advantages to estimating benchmarks. Specifically the optimization of a conventional utility function typically is optimization subject to various constraints. Thus in selecting an airplane design, Boeing's value function depends on customer convenience, speed, fuel efficiency and cost. Because of uncertainties in these variables, a utility function is defined over the value function. This expected utility is optimized subject to constraints screening out plane designs that

1. Require airports to dramatically extend their runways.
2. Are cost prohibitive because of the extensive use of exotic materials
3. Require technological breakthroughs that seem unlikely
4. Are strongly opposed by at least one key stakeholder

Since both the costs of future materials and technological breakthroughs are uncertain, Boeing cannot be sure about whether it is excluding some items from the choice set that are feasible. And conversely Boeing cannot be sure about whether certain options included in the choice set will eventually be found to be infeasible. Hence the constraint functions involve uncertainties. As a result, the constrained optimization problem must be formulated as the optimization of the product of

1. The expected utility of an airplane design given the constraints are satisfied
2. The probability of the constraints being satisfied

Using benchmarks, the expected utility can be written as the probability of the value function exceeding the uncertain benchmarks. If we define an optimal solution as a solution whose value exceeds the uncertain benchmarks and a feasible solution as a solution which satisfies the constraints, then the constrained utility maximization problem is simply the selection of a design which maximizes the joint probability of being both feasible and optimal. Thus target-oriented utility integrates the identification of the feasible choice set with the problem of selecting an optimal solution from that choice set. By making the constraints part of the objective function, target-oriented utility can reduce the computational effort require to solve utility maximization problems.

This emphasis on addressing the larger context in which utility functions are used is reflected in target-oriented utility elicitation's focus on

1. The benchmark alternative to the outcome of decision analysis
2. The larger goals which the decision maker feels their efforts should support.

3. The screening of alternatives to identify the choice set
4. The expectations based on prior experience with other choice sets

Thus target-oriented utility elicitation can be an invaluable part of the decision analyst's tool kit.

References

Anderson E, Simester D (2011) A step by step guide to smart business experiments. Harv Bus Rev 89:3–12

Andrews R, Srinivasan T (1994) Studying consideration effects in empirical choice models using scanner panel data. J Mark Res 32:30–41

Bandura A (1997) Self-efficacy: the exercise of control. W. H. Freeman, New York, CA

Bandura A, Cervone D (1986) Differential engagement of self-reactive influences in cognitive motivation. Organ Behav Hum Decis Process 38:92–113

Bazerman M, Gillespie J (1999) Betting on the future: the virtues of contingent contracts. Harv Bus Rev 77(5):155–160

Billingsley P (1995) Probability and measure. John Wiley & Sons, New York

Bordley R (2001) Integrating gap analysis in service research. J Serv Res 3(4):300–309

Bordley R (1992) An intransitive expectations-based bayesian variant of prospect theory. J Risk Uncertain 5(2):127–144

Bordley R, Kirkwood C (2004) Multiattribute preference analysis with performance targets. Oper Res 52(6):823–835

Bordley R, Hazen G (1991) SSB and weighted linear utility as expected utility with suspicion. Manag Sci 38(4):396–408

Bordley T (1986) Satiation and habit persistence (or the dieter's dilemma). J Econ Theory 38(1):178–184

Borch K (1968) Decision rules based on the probability of ruin. Oxf Econ Pap 20(1):1–30

Brett J (2007) Negotiating globally: how t negotiate deals, resolve disputes and make decisions across cultural boundaries. John Wiley & Sons, New York, p 74

Brodersen K, Galluser F, Koehler J, Remy N, Scott S (2015) Inferring causal impact using bayesian structural time-series models. Ann Appl Stat 9(1):247–274

Chinag J, Chib S, Narasimhan C (1999) Markov chain Monte Carlo and models of consideration set and parameter heterogeneity. J Econ 89:223–248

Davenport T (2009) How to design smart business experiments. Harv Bus Rev 87(2):68–76

Ding M (2007) An incentive-aligned mechanism for conjoint analysis. J Mark Res 54:214–223

Ding M, Rajdeep G, Liechty J (2005) Incentive-aligned conjoint analysis. J Mark Res 42:67–82

Erdem T, Swait J (2004) Brand credibility, brand consideration and choice. J Consum Res 31: 191–198

Gillbride T, Allenby G (2004a) A choice model with conjunctive, disjunctive and compensatory screening rules. Mark Sci 23(3):391–406

Gillbride T, Allenby G (2004b) Estimating heterogeneous EBA and economic screen rule choice models. Mark Sci 25:494–509

Gregory R, Failing L, Harstone M, Long G, McDaniels T, Ohlson D (2012) Structured decision making. John Wiley & Sons, Chichester

Grissom J, Kim J (2005) Effect sizes for research: a broad practical approach. Lawrence Erlbaum, Mahwah, NJ

Hauser J (1978) Testing the accuracy, usefulness and significance of probabilistic models: an information theoretic approach. Oper Res 26(3):406–421

Hauser J, Wernerfelt B (1990) An evaluation cost model of consideration sets. J Consum Res 16:398–403
Heath C, Larrick R, Wu G (1999) Goals as reference points. Cogn Psychol 38:79–109
Howard RA (1988) Decision analysis: practice and promises. Manag Sci 34(6):679–695
Huzak M, Perman M, Šikić H, Vondraček Z (2004) Ruin probabilities for competing claim processes. J Appl Probab 41(3):679–690
Keeney R (1992) Value-focused thinking. Harvard University Press, Cambridge
Koszegi B, Rabin M (2006) A model of reference-dependent priors. Q J Econ 121:1135–1165
Koszegi B, Rabin M (2007) Reference-dependent Priors. Am Econ Rev 97:1047–1073
Koszegi B, Rabin M (2009) Reference-dependent consumption plans. Am Econ Rev 99:909–936
Lancaster G, Withey F (2006) CIM Coursebook 06/07 marketing fundamentals. Butterworth-Heinemann, Oxford
Latham GP, Winters D, Locke E (1994) Cognitive and motivational effects of participation: a mediator study. J Organ Behav 15:49–63
Locke E, Latham G (2006) New directions in goal setting theory. Curr Dir Psychol Sci 15(5): 265–268
Menger C (1985) Investigations into the method of the social sciences with special reference to economics (ed. Louis Schneider, trans. Francis J. Nock). New York University Press, New York, NY
Miner J (1980) Theory of organizational behavior. Dryden, Hinsdale, IL
Pinder C (1984) Work motivation. Scott Foresman, Glenview, IL
Parnell GS, Bresnick TA, Tani SN, Johnson ER (2013) Handbook of decision analysis. John Wiley & Sons, New York, NY
Paulssen M, Bagozzi RP (2005) A self-regulatory model of consideration set formation. Psychol Mark 22(10):785–812
Payne J, Bettman J, Johnson R (1988) Adaptive strategy selection in decision making. J Exp Psychol Learn Mem Cogn 14:534–552
Payne J, Bettman J, Johnson R (1993) The adaptive decision maker. Cambridge University Press, Cambridge
Pugh S (1991) Total design: integrated methods for successful product engineering. Addison-Wesley, London
Rothkopf E, Billington M (1979) Goal-guided learning from text: inferring a descriptive processing model from inspection times and eye movements. J Educ Psychol 71(3):310–327
Samuelson W, Zeckhauser R (1988) Status quo bias in decision making. J Risk Uncertain 1:7–59
Seijts G, Latham G (2001) The effect of distal learning, outcomes and proximal goals on a moderately complex task. J Organ Behav 22:291–302
Swait J, Ben-Akiva M (1987) Incorporating random constraints in discrete models of choice set generation. Transp Res 21(B):92–102
Spetzler C, Stael von Holstein C (1975) Probability encoding in decision analysis. Manag Sci 22:340–358
Thomke S, Manzi J (2014) The discipline of business experimentation. Harv Bus Rev 92(12):70–79
Tsetlin I, Winkler R (2006) On equivalent target-oriented formulations for multiattribute utility. Decis Anal 3(2):94–99
Viscusi W (1989) Prospective reference theory: toward an explanation of the paradoxes. J Risk Uncertain 2:235–264
Von Neumann J, Morgenstern O (1944) Theory of games and economic behavior. Princeton University Press, Princeton, NJ
Wood R, Locke E (1990) Goal setting and strategy effects on complex tasks. In: Staw B, Cummings L (eds) Research in organizational behavior, vol 12. JAI, Connecticut

Chapter 12
Multiattribute Value Elicitation

Alec Morton

Abstract Multiattribute Value Theory (MAVT) methods are perhaps the most intuitive multicriteria methods, and have the most theoretically well-understood basis. They are employ a divide-and-conquer modelling strategy in which the value of an option is conceptualised as a function (typically the sum) of the scores associated with the performance of the option on different attributes. This chapter outlines the concept of preferential independence, which has a critical underpinning role of elicitation within the MAVT paradigm. I also present MAVT elicitation in the context of the overall Decision Analysis process, comprising three broad stages: Designing and Planning; Structuring the Model; and Analysing the Model. I outline some of the main practical methods for arriving at the partial values and weighting them to arrive at an overall value score, including both traditional methods relying on cardinal assessment, and the MACBETH approach which uses qualitative difference judgements. A running example of a house choice problem is used to illustrate the different elicitation approaches.

12.1 Background

The Multi-Attribute Value Theory (MAVT) approach, and in particular the additive model, is perhaps the most intuitive of all Multi-Criteria Decision Analysis (MCDA) methods. The decision aiding procedure suggested by MAVT is to line up the options, compare them according to a common set of criteria, assign scores to each option according to their performance on each criterion, weight these criteria and calculate an overall score for each option. The computations involved in applying MAVT are relatively straightforward compared to the methods of the outranking school—see Chap. 14 in this book (Dias and Mousseau 2018), and hence the method is transparent and easily understood. One of the insights of this MAVT paradigm

A. Morton (✉)
University of Strathclyde, 16 Richmond St, Glasgow, G1 1XQ, UK
e-mail: alec.morton@strath.ac.uk

L.C. Dias et al. (eds.), *Elicitation*, International Series in Operations Research & Management Science 261, https://doi.org/10.1007/978-3-319-65052-4_12

is that this seemingly simple procedure, involving nothing more than elementary arithmetic, actually requires quite a high level of conceptual sophistication to use well and appropriately.

The need for conceptual sophistication arises when one attempts to specify formally the meaning of the scores and weights in the procedure of the previous paragraph. The meaning of a probability, by contrast, is relatively clear, in the following sense. Although the exact interpretation of a probability statement depends on one's preferred axiomatics (French 1986; French and Ríos Insua 2000), probabilities are ultimately rooted in the procedure of counting which is a natural first step on the path to quantification. If an assessor is well-calibrated, of the class of events she assesses as having probability 50%, half will be realised, and half will not.

By contrast, value is not rooted in counting, but in preferring. However, whereas counting establishes an association between a set of things and a number, preferring merely establishes a relationship between two things: one thing is better than, more attractive than, or more desirable than, another. From such a binary relation, it is easy to see how to establish a ranking of objects. However, how might one go about associating *numbers* to options according to their criterion-wise performance in a principled way?

The central concept of MAVT is that as well as possessing an idea of *preference*, we also possess an idea of *strength of preference* (Dyer and Sarin 1979; Köbberling 2006). Thus, when thirsty on a hot day, I may have a slight preference for iced tea over iced coffee, but a strong preference for an iced drink over no iced drink. The difference between the scores I give to iced tea and iced coffee should therefore be relatively small, but the difference between these scores and no iced drink should relatively large. However, unlike preferences, which can be observed by an outside party who studies the elicitee's choice behaviour (I offer you a menu consisting of iced tea and coffee and see which, if either, you choose), strengths of preference are not observable. Nevertheless, the concept seems to be one which is intuitive and natural to most of us from casual introspection and ordinary discourse.

An alternative way to assign numbers to multiattributed options is the Multi-Attribute Utility Theory (MAUT) approach—dealt with in Chap. 10 of this book (González-Ortega et al. 2018). MAUT, like MAVT, provides a framework for deriving scores and weights. However, the interpretation of the scores and weights in MAUT does not use a strength of preference concept—rather it uses an approach based on equivalent gambles. MAUT is necessary if we are dealing with uncertain events, for instance in a multiattribute decision tree. However, while the MAUT mode of questioning can be appropriate in many settings, it presupposes a facility with probabilistic thinking which many people do not have, and involves asking questions which are often experienced as confusing and irrelevant.

In this chapter, I do the following. I begin with a discussion of the concept of preferential independence which is a foundational concept in the use of scoring and weighting methods based on MAVT. The main section presents MAVT elicitation in the context of the decision analysis process, from establishing aims through to sensitivity analysis and stress testing of the model. To assist readers who may be interested in using these procedures, I also provide some "troubleshooting" hints

and tips. I conclude with some suggestions for future prospects for MAVT methods. The interested reader is referred for comparison to other textbooks which deal with similar material such as Goodwin and Wright (2014) and Howard and Abbas (2016) as well as the seminal text of von Winterfeldt and Edwards (1986).

12.2 Preferential Independence: A Foundational Concept of Multiattribute Value Theory

A natural starting point is to ask the question: under what circumstances can MAVT be used? As it happens there is a very clear and mathematically well-specified answer to this question (Krantz et al. 1971; French 1986). To explain this answer I introduce the idea of a representation theorem. Representation theorems connect qualitative properties of preferences with functions which represent these preferences. (A function is said to represent preferences if it assigns a higher number to a more preferred object). Representation theorems have the following generic two-part form. The main action revolves around the relation $\succsim$, read "is weakly preferred to" or "is at least as good as".

12.2.1 Generic Representation Theorem

1. (Sufficiency) If the relation $\succsim$ has such and such properties, then there exists a real valued function $v(\bullet)$ of such and such a form such that $a \succsim b$ if and only if $v(a) \geq v(b)$.
2. (Necessity) If there exists a real valued function $v(\bullet)$ of such and such a form such that $a \succsim b$ if and only if $v(a) \geq v(b)$, then the relation $\succsim$ has such and such properties.

Note the differing role of these two parts of the theorem: the sufficiency part tells us that if an elicitee has a preference relation with certain characteristics, then there exists a real value function, whereas the necessity part tells us the opposite. (In general the sufficiency part is harder to prove than the necessity part.)

An example of a representation theorem (Krantz et al. 1971) is the following theorem which guarantees the existence of a general function.

12.2.2 Representation Theorem for the Existence of a Representing Function

1. (Sufficiency). If $\succsim$ is complete and transitive, then there exists a $v(\bullet)$ which represents $\succsim$.
2. (Necessity). If there exists a $v(\bullet)$ which represents $\succsim$, then $\succsim$ is complete and transitive.

What this tells if that if an elicitee has preferences which are non-transitive—she tells us she prefers tea to coffee and coffee to hot chocolate, and hot chocolate to tea, there is no representing function for her preferences. A moment's reflection shows why this is so: it would require finding three numbers x, y and z such that $x>y$, $y>z$, and $z>x$, which is plainly impossible.

A more interesting and subtle question is under what circumstances can scoring and weighting be used to arrive at an evaluation of options. Scoring and weighting implicitly involves the use of an additive value function $v(a) = \sum_j w_j v_j(a)$, where v_j is a scoring function which assigns scores for each criterion j to each option a and w_j is the weight of criterion j. Is there a representation theorem which tells us when this value function can be used? As it happens, there are several such representation functions. One useful illustrative example is the following.

12.2.3 Representation Theorem for the Existence of an Additive Representing Function

Let $\succsim$ be a preference ordering on a set of biattributed options with well-defined partial preferences $\succsim_i$ for i=1 and 2. Given certain technical assumptions, the Reidemeister condition is necessary and sufficient for the existence of a representing additive value function.

As this chapter aims for informality, I do not propose to explain this Theorem in detail here. In particular I ignore the role of technical conditions such as solvability and the Archimedean axiom in proving the result. However, the Reidemeister condition is insightful and it is worth taking some time to present in detail. To understand the condition, consider Fig. 12.1 (the illustration is based on that in Belton and Stewart 2002).

Figure 12.1 shows points in a biattribute space, with dimensions x and y. For example, in choosing a house, x and y could be *square footage* and (the negative of) *purchase price*. A and A'; B and B' and C and C' are pairs of points in this space (each pair representing a larger, more expensive house, and a smaller, cheaper house) between which the elicitee is indifferent, i.e. prefers neither one nor the other.

The Reidemeister condition is a condition on the elicitee's preferences. An elicitee's preferences obey this condition if, whenever she is indifferent between A and A'; B and B' and C and C' respectively, she is also indifferent between D and D'. To see why this condition is sufficient for the existence of a representing additive value function is hard: the proof involves using the condition iteratively to construct a grid of points which have the interpretation of a value function of the additive form. But to see the necessity is easy. Consider Fig. 12.2. If the elicitee's preferences are represented by an additive value function, then the formulae for the values of these indifferences can be written as shown on the grid. The reader can verify that by adding the equations corresponding to the indifferences between B and B' and C and C' respectively, and subtracting the equation corresponding to A and A', the result is the following:

$$w_x v_x (x_2) + w_y v_y (y_2) = w_x v_x (x_2 - m_2) + w_y v_y (y_2 + n_2)$$

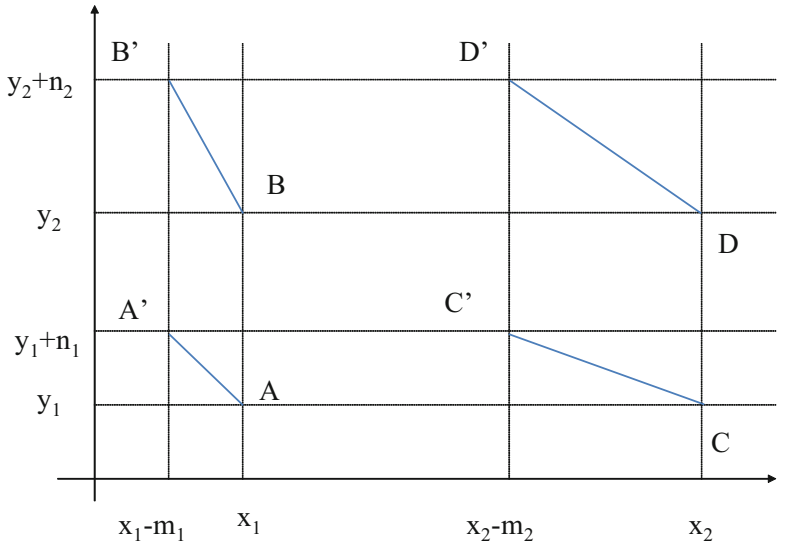


Fig. 12.1 Four pairs of points in a biattribute space, illustrating the Reidemeister condition

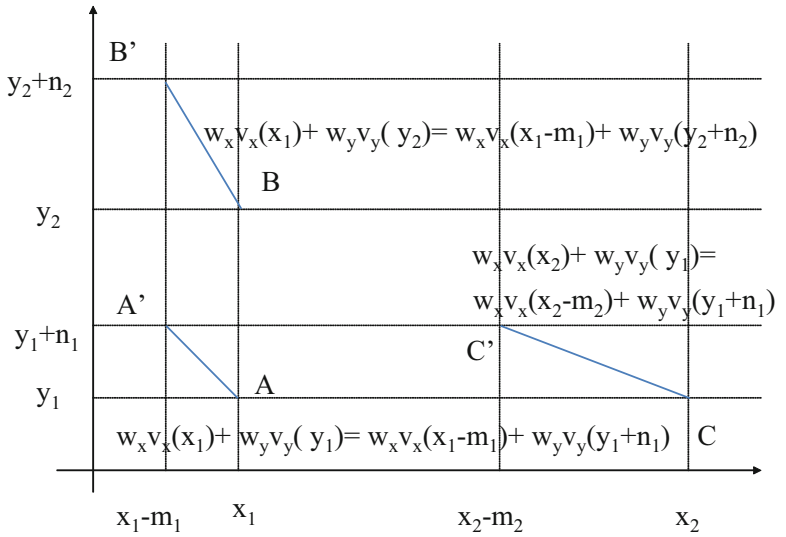


Fig. 12.2 Value functions associated with the indifferences between A and A', B and B', and C and C'

But this equation expresses nothing other than the idea that D is indifferent to D'. Hence any elicitee whose preferences are represented by an additive function must obey the Reidemeister condition.

It is not always or necessarily the case that the Reidemeister condition holds. In the case of buying a house, I may feel that the value of difference in space m_2 depends on the price which I am prepared to pay for the house. When I pay a lower price for the house, I can use the space to host fabulous parties, and hence the additional space has some value to me. But when I pay a higher price, I have no spare money for entertaining and the additional space just means that I have to spend more time cleaning. Hence, it does not make sense to give "points" to the additional space irrespective of the financial purchase price of the house.

If the Reidemeister condition or its equivalents fail to hold, that does not necessarily mean that all is lost. There are models which represent situations where there are interactions between criteria. The simplest and most intuitive example is that the "Quality Adjusted Life Year", or QALY, which has found widespread use in health economics as a measure of health benefit associated with a life extension or enhancement (for axiomatics, see Pliskin et al. 1980; Miyamoto et al. 1998). At its simplest the key idea of the QALY is that an individual's life can be considered as characterised in two dimensions: length and quality of life. Figure 12.3 illustrates two individuals, one of whom enjoys a short healthy life and the other of whom experiences a long miserable life.

For health gains, it makes no sense to calculate the value of a health gain as a weighted sum of duration and quality of life. To see why not, consider the extreme case of a life extension of zero (or infinitesimal) duration. Such a life extension

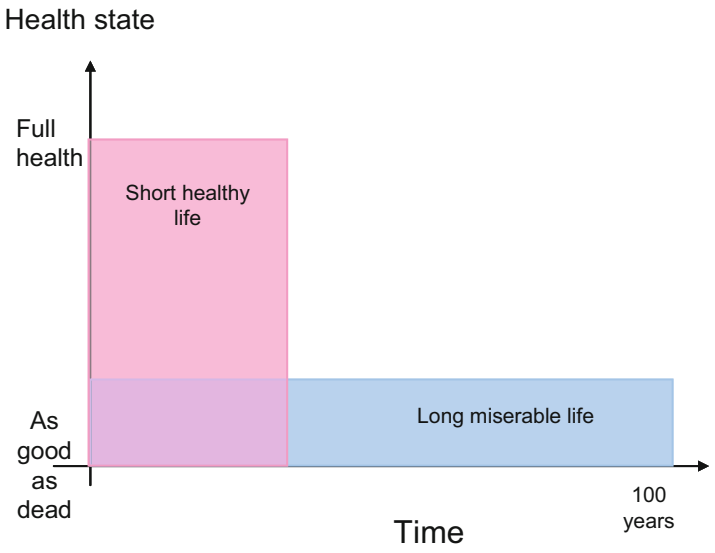


Fig. 12.3 Two possible lifecourses

clearly has no value, no matter how good the health state. This is not compatible with an additive model where the contribution of a set number of years of life to overall value is fixed, independently of the number of years lived in that health state. For this reason, QALYs are calculated as the length of life *multiplied* by a factor representing the quality of life (this can be visualised as the area of the rectangles in Fig. 12.3). Indeed, one popular way to elicit the value of a health state is to ask a so-called time tradeoff question, where a number is associated with a health state h (being blind, for example) by asking the elicitee for a number of years n such that they would be indifferent between n years in state h and 1 year in full health (Drummond et al. 2015).

12.3 The Decision Analysis Process

Having sketched the foundational concept of preferential independence, I now turn to the question of how to actually elicit scores and weights. Attempting to elicit scores and weights in the context of a poorly specified decision problem is a hopeless undertaking: before elicitation can take place, the problem context, and the basic elements of the model must be clearly specified and understood by all relevant parties in the elicitation. Accordingly I will structure this chapter through a map of the decision analysis process (see Fig. 12.4).

Design and planning	
Step 1.	Establish the aims of the analysis
Step 2.	Identify decision makers, stakeholders, and persons with relevant expertise
Step 3.	Design the intervention
Structuring the model	
Step 4.	Identify the options
Step 5.	Identify the criteria
Step 6.	Score the options on the criteria
Step 7.	Weight the criteria
Analysing the model	
Step 8.	Compute overall ranking
Step 9.	Conduct sensitivity analysis

Fig. 12.4 Schematic of the decision analysis process

12.3.1 Design and Planning

12.3.1.1 Step 1. Establish the Aims of the Analysis

A sensible starting point is to identify the objectives of the decision. For example, the objectives could be: to grow the organisation (in terms of revenues, reputation, market share, or profitability); to contribute to social welfare (e.g. through the provision of healthcare or recreation facilities); to contribute to equity objectives (for example health equity, equity in income distribution); to contribute to some other stated policy objective (such as reducing error in tax collection or benefits payment); or to help an organisation (e.g. a government agency or social enterprise) fulfil its mission.

The analysis may be intended to support different problem statements or *problématiques* (Roy 1985):

- Single choice (choose one option from n options)—for example choosing a site for a new airport.
- Multiple choice (choose k options from n options)—for example members of a team or a board.
- Budget allocation (choose options subject to a budget constraint of B)—for example determining a portfolio of R&D projects, or military equipment for purchase.
- Development of a priority ordering—for example ranking applicants for a scholarship in terms of their merit.
- Accepting or rejecting an option (for example, deciding whether a new drug can be provided by the national healthcare system).

Articulating both aims and the *problématique* is often a useful starting point for analysis.

12.3.1.2 Step 2. Identify Decision Makers, Stakeholders, and Persons with Relevant Expertise

It is important to identify early on both the decision makers, stakeholders, who may be individuals, organisational units, or organisations, and persons with relevant expertise. A decision maker is someone who has the authority to make a decision. A common definition of a decision is that it is "an irrevocable allocation of resources, in the sense that it would take additional resources, perhaps prohibitive in amount, to change the allocation" (Matheson and Howard 1983). Thus, to qualify as a decision maker, one must have the power to allocate resources. A stakeholder is someone who can affect or is affected by a decision (for interesting discussions of the stakeholder concept, see Bryson 2004; Ackermann and Eden 2011). An expert, by contrast, is someone who has knowledge relevant to the assessment of the characteristics of the options at hand (for more discussion, see European Food Safety Authority 2014, Appendix A.2.2).

12.3.1.3 Step 3. Design the Intervention

Often, MAVT is used in a participative way—in what Franco and Montibeller (2010) call the "facilitated mode" of analysis. Sometimes, the entire decision analysis process will take place in a workshop or series of workshops (this is sometimes known as "decision conferencing"—Phillips 2007). Workshops are often valuable as they build consensus and enable disagreements to be explored and sometimes resolved, however, they can be time-consuming and expensive. On other occasions, analysis may be done entirely "in the backroom"—such behind the scenes analysis can still be valuable contribution to clarifying the problem and guiding a path to a decision.

Sometimes, it may be most useful to have a hybrid process. For example, scoring can be done "offline" by individuals, so that when face-to-face discussion takes place it can focus on where there are differences of opinion in the scoring. In thinking through the design of an intervention, it may be useful to fill in a matrix of the form shown in Table 12.1.

Different modes of working may make sense in different contexts. For example, when options are scientific projects which contribute to public welfare, it may make sense to have scientists identify the options and perform the scoring on an individual basis, but for representatives of the relevant stakeholders to do weighting in workshop.

12.3.2 Structuring the Model

12.3.2.1 Step 4. Identify the Options

Options (sometimes called alternatives or actions) are things which could be done.

Options should be:

- Creative. It is important to canvass a wide range of options, even options which are not immediately doable.
- Manageable in the time available. The number of options drives the length of time required by the analysis.

Table 12.1 Matrix for determining involvement in a MAVT application

	Who to involve?	How to involve?
Options		
Criteria		
Scoring		
Weighting		
Sensitivity analysis		

- Homogeneous—they should be the same sort of thing. For example
  - A facility which will deliver benefits over a 5 year timeframe cannot be directly compared with a facility which will deliver benefits over a 100 year timeframe.
  - An investment option which costs £50 cannot be directly compared with an investment option which costs £1,000,000.
- If more than one option can be done, options should be evaluatively independent, that is, it should be possible to evaluate an option *a* without knowing whether a second option *b* is to be implemented.

Let us look at an example where evaluative independence might fail. I cannot evaluate "coffee" without knowing whether I am also to receive "milk" (as it happens, I prefer not to drink my coffee black) and vice versa. If options are not evaluatively independent they can sometimes be restructured to achieve evaluative independence (e.g. I combine "coffee" and "milk" into a single option). Sometimes this is not possible, and more complex approaches are required, such as the use of mathematical programming methods.

It is good practice to identify a baseline level of activity ("do nothing"). This is particularly important where the problem is not a problem of single choice. In multiple choice contexts, failure to identify an appropriate baseline can lead to paradoxical behaviour where model results change depending on seemingly arbitrary features of model specification—see Morton (2015) for more details.

12.3.2.2 Step 5. Identify the Criteria

Criteria are the measures of performance by which an option is judged. Just because in MAVT—and indeed in Multi-Criteria Decision Analysis (MCDA) procedures more generally—the aim is to identify criteria which can be used to guide choice, this does not mean that that these criteria really "exist" in the world: they have to be discussed, negotiated and agreed between the various decision makers. Indeed, research tells us that people are often not even sure what their own objectives are, even in problems which are quite important to them (Bond et al. 2010): this is a reason why there needs to be a structured process to discuss these objectives and arrive at a model which everyone can sign up to.

A useful question to identify criteria is consider the options and ask the question "what would distinguish between a good and bad choice in this decision problem?" Criteria thus form a bridge between the options and the objectives.

Criteria have a sense or direction of preference:

- If one prefers more of the criterion to less (e.g. revenue), one says it has an *increasing direction of preference*
- If one prefers less to more (e.g. cost) one says it has a *decreasing direction of preference*.

Table 12.2 Performance matrix for the house choice problem

	Criteria			
House	Financial cost (£)	Closeness (zone)	Character	Size (Sq footage)
1	220	A	Yes	600
2	180	B	Yes	600
3	130	C	No	700
4	120	C	No	500
5	180	B	No	600

Once criteria have been identified, it should be possible to describe how the options perform against the criteria. This can be done by specifying a performance matrix, with options along the vertical dimension and criteria along the horizontal direction. The individual performances are described in the cells: these can be described either in terms of natural attributes (e.g. number of lives saved); constructed attributes (e.g. numbers of stars which summarise further disaggregate information); or qualitative descriptions (e.g. "very good"; "barely adequate").

Suppose one is choosing a house to purchase. Table 12.2 shows an example of a performance matrix (this example also appears in Morton and Fasolo 2009). Here, Financial Cost is operationalised through money (in £); Closeness to the city centre is operationalised through the zone of the city in which the house is located (A is closest to the centre and C is furthest way); Character is assessed as a simple "yes" or "no"; and size is measured in square footage. Size thus has increasing direction of preference (more preferred to less) whereas Financial cost has decreasing direction of preference (less preferred to more). The measures which are used to operationalise the criteria are called *attributes*: unlike criteria which are expressions of a decision maker's aspirations in a decision problem, attributes are objective characteristics which can be "read off" from a description of the options themselves.

Sometimes it is possible to identify options which are *dominated*. An option a is said to be dominated by a second option b if b is at least as good as a on each criterion and strictly better than a on at least one criterion. In single choice problems, dominated options will always be ranked at least second, and so can be eliminated from consideration. For example, in the house choice problem, House 5 is dominated by House 2. It performs the same as House 2 on every criterion except Character: House 2 has character and House 5 has no character.

If there are a large number of criteria, it may be worthwhile structuring the criteria as a hierarchical value tree—see e.g. Fig. 10.2 of Chap. 10 (González-Ortega et al. 2018). As a whole, the set of criteria should be (Keeney and Raiffa 1976):

- *Discriminatory*. They should distinguish between options. Sometimes there may be objectives which are felt to be very important but which do not distinguish between the options under consideration (e.g. how a software program is designed may have no impact on climate change). In this case, there will be no criterion associated with this objective in this decision problem.

- *Complete.* Criteria should capture everything which the decision makers and stakeholders care about.
- *Small in number.* As with options, a large number of criteria result in options will increase time and care should be taken not to list too many criteria.
- *Non-redundant.* Criteria should not duplicate each other: there should be no double counting.
- *Preferentially Independent* (as discussed earlier in this chapter). One useful way to test whether preferential independence holds in practice is to see whether it is possible for the elicitee to assess the value of performance on one criterion independently of the level of performance on another criterion. If not, this suggests that preferential independence does not apply and so the model should be restructured, or a non-additive value model should be applied.

As I have stressed above, preferential independence is critical if scoring and weighting approaches are to be used. Here is an example where preferential independence might fail in our house choice setting. In choosing a house a purchaser may care about whether there is a park nearby, and about whether there is a swimming pool nearby: but if there is a park, she no longer care so much about the swimming pool (and *vice versa*). Often, as in this case, failure of preferential independence indicates that there is a higher order value (in this case, whether there are facilities for exercise), and if the two preferentially dependent criteria are replaced with the single more fundamental one, the problem is resolved. For a discussion of models which make implausible preference independence assumptions in the health domain, see Morton (2017).

12.3.2.3 Step 6. Score the Options on the Criteria

MAVT involves making numerical assessments of value and of relative importance. Sometime this can be hard for people to do because they are used to thinking of numbers as representing data about things which are "out there in the world". This is the wrong way to think about the numbers which are used in MAVT: numbers are used but as part of a language to express how people feel about their values. Questions which are mathematically equivalent from the point of view of the multicriteria model can often be experienced psychologically as being quite different (Morton and Fasolo 2009). For this reason it is often useful to have different ways to ask MAVT elicitation questions: I will review some of these different ways in this subsection.

It is conventional to use a scale bounded by 0 and 100 within each criterion to score options. The performance levels which are defined as 0 and 100 are called the lower and upper reference points. In single choice problems, a common approach is to set the worst performance level in each criterion as 0 and the best as 100; an alternative approach is to anchor the scale at 0 by some absolute idea of a "neutral" level of performance and at 100 by some absolute idea of a "good" level of performance. In problems other than single choice problems, it is good practice

to set the *do nothing* baseline level of performance equal to zero (this may mean that some options have negative scores). This is required in order to ensure that the value of two options together (against the baseline) is equal to the sum of the individual values of the options (against the baseline) (see Morton 2015, for more details).

Once 0 and 100 have been assigned, it remains to score the remaining options. The scores should have a preference intensity interpretation. This means, they should represent how intensely option *a* is preferred to be *b* relative to how intensely *c* is preferred to *d*. For instance, if the difference between the scores of *a* and *b* is 40 points, and the difference between the scores of *c* and *d* is 20 points, then *a* is preferred to *b* twice as strongly as *c* is preferred to *d*.

To actually establish the value scores of these intermediate points, it is helpful to have multiple ways to help the elicitee access their values. For example, one can ask the elicitation question as follows:

> Suppose you living in a house in Zone C and you woke up one morning to find your house had been moved to Zone A. You would feel happy, right? Fix in your mind how happy you would feel . . . Now, suppose instead of your house moving from Zone C to Zone A, it only moves to Zone B . . . . You would still feel happy, but you would feel less happy, right? Now, can you tell me how big is the second amount of happiness as a fraction of the first amount of happiness?

If the answer to this line of questioning is, say, "I would feel two thirds as happy", then the value Zone B should be 67 (on a scale where Zone C is zero and Zone A is 100). It is normally to do "consistency checks" on such number. For example, if Zone B does indeed have a score of 67, this means that a move from Zone C to Zone B should give twice as much happiness as a move from Zone B to Zone A. It is generally worth checking out with the elicitee whether this does indeed correspond to how they feel about the options.

Often there is a certain amount of initial resistance to expressing such quantitative judgements. The elicitor should give the elicitee time to surface the qualitative arguments which may support a judgement of preference intensity. To facilitate the expression of a preference judgement, it is often useful to draw measurement scales and or different numbers or smileys to represent different degrees of happiness (see Fig. 12.5). In group settings, a useful way to get a discussion going is to ask each member of the group to privately assess a score and then compare and discuss differences.

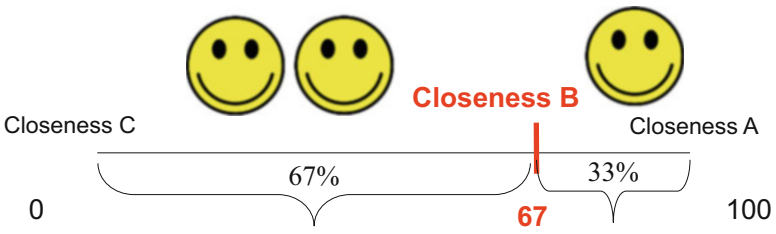


Fig. 12.5 Assigning a score for the intermediate level of closeness

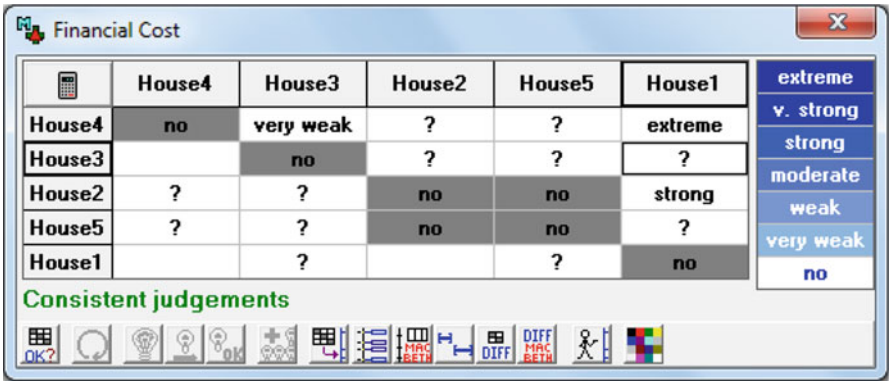


Fig. 12.6 Establishing value scores for cost: the MACBETH approach

One way to avoid the reluctance which many people feel to putting numbers on their feelings is to ask not for quantitative scores but for qualitative statements about strength of preference. This is the approach of Analytic Hierarchy Process or AHP approach and the MACBETH approach (see Belton and Stewart 2002 for a presentation of both approaches in a comparative context). MACBETH is fully compatible with the MAVT paradigm, whereas AHP has been criticised in the decision analysis literature on the grounds that it can lead to rank reversals (Dyer 1990).

A screenshot from the MACBETH software is shown in Fig. 12.6. In the software, options are arranged in a matrix, and elicitees are invited to make statements about the qualitative strength of preference between a number of different pairs. For example, the elicitee may state that the difference in preference in terms of cost between House 4 (the cheapest) and House 1 (the most expensive) is "extreme", whereas the difference between House 4 and House 3 (the next cheapest) is merely "very weak". The MACBETH software will then construct a value scale placing the options at appropriate points on the scale, by using linear programming optimisation in which the variables are the scores. The software also facilitates other forms of analysis. In particular the software has an inbuilt function which performs consistency checks on the matrix of comparisons (to identify situations where e.g. a is strongly preferred to b and b is strongly preferred to c but a is only weakly preferred to c) and suggests how consistencies can be resolved. For further introduction to MACBETH, see Bana e Costa and Chagas (2004) of Bana e Costa et al. (2012).

Table 12.3 shows some possible scores in the house choice problem, with the lower reference point set as the worst level of performance and the upper reference point set as the best level of performance.

Note that the criterion-specific scores as depicted in Table 12.3 are simply vectors of numbers. If the underlying attribute is continuous (e.g. money, quantity of emissions etc.), it may be possible to draw a value function. A value function

Table 12.3 Attribute scores for the house choice problem

	Criteria			
House	Financial cost	Closeness	Character	Size
1	0	100	100	67
2	50	70	100	67
3	95	0	0	100
4	100	0	0	0
5	50	70	0	67

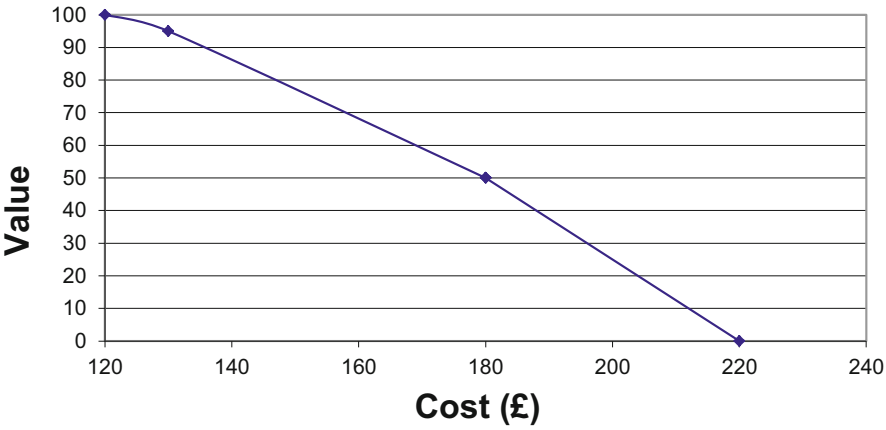


Fig. 12.7 A possible value function for cost

captures graphically how incremental value changes as the level of performance changes. Figure 12.7 shows a possible value function for cost. Note that this value function is decreasing, capturing the idea that lower costs are preferred to higher ones; it is also non-linear, capturing the idea that the decision maker cares more about an increment of £40,000 in cost from a base of £180,000 than from a base of £120,000 (i.e. the difference in value between £120,000 and 160,000 is about 30 whereas the difference in value between £180,000 and 220,000 is 50).

One natural way to elicit a value function is to use the *bisection method.* When using this method, one asks the elicitee to find a price point x such that a reduction in cost from the highest level (£220 K) to £x yields the same amount of value as a reduction in cost from £x to the lowest price level (£120 K). Since the most preferred price point has a score of 100 and the least preferred a score of 0, £x should therefore have a score of 50. By iterating this procedure, price points corresponding to value scores of 25, and 75 can be found, and then corresponding to 12.5, 27.5, 62.5 and 87.5 .... to any required degree of articulation.

It should be noted that value functions are quite different from performing an (often arbitrary) normalisation of the attribute scales. Normalisations are an automatic mathematical operation that does not represent preferences. A value function represents preferences and therefore must result from an elicitation process.

12.3.2.4 Step 7. Weight the Criteria

Once scores have been established, the next step is to weight the criteria. The reason for weighting is that although options have been scored on individual criteria, criteria scales are not commensurable: a unit of value on one criterion scale is not the same as a unit of value on another scale. It is as if the options had been valued in terms of different currencies: UK pounds, euros, US dollars, etc..

Weighting thus sets the "exchange rates" between the different criteria. It is critical to do weighting properly as this is what distinguishes MAVT from *ad hoc* approaches. In *ad hoc* approaches, people often set weights by asking questions such as "how important is this criterion relative to that criterion?". Although people can answer such questions, the questions themselves are meaningless (Morton and Fasolo 2009). In MAVT, the weighting questions are phrased in terms of increments on different scales.

To see this the difference, consider the question "Which is more important, saving money or saving lives?". This question as posed is ill-formed. However, the question of how much one is prepared to pay to correct implement a safety feature which will save on average such-and-such a number of lives is a well-formed question. MAVT relies on questions of this latter type.

The most popular method of weighting in MAVT depends on the concept of *swings*. A swing is typically defined as an increase in performance from the level of performance associated with the lower reference point on some criterion to the level of performance associated with the upper reference point. A weight reflects the value of a swing, i.e. the value of improving an option which performs at the lower reference point level on some criterion, so that it performs at the upper reference point level on that criterion. Conventionally the weight of the most valued swing is set as 1 and the weights of the other swings are set as fractions of the most valued swing.

Just as in scoring, swing weighting involves asking questions about hypothetical changes in options. The following question can be used to produce a ranking of the swings.

> Imagine you are going to buy a house which has the worst performance levels on all criteria (it costs £220K, is situated in Zone C, has no character, and is only 500 square feet in size). One day, your fairy godmother appears and offers to grant you some wishes. She is unsure how many wishes she has to grant and asks you to prioritise. You may reduce the cost to £120K, change the location from Zone C to Zone A, bestow the flat with character, and increase the size to 700 sq feet. Which do you choose first, which second, which third, and which fourth?

This procedure generates a ranking of the swings. (In our case, suppose the ranking is Financial Cost, Closeness, Size and Character.) The next step is to ask the "how much" question: how much do you like the second swing as a proportion of how much you like the first? how much do you like the third swing as a proportion of how much you like the first? how much do you like the fourth swing as a proportion of how much you like the first?

The principles behind asking and answering such questions are exactly the same as, and build on the scoring questions: allow elicitees time to reflect and debate, visualise, and make consistency checks to ensure that results "feel right". The MACBETH software can also be used for weighting, by eliciting qualitative statements about strength of preference ("extreme", "very strong", etc.) between the possible swings. a particular advantage of this software is that it also incorporates dominance checks which can supplement quantitative scores by showing how strong the evidence is that one option is overall more highly ranked than another.

Table 12.4. shows swings and associated swing weights for the house choice problem.

As in the case of scoring, where attributes are continuous, this allows an alternative procedure for weighting, called tradeoff weighting. The idea in tradeoff weighting is to adjust the more preferred swing until it yields as much value as the less preferred swing. The concept is depicted in Fig. 12.8. Suppose we have two options, Option 1 which is cheap but poky (£120 K, 500 sq ft) and Option 2 which is roomy but expensive (£220 K, 700 sq ft). We like both of these flats better than an expensive and poky flat (£220 K, 500 sq ft, called the "nadir"). Moreover, we know from the answer to our fairy godmother question that we would prefer the Option 1 to Option 2: Financial Cost is our most valued swing.

Now we want to ask the "how much" question. But instead asking it directly, we can ask in the following way. Suppose that I adjust Option 1 downwards, in the direction of the nadir, by increasing the price. At some point, Option 1 will cease to be better than Option 2, and become first indifferent and then worse. By locating the point at which indifference occurs, I can find a weight for Size in terms of Financial

Table 12.4 Swings and weights for the house choice problem

	Criteria			
	Financial cost	Closeness	Character	Size
Worst performance level	220	3	No	500
Best performance level	120	1	Yes	700
Swing	220 → 120	3 → 1	no → yes	500 → 700
Unnormalised swing weight:	1.00	0.85	0.30	0.50

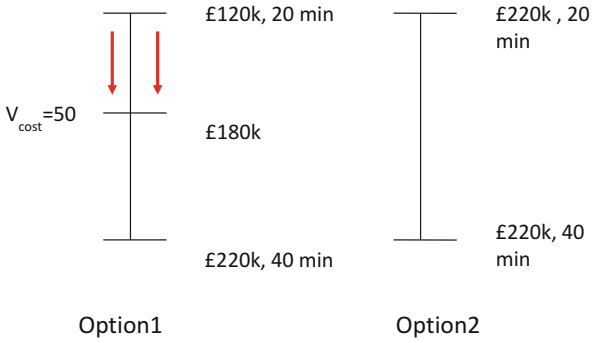


Fig. 12.8 Sketch of the procedure for tradeoff weighting

cost. The reasoning works as follows: I read the price level of the indifference price (£180 K, say), and look it up on my value function for cost. From this I see that a price of £180 K as compared to £120 K is worth 50 value points, measured on the scale of the value function for cost. Since the value of the swing from 500 to 700 sq ft is 100, measured in the scale of the value function for size, if I want to express the value of square footage in a way which is commensurable with the value of cost, I must divide the value scores for size by 2, i.e. use a weight of 0.5.

12.3.3 Analysing the Model

12.3.3.1 Step 8. Compute Overall Rankings

Given the scores and weights, and if the options and criteria have the properties outlined in Steps 4 and 5, then it is legitimate to compute an overall value score for each option a using the following formula

$$v(a) = \sum_j w_j v_j(a)$$

where w_j is the weight of criterion j and $v_j(a)$ is the score of option a on criterion j. This provides a ranking of all options, and can be used to identify the best option, or k best options.

It should be noted that "weight of criterion j" is something we often say in common language, but more formally it should be called "the scaling constant associated with value function v_j". Since these weights might not match the decision makers's intuition (e.g., "how come safety has such a low weight?") it might be useful to communicate it as the weight of value function v_j (or the weight of swing j). Bana e Costa et al. (2008) present an interesting and instructive application where particular attention was paid to designing the swing weighting procedure so that the swing weights corresponded closely to the decision makers' natural prior understanding of criterion importance.

Sometimes where there multiple options can be implemented together and there is a concern for value for money, an alternative formula

$$vfm(a) = \frac{\sum_j w_j v_j(a)}{c(a)}$$

may be used, where j indexes the criteria on the benefit side of the value tree only, and $c(a)$ is the cost of option a (excepting the "do-nothing" option which has cost of zero). This formula has the advantage that ordering the options according to this formula and proceeding down the list until the budget is exhausted, will give a good solution to the budget allocation problem, especially if there are many options. For

more ideas on how to deal with this particular *problématique*, see Salo et al. (2011) and Morton et al. (2016).

12.3.3.2 Step 9. Conduct Sensitivity Analysis

Often people consider that a multicriteria analysis is complete when they have scored options and weighted criteria and arrived at a ranking of options. Nothing could be further from the truth. The aim of MAVT is not to find the "right answer"—where there are conflicting objectives, no right answer exists—but to enable decision makers and stakeholders to explore the problem and come to a considered decision. Sensitivity analysis involves varying scores or weights in an interval and noting the impact on the model results. Sensitivity analysis can reveal how important uncertainties or disagreements (such as those identified in Steps 6–7) are on the final results.

I now present three sensitivity analysis displays for the house choice problem: the stacked bar chart, the Pareto chart, and the parameter-wise sensitivity analysis.

Figure 12.9, the stacked bar chart, shows the composition of aggregate value for the different options. From this it can readily be seen what options are cheap (a lot of the value of Houses 3 and 4 is due to their strong performance on the cost criterion). House 4 in particular has nothing to recommend it except that it is cheap. House 1 gets a great deal of value from closeness and if the elicitee really cared about closeness she would choose this option, but the winner seems to be—given these scores and weights—House 2 which has the advantage that it is a good all-rounder, with cost, closeness, character and size reasons to recommend it.

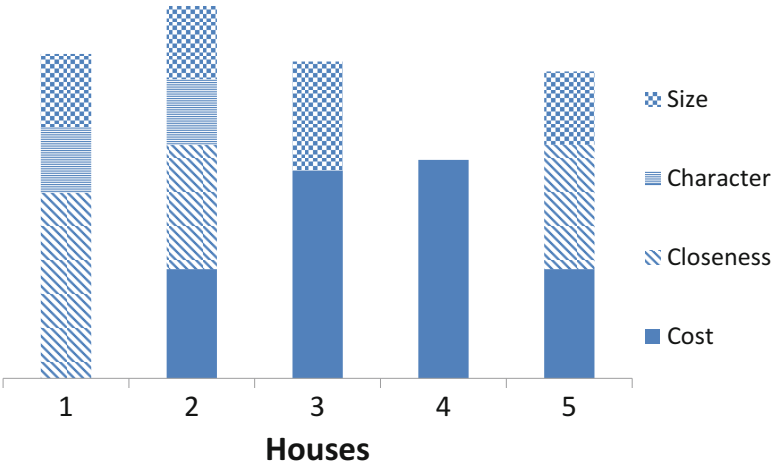


Fig. 12.9 Stacked bar chart for the house choice problem

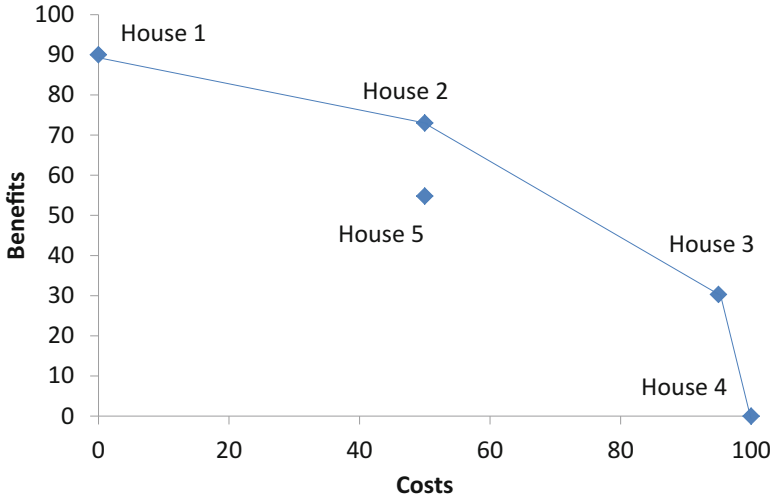


Fig. 12.10 Efficient front display for house choice problem

Figure 12.10 shows a Pareto chart. In this display, the scores for Financial cost are plotted against a weighted combination of the scores of all other criteria ("Benefits"). Houses on the frontier of the enclosed area are efficient in the sense that for each house, that there is some assignment of weights to "Costs" and "Benefits" which makes that house the highest valued house. House 1 is the point on the vertical access (it has all the benefits but is expensive); House 4 is the house on the horizontal axis (it has no benefits but is cheap) and Houses 2 and 3 are the points on the curve, both representing a compromise between costs and benefits. Note that House 5 is not efficient in this display. This is a consequence of house 5 being dominated. It is however possible for an option to be not efficient even if it is not dominated.

Figure 12.11 shows a parameter-wise sensitivity analysis for the criterion Closeness. This display shows how the valuation of the options changes as the weight on Closeness is varied relative to the weight on the other criteria whilst holding the relative weights on the Benefit criteria fixed. From this display it can be easily seen that: House 1 is a good option if Closeness is high weighted relative to the other criteria; House 2 is a good option if Closeness is intermediate weighted relative to the other criteria; and House 3 is a good option if Closeness is low weighted relative to the other criteria. The other two options do not, for this analysis and given these numbers, make it into the running.

Although sensitivity analysis in MAVT can be done using spreadsheets, it is often more efficient to use software (for example, Hiview or VISA or MACBETH or WISED for single choice problems; Equity or PROBE for multiple choice and budget allocation problems) as these softwares have built-in sensitivity analysis tools. The technical literature has a wider selection of ideas and tools for performing

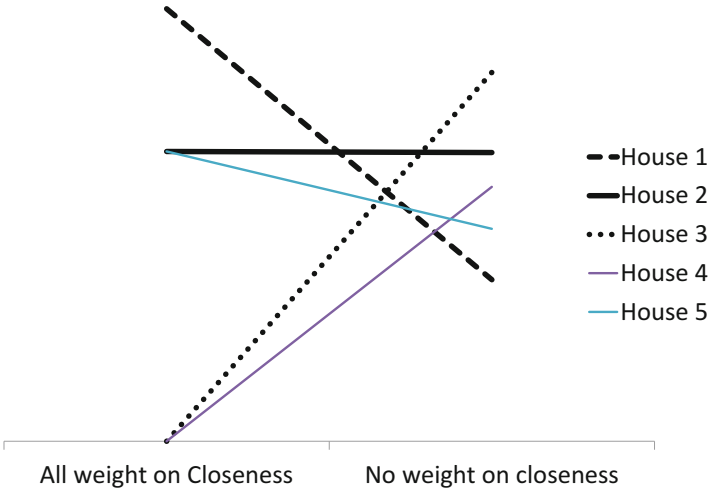


Fig. 12.11 Parameter-wise sensitivity analysis: value of options varying weight on Closeness for house choice problem

sensitivity analysis (e.g. Ríos Insua and French 1991; Dias and Clímaco 2000; Borgonovo and Plischke 2016) but these have not yet generally been incorporated in professional commercially-available user-friendly software.

12.3.4 Troubleshooting

In this section of the chapter, I consider some commonly occurring problems in applying MAVT in practice, and suggest some ways to approach such problems.

1. *There are too many options.* Consider using a small number of screening criteria to establish a shortlist (e.g. would this option require new legislation to implement? Would it cost more than £x?). If several options are similar (e.g., small variations), consider evaluating only one from each group/cluster and, if it turns out to be among the best, only then evaluate the ones similar to it.
2. *There are not enough options.* Look for solutions which other organisations have implemented when faced this or similar decisions. Consider holding a brain storming session. Consider enlarging the scope of the analysis. E.g., someone suggests you do voluntary work at an hospital 2 h per week. Instead of considering the alternatives "yes" and "no", you might consider the alternatives are how many hours will you devote to this organisation, or consider the alternatives are different organisations where you could do voluntary work.

3. *The options do not seem to be comparable.* Come up with a description of what options should be (e.g. "facilities"; "development plans"). Restructure the options by merging some or deleting some.
4. *The options cannot be evaluated independently of each other*. Consider restructuring the options (e.g. merging options which have a dependence relation; assuming that one option on which several others depend will be done). Alternatively, consider using more complicated analytic techniques such as mathematical programming. Consider using Portfolio Decision Analysis methods, see Salo et al. (2011) and Morton et al. (2016).
5. *There are too many criteria.* Look for criteria which are redundant, i.e. which duplicate each other; which do not discriminate between options. Consider merging similar criteria into higher level criteria.
6. *There are not enough criteria.* Look for criteria which other organisations have implemented when facing this or similar decisions. Consult published documents such as strategic plans. Consider holding a brain storming session. Consider what important attributes might differentiate two alternatives that are similar on the criteria you already have.
7. *The criteria are not preferentially independent.* Consider restructuring the criteria (e.g. merging two criteria which are dependent because they are alternative ways of achieving some higher order goal). Use a non-additive value model.
8. *Participants don't understand scoring and/or weighting.* Use software, or draw pictures on flip charts to help participants visualise. Ask questions in different ways, using the different questioning modes listed in this chapter. Use analogies to communicate weight and scale concepts (e.g. exchange rates; metric and imperial scales; Celsius and Fahrenheit). Build models in real-time allowing to observe how outputs change as inputs also change.
9. *The overall values don't "feel right"*. Ask yourself and your decision maker why the answers don't feel right. Is there a missing criterion? Do you really believe the scores and weights? Use sensitivity analysis to explore the model.
10. *There isn't enough time to do everything properly.* One option is to proceed with incomplete information and check what is robust to save time, see Dias (2007). However, a decision analysis can take various forms—from a quick back-of-the-envelope analysis in an hour or two to workshops spread over several days. Use the time you have, and be realistic about what you can achieve.
11. *The decision makers or significant parties do not have time to participate.* Do not demand very exact answers (e.g are you sure the score is 50 and not 49 or 51?). Often, sensitivity analysis shows that small imprecisions do not matter ("flat maxima principle" of von Winterfeldt and Edwards 1986). Remind the decision maker that the analysis is a tool to help them structure and think through the decision, not something which will or should try to take the decision for them. If time is an issue, not everyone has to be involved in every stage of the decision process (for example a small working group may define criteria and options which can be scored by a larger group).

12. *The decision makers or significant parties are afraid of losing control of the decision.* Not everyone has to be involved in every stage of the decision process (for example weights may be defined by the management team or by a single client). Control is not absolute in any case, and often decisions which are arrived at by a non-transparent process are hard to implement because of stakeholder resistance.
13. *The decision makers do not agree on some inputs.* Build different models in parallel or use incomplete information they agree with (e.g., they do not agree on the weights w_1 and w_2, but agree that $w_1 > w_2$. Assess what common results can be obtained. Often, different inputs lead to the same outputs.
14. *The decision makers refuse the idea of trade-offs* (e.g., harm to the environment vs. harm to health vs. costs). This may be caused by options with unacceptable performance on key criteria that the decision makers feel cannot be compensated by good performance in another criterion. In such cases, consider removing these unacceptable alternatives. Otherwise, using MAVT might not be the best option and outranking methods (see Chap. 14 of this book (Dias and Mousseau 2018)) or other approaches might be appropriate to such type of decision makers.

12.4 Concluding Remarks

The founding texts of MAVT (Keeney and Raiffa 1976; von Winterfeldt and Edwards 1986) are now 40 and 30 years old respectively. Although younger by several decades (or centuries, depending on how one counts) than probability theory, MAVT can therefore also be considered to be a mature technology.

Is it a successful technology? Considered in its broadest sense, the answer has to be yes: the scoring and weighting approach is (as far as one can tell) very widely used in applied settings, such as R&D prioritisation and procurement. However, many users of scoring and weighting have never heard of MAVT, and are unaware that that a body of theory-based knowledge exists about how to perform elicit scores and weights. To some extent this is also true of probabilistic modelling also. However, much of the use of probabilistic concepts is mediated by software such as spreadsheet simulation packages and such software provides an easy bridge for users to learn more about probabilistic concepts. Software based around MAVT concepts has not (yet) enjoyed such widespread success.

Like the authors of the Chap. 9 (González-Ortega et al. 2018, in their Discussion section) I see huge potential for MAVT methods in an increasingly digital and data-rich world. Currently if one is shopping online for hotel rooms or flights, the search engines allow one to rank order options on the basis of holistic assessments, or on the basis of individual criteria, but provide little in the way of support for locating the option which has the ideal balance of attributes given one's preferences. It is plausible that increasingly demanding online consumers will at some point start to

ask for and expect better decision support to enable them to cope with the vast mass of undigestable information which is regularly served up to them.

However, the original promise of MAVT as a rigorous yet transparent framework for choice was to help support big policy decisions as well as small personal ones. There are some signs in some domains that multicriteria methods are meeting with increasing favour. In the area of health technology regulation and assessment, for example, there has been a recent upsurge in interest in the use of multicriteria methods to support medicines regulation and reimbursement decisions (Thokala et al. 2016; Marsh et al. 2016). However, there is still a substantial gap between the potential for the formal use of MAVT to beneficially support substantial decisions in government and business, and actual current practice. Hopefully that gap will close in the years and decades ahead.

Acknowledgements I would particularly like to acknowledge the many helpful comments of Luis Dias, many of which have been directly incorporated in the text.

References

Ackermann F, Eden C (2011) Strategic management of stakeholders: theory and practice. Long Range Plan 44(3):179–196

Bana e Costa CA, Chagas MP (2004) A career choice problem: an example of how to use macbeth to build a quantitative value model based on qualitative value judgments. Eur J Oper Res 152(2):323–331

Bana e Costa CA, de Corte J-M, Vansnick JC (2012) MACBETH. Int J Inf Technol Decis Mak 11(2):359–387

Bana e Costa CA, Lourenço JC, Chagas MP, Bana e Costa JC (2008) Development of reusable bid evaluation models for the Portuguese electric transmission company. Decis Anal 5:22–42

Belton V, Stewart TJ (2002) Multiple criteria decision analysis: an integrated approach. Kluwer, Boston, MA

Bond SD, Carlson KA, Keeney RL (2010) Improving the generation of decision objectives. Decis Anal 7:238–255. doi:10.1287/deca.1100.0172

Borgonovo E, Plischke E (2016) Sensitivity analysis: a review of recent advances. Eur J Oper Res 248(3):869–887

Bryson JM (2004) What to do when stakeholders matter stakeholder identification and analysis techniques. Public Manag Rev 6(1):21–53

Dias LC (2007) A note on the role of robustness analysis in decision-aiding processes. In: Roy B, Ali Aloulou M, Kalaï R (eds) Robustness in OR-DA, Annales du LAMSADE, No. 7. Université-Paris Dauphine, Paris, pp 53–70

Dias LC, Clímaco JN (2000) Additive aggregation with variable interdependent parameters: the VIP analysis software. J Oper Res Soc 51(9):1070–1082

Dias LC, Mousseau V (2018) Eliciting multi-criteria preferences: ELECTRE models. In: Dias LC, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgment. Springer, New York, NY

Drummond MF, Sculpher MJ, Claxton K, Stoddart GL, Torrance GW (2015) Methods for the economic evaluation of health care programmes. Oxford University Press, Oxford

Dyer JS, Sarin RK (1979) Measurable multiattribute value functions. Oper Res 27(4):810–822

Dyer JS (1990) Remarks on the analytic hierarchy process. Manag Sci 36(3):249–258

European Food Safety Authority (2014) Guidance on expert knowledge elicitation in food and feed safety risk assessment. EFSA J 12(6):3734
French S (1986) Decision theory: an introduction to the mathematics of rationality. Ellis Horwood, Chichester
French S, Ríos Insua D (2000) Statistical decision theory. Kendall's Library of Statistics. Arnold, London
González-Ortega J, Radovic V, Ríos Insua D (2018) Utility elicitation. In: Dias LC, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgment. Springer, New York, NY
Goodwin P, Wright G (2014) Decision analysis for management judgement, 5th edn. Wiley, Chichester
Howard RA, Abbas AE (2016) Foundations of decision analysis. Pearson, Harlow, Essex
Keeney RL, Raiffa H (1976) Decisions with multiple objectives: preferences and value tradeoffs. Wiley, Chichester
Köbberling V (2006) Strength of preference and cardinal utility. Economic Theory 27(2):375–391
Krantz DH, Luce RD, Suppes P, Tversky A (1971) Foundations of measurement vol 1. Academic, New York
Marsh K, IJzerman M, Thokala P, Baltussen R, Boysen M, Kalo Z, Lonngren T, Mussen F, Peacock S, Watkins J, Devlin N (2016) Multiple criteria decision analysis for health care decision making-emerging good practices: report 2 of the ISPOR MCDA emerging good practices task force. Value Health 19(2):125–137. doi:10.1016/j.jval.2015.12.016
Matheson JE, Howard RA (1983) An introduction to decision analysis. In: Howard RA, Matheson JE (eds) The principles and applications of decision analysis. SDG, Menlo Park, CA
Miyamoto JM, Wakker PP, Bleichrodt H, Peters HJM (1998) The zero-condition: a simplifying assumption in QALY measurement and multiattribute utility. Manag Sci 44(6):839–849
Franco LA, Montibeller G (2010) Facilitated modelling in operational research. Eur J Oper Res 205(3):489–500
Morton A (2015) Measurement issues in the evaluation of projects in a project portfolio. Eur J Oper Res 245(3):789–796
Morton A (2017) Treacle and smallpox: two tests for multicriteria decision analysis models in health technology assessment. Value Health 30(3):512–515
Morton A, Fasolo B (2009) Behavioural decision theory for multi-criteria decision analysis: a guided tour. J Oper Res Soc 60(2):268–275
Morton A, Keisler J, Salo A (2016) Multicriteria portfolio decision analysis for project selection. In: Ehrgott M, Figueira JR, Greco S (eds) Multiple criteria decision analysis: state of the art surveys, 2nd edn. Springer, New York, NY
Phillips LD (2007) Decision conferencing. In: Edwards W, Miles RF, Von Winterfeldt D (eds) Advances in decision analysis: from foundations to applications. CUP, Cambridge
Pliskin JS, Shepard DS, Weinstein MC (1980) Utility functions for life years and health status. Oper Res 28(1):206–224
Ríos Insua D, French S (1991) A framework for sensitivity analysis in discrete multi-objective decision-making. Eur J Oper Res 54:176–190
Roy B (1985) Méthodologie multicritère d'aide à la décision. Economica, Paris
Salo A, Keisler J, Morton A (eds) (2011) Portfolio decision analysis: methods for improved resource allocation. Springer, New York, NY
Thokala P, Devlin N, Marsh K, Baltussen R, Boysen M, Kalo Z, Longrenn T, Mussen F, Peacock S, Watkins J, Ijzerman M (2016) Multiple criteria decision analysis for health care decision making-an introduction: report 1 of the ISPOR mcda emerging good practices task force. Value Health 19(1):1–13. doi:10.1016/j.jval.2015.12.003
von Winterfeldt D, Edwards W (1986) Decision analysis and behavioral research. CUP, Cambridge

Chapter 13
Disaggregation Approach to Value Elicitation

Nikolaos F. Matsatsinis, Evangelos Grigoroudis, and Eleftherios Siskos

Abstract The philosophy of preference disaggregation in multicriteria decision analysis encapsulates the assessment/inference of preference models, from given preferential structures, and the implementation of decision aid activities through consistent and robust operational models. This chapter presents a new outlook on the well-known UTA method, which is devoted to the elicitation of values through the inference of multiple additive value models. On top of that, it incorporates the latest theoretical developments, related to the robustness control of both the decision model and the surfacing decision aiding conclusions. An application example on job evaluation is elaborated as an educative example, while other potential areas for future use applications of the methodological framework are listed. The chapter concludes with several promising directions for future research.

13.1 Introduction

The philosophy of preference disaggregation in multicriteria analysis is to assess/infer preference models from given preferential structures and to address decision-aiding activities through operational models within the aforementioned framework. In simple words, assuming that a decision is given, the preference disaggregation approach is focused on finding rational basis for the decision being made. Therefore, it is possible to assess the Decision-Maker's (DM's) preference model leading to exactly the same decision as the actual one (Siskos et al. 2016).

Preference disaggregation has been proven especially competent for complex decision making systems, in the presence of multiple conflicting and heterogeneous criteria (Jacquet-Lagrèze and Siskos 2001; Siskos et al. 2016). In such cases, the

N.F. Matsatsinis (✉) • E. Grigoroudis
Decision Support Systems Laboratory, School of Production Engineering and Management, Technical University of Crete, University Campus, Kounoupidiana, 73100, Chania, Crete, Greece
e-mail: nikos@ergasya.tuc.gr; vangelis@ergasya.tuc.gr

E. Siskos
School of Electrical and Computer Engineering, National Technical University of Athens, 9, Iroon Polytechniou Str., 15780, Athens, Zografou, Greece
e-mail: lsiskos@epu.ntua.gr

L.C. Dias et al. (eds.), *Elicitation*, International Series in Operations Research & Management Science 261, https://doi.org/10.1007/978-3-319-65052-4_13

standard explicit elicitation of preferential parameters, especially when the DMs are ignorant of the rationale and methods of Multiple Criteria Decision Aid (MCDA), is a complicated task, which poses a heavy cognitive burden, and often leads to results of questionable value and acceptance. These decision making systems can be alternatively addressed with the aid of the implicit procedures of preference disaggregation. Such procedures are also suitable and convenient for the case of multiple DMs (Group Decision Making), where the preference models of each are aggregated to a global one (Siskos and Grigoroudis 2010; Stavrou et al. 2018).

In the context of preference disaggregation, goal programming techniques have been the first approaches applied in order to assess/infer preference/aggregation models or develop linear or nonlinear multidimensional regression analyses (Siskos 1983). Among other, these first research efforts include the works of Charnes et al. (1955) and Karst (1958), who applied goal programming approaches in order to assure the consistency of the developed models with available data. In particular, Charnes et al. (1955) developed a linear model of optimal estimation of executive compensation (salaries), as consistent as possible with the data from the goal programming point of view, while the goal programming approach of Karst (1958) was a single linear regression model, minimizing the sum of absolute deviations. Later Wagner (1959) generalizes the Karst's model in the case of multiple linear regression and Kelley (1958) proposed an alternative optimality criterion (i.e., minimize Tchebycheff's criterion). Other early important efforts may refer to the works of Srinivasan and Shocker (1973) who proposed a linear value function assessment approach based on ordinal regression and pairwise judgments and the study of Freed and Glover (1981) who developed an inference approach for estimating the weights of linear value functions in the context of discriminant analysis using goal programming techniques.

The case of ordinal criteria in preference disaggregation is considered by the early works of Young et al. (1976) and Jacquet-Lagrèze and Siskos (1978) which focused on the inference of additive value functions by disaggregating a ranking of reference alternatives. In particular, Jacquet-Lagrèze and Siskos (1978) in the "Cahiers du LAMSADE" series present the UTA method ensuring that the additive value function is optimally consistent with the given ranking through linear programming (LP) techniques, contrary to Young et al. (1976) where optimality is not ensured given the adopted least squares techniques. The research presented in the "Cahiers du LAMSADE" series may be considered as the actual initiation of the development of disaggregation methods.

In the context of MCDA, the general decision-making methodology includes the modeling process of a consistent family of criteria $\{g_1, g_2, \ldots g_n\}$, where each criterion is a non-decreasing real valued function defined on A, as follows:

$$g_i : A \rightarrow \left[g_{i*}, g_i^*\right] \subset \mathbb{R}/a \rightarrow g(a) \in \mathbb{R} \tag{13.1}$$

where $\left[g_{i*}, g_i^*\right]$ is the criterion evaluation scale, g_{i_*} and g_i^* are the worst and the best level of the i-th criterion respectively, $g_i(a)$ is the evaluation or performance of action a on the i-th criterion and $\mathbf{g}(a)$ is the vector of performances of action a on the n criteria.

From the above definitions, the following preferential situations can be determined:

$$\begin{cases} g_i(a) > g_i(b) \iff a \succ b \ (a \text{ is preferred to } b) \\ g_i(a) = g_i(b) \iff a \sim b \ (a \text{ is indifferent to } b) \end{cases} \tag{13.2}$$

So, having a weak-order preference structure on a set of actions, the problem is to adjust additive value or utility functions based on multiple criteria, in such a way that the resulting structure would be as consistent as possible with the initial structure. This principle underlies the disaggregation approach, where the preference models are inferred given a set of global preference, contrary to the traditional aggregation paradigm, where the criteria aggregation model is known a priori, while the global preference is unknown.

In this context, the preference disaggregation approach (Jacquet-Lagrèze and Siskos 1982, 2001; Siskos 1980; Siskos and Yannacopoulos 1985) aims at analyzing the behavior and the cognitive style of the DM.

As noted by Jacquet-Lagrèze and Siskos (2001) the clarification of the DM's global preference necessitates the use of a set of reference actions A_R, which may include:

(a) A set of past decision alternatives (A_R: past actions);
(b) A subset of decision actions, especially when A is large ($A_R \subset A$) or
(c) A set of fictitious actions, consisting of performances on the criteria, which can be easily judged by the DM to perform global comparisons (A_R: fictitious actions).

As shown in Fig. 13.1, a combination of the previous options may also be applicable (i.e., A_R may include a subset of A, as well as a set of fictitious actions). According to Fig. 13.1, the preference disaggregation focuses on the following

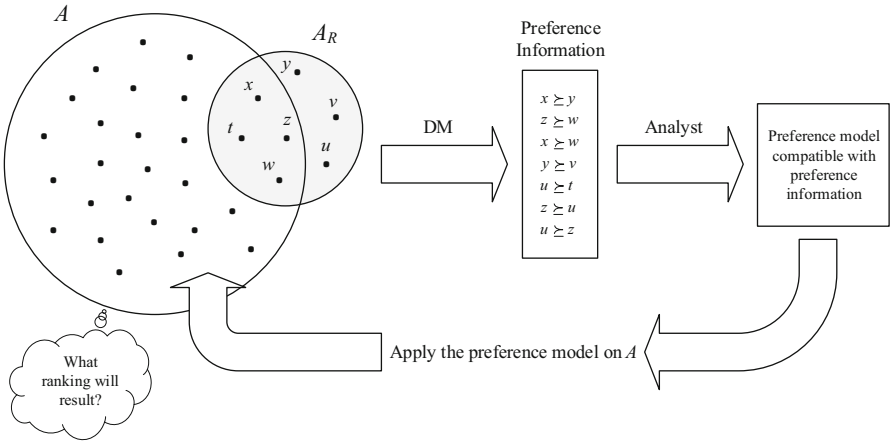


Fig. 13.1 Preference disaggregation procedure

question: based on preference information transformed to a compatible preference model, what is the consequence of using the whole set A? In the procedure the DM is asked to externalize and/or confirm his/her global preferences on the set A_R taking into account the performances of the reference actions on all criteria. The resulting DM's aggregated value system is then applied to A, thus the main aim of such an approach is to aid the DM to improve his/her knowledge on the decision situation and his/her way of preferring, which entails a consistent decision to be achieved.

The main aim of this chapter is to present a new perspective on the well-known UTA method, emphasizing on the elicitation of values through the inference of multiple additive value models. For this reason, the latest theoretical developments, related to the robustness control of both the decision model and the surfacing decision aiding conclusions are discussed and an educative example referring to an application on job evaluation is presented. The chapter concludes with the presentation of future research directions, as well as existing and potential applications of the proposed methodological framework.

13.2 A New Look on the UTA Method

13.2.1 Problem Statement and Notation

The UTA (UTilité Additive) method proposed by Jacquet-Lagrèze and Siskos (1982) aims at inferring one or more additive value functions from a given ranking or other preference statements (e.g., pairwise comparisons) expressed on a reference set A_R. The method uses LP techniques to assess these functions so that the ranking(s) obtained through these functions on A_R is (are) as consistent as possible with the reference preference statements.

The criteria aggregation model in UTA is assumed to be an additive value function of the following form (Jacquet-Lagrèze and Siskos 1982):

$$u(\mathbf{g}) = \sum_{i=1}^{n} u_i(g_i) \tag{13.3}$$

subject to normalization constraints:

$$\begin{cases} \sum_{i=1}^{n} u_i\left(g_i^*\right) = 1 \\ u_i\left(g_{i*}\right) = 0 \quad \forall i = 1, 2, \ldots, n \end{cases} \tag{13.4}$$

where u_i, $i = 1, 2, \ldots, n$ are non-decreasing real valued functions, named marginal value or utility functions.

Both the marginal $u_i(g_i)$ and the global $u(\mathbf{g})$ value functions have the monotonicity property of the so-called "true criterion". For instance, in the case of the global value function the following properties hold:

$$\begin{cases} u\left[\mathbf{g}(a)\right] > u\left[\mathbf{g}(b)\right] \iff a \succ b \ (preference) \\ u\left[\mathbf{g}(a)\right] = u\left[\mathbf{g}(b)\right] \iff a \sim b \ (indifference) \end{cases} \tag{13.5}$$

13.2.2 The UTASTAR Algorithm

The UTASTAR method proposed by Siskos and Yannacopoulos (1985) is an improved version of the original UTA model (Jacquet-Lagrèze and Siskos 1982). UTASTAR uses a double positive error function, so that the value of each alternative $a \in A_R$ can be written as:

$$u'\left[\mathbf{g}(a)\right] = \sum_{i=1}^{n} u_i\left[g_i(a)\right] - \sigma^+(a) + \sigma^-(a) \quad \forall a \in A_R \tag{13.6}$$

where σ^+ and σ^- are the underestimation and the overestimation error, respectively.

In addition, the monotonicity constraints in the UTASTAR method are taken into account through the following transformations:

$$w_{ij} = u_i\left(g_i^{j+1}\right) - u_i\left(g_i^{j}\right) \geq 0 \quad \forall i = 1, 2, \ldots, n \text{ and } j = 1, 2, \ldots, \alpha_i - 1 \tag{13.7}$$

and thus, the monotonicity conditions may be replaced by the non-negative constraints for the variables w_{ij} (α_i is the number of points, on which the value function u_i is assessed).

Based on the above, the UTASTAR algorithm may be summarized in the following steps:

13.2.2.1 Step 1

Express the global value of reference actions $u[\mathbf{g}(a_k)]$, $k = 1, 2, \ldots, m$, first in terms of the marginal values $u_i(g_i)$, and then in terms of the variables w_{ij}:

$$\begin{cases} u_i\left(g_i^1\right) = 0 & \forall i = 1, 2, \ldots, n \\ u_i\left(g_i^j\right) = \sum_{t=1}^{j-1} w_{it} & \forall i = 1, 2, \ldots, n \text{ and } j = 2, 3, \ldots, \alpha_i \end{cases} \tag{13.8}$$

In order to estimate the corresponding marginal value functions, when necessary, linear interpolation is applied. For example, if $g_i(a) \in \left[g_i^q, g_i^{q+1}\right]$ the marginal value function is given by:

$$u_i\left[g_i(a)\right] = u_i\left(g_i^q\right) + \frac{g_i(a) - g_i^q}{g_i^{q+1} - g_i^q}\left[u_i\left(g_i^{q+1}\right) - u_i\left(g_i^q\right)\right] = \sum_{t=1}^{q-1} w_{it} + \frac{g_i(a) - g_i^q}{g_i^{q+1} - g_i^q} w_{iq} \tag{13.9}$$

13.2.2.2 Step 2

Introduce two error functions σ^+ and σ^- on A_R by writing for each pair of consecutive actions in the ranking the analytic expressions:

$$\begin{aligned}\Delta\left(a_k, a_{k+1}\right) = u\left[\mathbf{g}\left(a_k\right)\right] - \sigma^+\left(a_k\right) + \sigma^-\left(a_k\right) - u\left[\mathbf{g}\left(a_{k+1}\right)\right] \\ + \sigma^+\left(a_{k+1}\right) - \sigma^-\left(a_{k+1}\right)\end{aligned} \tag{13.10}$$

13.2.2.3 Step 3

Solve the following LP:

$$[\min] z = \sum_{k=1}^{m}\left[\sigma^+\left(a_k\right) + \sigma^-\left(a_k\right)\right] \tag{13.11}$$

Subject to:

$$\left.\begin{array}{c}\Delta\left(a_k, a_{k+1}\right) \geq \delta \text{ if } a_k \succ a_{k+1} \\ \Delta\left(a_k, a_{k+1}\right) = 0 \text{ if } a_k \sim a_{k+1}\end{array}\right\} \forall k$$
$$\sum_{i=1}^{n}\sum_{j=1}^{\alpha_i - 1} w_{ij} = 1$$
$$w_{ij} \geq 0, \sigma^+\left(a_k\right) \geq 0, \sigma^-\left(a_k\right) \geq 0 \forall i, j \text{ and } k \tag{13.12}$$

where a_k and a_{k+1} are two successive actions in the DM's ranking and δ is a small positive number.

13.2.2.4 Step 4

Test the existence of multiple or near optimal solutions of the LP (12) (stability/robustness analysis); in case of non-uniqueness, find the mean additive value function as the most representative (barycenter) of those (near) optimal solutions which maximize/minimize the objective functions:

$$u_i\left(g_i^j\right) = \sum_{t=1}^{j-1} w_{it} \text{ for } i = 1, 2, \ldots, n \text{ and } j = 2, 3, \ldots, \alpha_i \tag{13.13}$$

on the polyhedron of the constraints of the LP (12) bounded by the new constraint:

$$\sum_{k=1}^{m} \left[\sigma^{+}(a_k) + \sigma^{-}(a_k)\right] \leq z^{*} + \varepsilon \tag{13.14}$$

where z^{*} is the optimal value of the LP in step 3 and ε is a very small positive number.

The number of LPs that have to be solved in this step (and the corresponding value functions obtained) is $2 \cdot \sum_{i=1}^{n} (\alpha_i - 1)$. In most of the UTASTAR applications one usually seeks value functions that are free of errors (all errors variables σ are zero) and no relaxation from the minimal error is allowed ($\varepsilon = 0$).

13.3 Interactive Disaggregation and Robustness Control

13.3.1 Bipolar Robustness Control

The UTASTAR inference engine shows that the DM's preference model may not be a unique additive value function but a set of functions, all being compatible with the holistic preference statements provided to the analyst. This infinite set of functions comprises a polyhedral set, confined under some linear constraints, in the $\sum_{i=1}^{n} (\alpha_i - 1)$ dimension space, where a_i is the number of points on which the value function u_i is assessed.

Greco et al. (2010) proposed a general methodological framework, named Robust Ordinal Regression (ROR), which can be implemented synergistically to the disaggregation methods and aims at enhancing the robustness of the estimated results. ROR is based on the principle, according to which the decisions and proposals emerge after considering all those parameters that are compatible with the preferences of the DM. This principle contradicts the theoretical approaches of many MCDA methods, which select only specific parameters for the estimation of the results. The latter is considered theoretically arbitrary and at the same time excludes potential additional information regarding the whole set of alternative actions. On the other hand, ROR considers all value functions, which are consistent with the information provided by the DM and calculates necessary weak and possible weak preference relations. The former preference relation holds when an alternative is at least as good as another one for all instances, which are compatible with the DM's preference information, and the latter when there exists at least one instance that an alternative is at least as good as another one.

Recently Siskos and Psarras (2016) proposed an interactive bipolar robustness control, which manages robustness in both phases/poles of the interactive decision support process, namely the disaggregation and the aggregation one. Through

this integrated procedure, the analyst has the possibility to examine, measure and analyze in a systematic way the robustness of the decision model's parameters and the results that emerge after the implementation of the additive value model. Although bipolar robustness control is coupled perfectly to the UTA-type methods, it can be just as well implemented under a synergy with several other MCDA methods.

Regarding the family of UTA methods, the robustness control process is initiated after the inference of the additive value model, which leads to the ranking of the reference actions. It then proceeds to the analysis of the robustness of the model, with the option of discontinuing the modeling process, if the results are not satisfactory. In this case, the analyst asks the DM to enrich the reference set with additional reference actions or add other new preference statements. Ciomek et al. (2016) note that this additional information can impose excessive cognitive burden on the DM and they proposed heuristics for prioritizing pairwise elicitation questions.

In the reverse direction, the process moves from the disaggregation to the aggregation pole, where the MCDA model is implemented and the ranking of the real actions is achieved. Robustness is again measured in this pole, in terms of the stability of the ranking positions of each action. If the robustness of the results is adequate enough to support a sound decision, the algorithm ends, otherwise the analyst returns to the disaggregation pole and asks the DM for the acquisition of additional preferential information. Figure 13.2 illustrates the algorithm that an analyst may apply during the implementation of bipolar robustness control.

The robustness control framework, when coupled with any UTA family method, uses two separate sets of robustness indices to judge: (1) the efficacy of the additive model in the disaggregation pole and (2) the robustness of the final results, achieved after the extrapolation of the model on the whole set A, in the aggregation pole. The calculation of these indices requires the implementation of certain techniques and standalone methods, in parallel with the decision support procedure.

13.3.2 Robustness Indices

The indices, related to the disaggregation pole of the robustness control framework, focus on the efficacy/stability of the model to produce results that are stable and not misleading or ambiguous. The objective of these indices is to build a robust decision model that accurately reflects the preferences of the DM. On top of that, these indices have a practical meaning, since they prevent the analyst from performing heavy, pointless computations, which are certain to reach results of low quality. The whole computational effort is therefore decreased and the goals of the DM are reached by spending fewer resources.

The robustness indices proposed by Siskos and Psarras (2016) are categorized, based on which pole they apply to. Certain representative indices, which are also calculated in the application example of Sect. 13.4, are presented below.

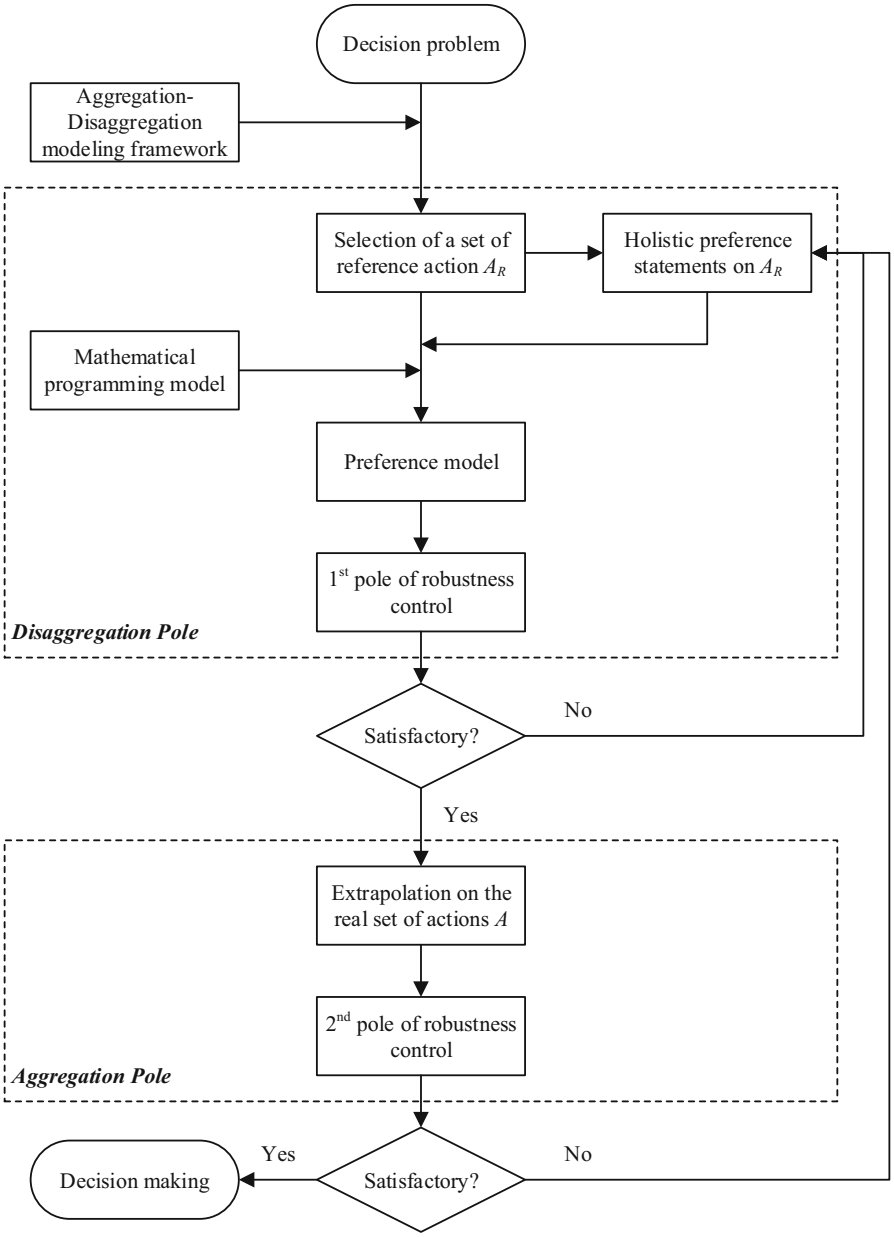


Fig. 13.2 Algorithm of the bipolar robustness control procedure (Stavrou et al. 2018)

13.3.2.1 Robustness Indices on the Disaggregation Pole

Two indices can be recognized in this category. The use of these indices presupposes the production of multiple sets of preferential parameters. A usual way to achieve this, when implementing the UTA-type methods is the max–min LPs technique (see for example step 4 in Sect. 13.2.2). During this procedure, all or a subset of model's parameters are successively minimized and maximized, under the set of feasibility constraints, and then visualized.

Let p_{rs} denote the set of the model's parameters produced by a robustness disaggregation technique, where r denoted a specific instance in which the parameter is estimated ($r = 1, 2, \ldots, R$) and s denotes a specific parameter ($s = 1, 2, \ldots, S$). For example, in the proposed approach, where $u_i\left(g_i^j\right)$ are examined during step 4 of the UTASTAR algorithm, the number of instances is $R = 2\sum_{i=1}^{n}(\alpha_i - 1)$ and the number of parameters is $S = \sum_{i=1}^{n}(\alpha_i - 1)$. In the case of the UTA II method, where only $u_i\left(g_i^*\right)$ are minimized and maximized we have $R = 2n$ and $S = n$ when $u_i\left(g_i^*\right)$ are examined or $S = \sum_{i=1}^{n}(\alpha_i - 1)$ when $u_i\left(g_i^j\right)$ are examined.

Average Range of the Preferential Parameters (ARP)

This index reveals the range of an average preferential parameter, after considering the preference information extracted by the DM. The calculation of the *ARP* requires the a priori implementation of the max–min LPs technique and is defined as follows:

$$ARP = \frac{1}{S}\sum_{s=1}^{S}\left[\max_{r}(p_{rs}) - \min_{r}(p_{rs})\right] \tag{13.15}$$

where p_{rs} is the r-th instance of the s-th preferential parameter and S is the number of all the different instances considered during the max–min LPs procedure.

This index ranges in $[0, 1]$ and receives lower values, when the robustness of a model increases. *ARP* receives the value of 0 when a unique preference model reflects the preference statements of the DM.

Average Stability Index (ASI)

The average stability index is a robustness index proposed by Siskos and Grigoroudis (2010) and indicates the average value of the normalized standard deviation of the preferential parameters (see also Grigoroudis and Siskos 2010 for a theoretical discussion and Delias et al. 2013b and Delias and Matsatsinis 2013 for some indicative applications). *ASI* is assessed as:

$$ASI = 1 - \frac{1}{S}\sum_{s=1}^{S}\sqrt{\frac{R\sum_{r=1}^{R}p_{rs}^2 - \left(\sum_{r=1}^{R}p_{rs}\right)^2}{R\sum_{r=1}^{R}p'^2_{rs} - \left(\sum_{r=1}^{R}p'_{rs}\right)^2}} \tag{13.16}$$

where p'_{rs} is the possible value of the r-th instance of the s-th preferential parameter that maximizes the variance of a particular parameter during the max–min LPs procedure. *ASI* also ranges in $[0, 1]$ and returns the value of 1 when perfect robustness is achieved.

In this chapter, *ASI* has the following form:

$$ASI = 1 - \frac{1}{\sum_{i=1}^{n}(\alpha_i - 1)}\sum_{i=1}^{n}\sum_{j=2}^{\alpha_i}\frac{\sqrt{\sum_{i=1}^{n}(\alpha_i - 1)\sum_{r=1}^{R}\left(u_{ij}^r\right)^2 - \left(\sum_{r=1}^{R}u_{ij}^r\right)^2}}{2\sqrt{\sum_{i=1}^{n}\alpha_i - (n+1)}} \tag{13.17}$$

where u_{ij}^r is the r-th instance of u_{ij} during the max–min LPs procedure with $R = 2\sum_{i=1}^{n}(\alpha_i - 1)$.

13.3.2.2 Robustness Indices on the Aggregation Pole

The exploitation of the indices related to the disaggregation pole offers a comprehensive view of the robustness of the decision model. However, this does not guarantee the acquisition of robust results after the implementation of the decision model. The proposition of appropriate indices in the aggregation pole (2nd pole) is therefore necessary. Again, these indices work under the condition that certain techniques are implemented.

Average Range of the Ranking (ARRI) and Ratio of the Average Range of the Ranking (RARR)

The average range of the ranking index and the ratio of the average range of the ranking are coupled with the Extreme Ranking Analysis technique proposed by Kadziński et al. (2012). *ARRI* depicts the possible number of positions that an average action can occupy in the whole ranking, while *RARR* reflects the ratio of the aforementioned deviation, with respect to the whole number of the alternatives under evaluation. The optimal values of *ARRI* and *RARR* are 1 and 0%, respectively, and they are calculated using the following formulae:

$$ARRI = \frac{1}{m}\sum_{k=1}^{m}\left(|R_*(k) - R^*(k)| + 1\right) \tag{13.18}$$

$$RARR = \frac{ARRI - 1}{m - 1} \cdot 100\% \tag{13.19}$$

Where $R_*(k)$ and $R^*(k)$ are the worst and best possible ranking positions, respectively for the k-th alternative and m is the number of reference actions.

Statistical Preference Relations Index (SPRI)

The statistical preference relations index (*SPRI*) offers a comprehensive way to examine the stability of all the ranking positions achieved by the whole set of alternatives. It is performed in synergy with random sampling techniques, the Manas-Nedoma algorithm (Manas and Nedoma 1968) that extracts all the vertices of the model's polyhedron, and generally methods that provide a statistically adequate number of sets of preferential parameters, within the feasible area. *SPRI* calculates the separate probabilities, that each alternative occupies a single ranking position in the final ranking, and forms a meaningful measure, which gives a clear insight of the robustness of the final results.

Specifically, the estimation of the probability that an alternative a_k gets ranked in the t-th position is performed using the following relation:

$$P_t^k = \frac{c_t^k}{R} \cdot 100\% \tag{13.20}$$

where c_t^k is the number of samples/instances that position an alternative a_k in the t-th position ($t = 1, 2, \ldots, m$) and R is the number of all the samples/instances.

The statistical preference relations index is then calculated using the following equation:

$$SPRI = \frac{1}{R}\sum_{k=1}^{m}\sum_{t=1}^{m} P_t^k \tag{13.21}$$

SPRI reaches the optimal value of 100% when all the alternatives occupy a single ranking position with a probability of 100%. In other words, the same ranking exactly occurs for all the preferential parameters samples/instances under consideration.

13.4 An Application Example

13.4.1 Problem Presentation

The example presented in this section is inspired from a successful real world application of the UTA method in a job evaluation problem in a leading Greek organization (Spyridakos et al. 2000). Job evaluation is a systematic process that enables the design and establishment of human resources improvement procedures and fair reward systems.

In the organization under examination, job evaluation concerns the assessment of a value system that encapsulates the importance of the parameters that reflect the global responsibility and duties of each different job position. It should be noted that this evaluation does not concern the real persons in these positions, but the jobs themselves, the responsibilities associated with them, and their contribution to productivity and profitability. The evaluation positively influences the competence and performance management, since it: (1) aids the establishment of a reward system that links the importance of the jobs to the payment offered, and (2) supports the design of human resources development requirements, in order to improve the effectiveness of the positions' operations. Three evaluation criteria are the following (see details in Table 13.1):

- *Criterion* 1 (input criterion): Required qualifications and skills (i.e., basic knowledge, expertise, skills, experience), measurable in the numerical scale (5, 20).
- *Criterion* 2 (process criterion): Contribution to decision making (e.g., participation to committees, position in the hierarchy, problem solving, quantity and importance of the decisions), measured using an ordinal scale: (limited, medium, high, very high).
- *Criterion* 3 (output criterion): Responsibility (e.g., qualitative results, geographical area, degree of responsibility, perspectives, strategic role in development activities, and support to other units), measured using an ordinal scale: (limited, medium, high, very high).

Using the aforementioned criteria, ten job positions are evaluated, as shown in Table 13.2.

Table 13.1 Job evaluation criteria for the application example

Criteria name	Point of view	Type	Evaluation scale
g_1: Required qualifications and skills	Input	Measurable	Numerical scale (5, 20)
g_2: Contribution to decision making	Process	Ordinal	(limited, medium, high, very high)
g_3: Responsibility	Output	Ordinal	(limited, medium, high, very high)

Table 13.2 Multicriteria evaluation of ten job positions

Job position	Criterion 1 (required qualifications and skills)	Criterion 2 (contribution to decision making)	Criterion 3 (responsibility)
A	7	Medium	High
B	12	High	Medium
C	15	Limited	Limited
D	5	Medium	Medium
E	10	Limited	Very high
F	19	Very high	Limited
G	12	Limited	High
H	8	High	High
I	16	Limited	Medium
J	6	Medium	Very high

13.4.2 Reference Set and Preference Elicitation

The decision analyst develops a dialogue with the DM in order to construct a reference set of job positions and help the DM to articulate his preference statements. An excerpt of the dialogue between the analyst and the DM is the following:

Analyst: Let's take the job position E which requires graduate studies but no experience and no special skills. According to the job description, this position does not require the participation to committees but has very high responsibilities in the organization (see Table 13.2). Comparing to a fictitious job position, Z_1, which has the same responsibilities, a high contribution to decision processes and requires only a high school degree ($g_1 = 5$), which one is globally most important for the organization?

DM: I think the second one is most important.

Analyst: Let's compare now the same job position E to a new fictitious job, namely Z_2, which requires the same qualifications and has a high contribution to decision processes and high responsibilities. Which one is globally most important for the organization?

DM: It seems to me that the two jobs are globally equivalent.

Analyst: Would you now rate a fictitious job position, namely Z_3, which requires significant qualifications and skills ($g_1 = 15$) but with medium scoring to both the contribution to decision making and the responsibility?

DM: In my opinion this position is clearly inferior to job position E.

Analyst: Perfect. Let's summarize the comparisons. Your complete ranking of the four jobs is the one that appears in Table 13.3, right?

DM: Yes.

Consequently, the constructed reference set of reference jobs includes one real job position from the set A and three fictitious job positions, i.e., $A_R = \{Z_1, E, Z_2, Z_3\}$ as presented in Table 13.3.

Table 13.3 DM's ranking of the four reference job positions

Reference job position	Criterion 1 (required qualifications and skills)	Criterion 2 (contribution to decision making)	Criterion 3 (responsibility)	Ranking position
Z_1	5	High	Very high	1
E	10	Limited	Very high	2
Z_2	10	High	High	2
Z_3	15	Medium	Medium	4

13.4.3 Preference Disaggregation Using UTASTAR Method

13.4.3.1 Step 1

According to the first step of the UTASTAR algorithm, the following expressions are calculated:

$$\begin{aligned} u[\mathbf{g}(Z_1)] &= u_1(5)+u_2(\text{high})+u_3(\text{very high}) = 0+(w_{21}+w_{22})+(w_{31}+w_{32}+w_{33}) \\ &= w_{21}+w_{22}+w_{31}+w_{32}+w_{33} \end{aligned}$$

$$\begin{aligned} u[\mathbf{g}(E)] &= u_1(10)+u_2(\text{limited})+u_3(\text{very high}) = w_{11}+0+(w_{31}+w_{32}+w_{33}) \\ &= w_{11}+w_{31}+w_{32}+w_{33} \end{aligned}$$

$$\begin{aligned} u[\mathbf{g}(Z_2)] &= u_1(10)+u_2(\text{high})+u_3(\text{high}) = w_{11}+(w_{21}+w_{22})+(w_{31}+w_{32}) \\ &= w_{11}+w_{21}+w_{22}+w_{31}+w_{32} \end{aligned}$$

$$\begin{aligned} u[\mathbf{g}(Z_3)] &= u_1(15)+u_2(\text{medium})+u_3(\text{medium}) = (w_{11}+w_{12})+w_{21}+w_{31} \\ &= w_{11}+w_{12}+w_{21}+w_{31} \end{aligned}$$

13.4.3.2 Step 2

For each pair of consecutive actions in the ranking, the following differences are obtained:

$$\begin{aligned} \Delta(Z_1,E) &= u[\mathbf{g}(Z_1)]-\sigma^+(Z_1)+\sigma^-(Z_1)-u[\mathbf{g}(E)]+\sigma^+(E)-\sigma^-(E) \\ &= (w_{21}+w_{22}+w_{31}+w_{32}+w_{33})-\sigma^+(Z_1) \\ &\quad +\sigma^-(Z_1)-(w_{11}+w_{31}+w_{32}+w_{33})+\sigma^+(E)-\sigma^-(E) \\ &= -w_{11}+w_{21}+w_{22}-\sigma^+(Z_1)+\sigma^-(Z_1)+\sigma^+(E)-\sigma^-(E) \end{aligned}$$

$$\begin{aligned} \Delta(E,Z_2) &= u[\mathbf{g}(E)]-\sigma^+(E)+\sigma^-(E)-u[\mathbf{g}(Z_2)]+\sigma^+(Z_2)-\sigma^-(Z_2) \\ &= (w_{11}+w_{31}+w_{32}+w_{33})-\sigma^+(E)+\sigma^-(E) \\ &\quad -(w_{11}+w_{21}+w_{22}+w_{31}+w_{32})+\sigma^+(Z_2)-\sigma^-(Z_2) \\ &= -w_{21}-w_{22}+w_{33}-\sigma^+(E)+\sigma^-(E)+\sigma^+(Z_2)-\sigma^-(Z_2) \end{aligned}$$

$$\begin{aligned}\Delta(Z_2, Z_3) &= u[\mathbf{g}(Z_2)] - \sigma^+(Z_2) + \sigma^-(Z_2) - u[\mathbf{g}(Z_3)] + \sigma^+(Z_3) - \sigma^-(Z_3)\\ &= (w_{11} + w_{21} + w_{22} + w_{31} + w_{32}) - \sigma^+(Z_2) + \sigma^-(Z_2)\\ &\quad - (w_{11} + w_{12} + w_{21} + w_{31}) + \sigma^+(Z_3) - \sigma^-(Z_3)\\ &= -w_{12} + w_{22} + w_{32} - \sigma^+(Z_2) + \sigma^-(Z_2) + \sigma^+(Z_3) - \sigma^-(Z_3)\end{aligned}$$

13.4.3.3 Step 3

The following LP is solved:

$$\begin{aligned}[\min] z = {} & \sigma^+(Z_1) + \sigma^-(Z_1) + \sigma^+(E) + \sigma^-(E) + \sigma^+(Z_2) + \sigma^-(Z_2)\\ & + \sigma^+(Z_3) + \sigma^-(Z_3)\end{aligned}$$

Subject to:

$$\begin{aligned}&-w_{11} + w_{21} + w_{22} - \sigma^+(Z_1) + \sigma^-(Z_1) + \sigma^+(E) - \sigma^-(E) \geq 0.05\\ &-w_{21} - w_{22} + w_{33} - \sigma^+(E) + \sigma^-(E) + \sigma^+(Z_2) - \sigma^-(Z_2) = 0\\ &-w_{12} + w_{22} + w_{32} - \sigma^+(Z_2) + \sigma^-(Z_2) + \sigma^+(Z_3) - \sigma^-(Z_3) \geq 0.05\\ &w_{11} + w_{12} + w_{13} + w_{21} + w_{22} + w_{23} + w_{31} + w_{32} + w_{33} = 1\\ &w_{11}, w_{12}, w_{13}, w_{21}, w_{22}, w_{23}, w_{31}, w_{32}, w_{33} \geq 0\\ &\sigma^+(Z_1), \sigma^-(Z_1), \sigma^+(E), \sigma^-(E), \sigma^+(Z_2), \sigma^-(Z_2), \sigma^+(Z_3), \sigma^-(Z_3) \geq 0\end{aligned}$$

where δ is initially set to 0.05.

13.4.3.4 Step 4

The previous LP has a zero error solution ($z = 0$), which means that there exists at least one additive value function that is fully compatible with the DM's ranking of the four reference jobs. The results, obtained after the first solution of the LP in Step 3, appear in the first row of Table 13.4. All the results have been rounded to the third decimal place.

Next, according to step 4 of the UTASTAR algorithm, the analyst seeks for a set of $2 \times (3 + 3 + 3) = 18$ extreme solutions of the solution polyhedral set, by successively solving the LPs of the following type:

$$[\max] \text{ or } [\min] \sum_{t=1}^{j-1} w_{it} \text{ for } i = 1, 2, 3 \text{ and } j = 2, 3, 4$$

(see Table 13.4, rows 3–20).

Table 13.4 Value function solutions of the UTASTAR method

Type of solution	w_{11}	w_{12}	w_{13}	w_{21}	w_{22}	w_{23}	w_{31}	w_{32}	w_{33}
$\delta = 0.05$	0.3	0	0	0.3	0.05	0	0	0	0.35
[min]w_{11}	0	0	0	0.45	0.05	0	0	0	0.5
[max]w_{11}	0.3	0	0	0.3	0.05	0	0	0	0.35
[min]$w_{11} + w_{12}$	0	0	0	0.45	0.05	0	0	0	0.5
[max]$w_{11} + w_{12}$	0.225	0.225	0	0	0.275	0	0	0	0.275
[min]$w_{11} + w_{12} + w_{13}$	0	0	0	0.45	0.05	0	0	0	0.5
[max]$w_{11} + w_{12} + w_{13}$	0	0	0.9	0	0.05	0	0	0	0.05
[min]w_{21}	0.3	0	0	0	0.35	0	0	0	0.35
[max]w_{21}	0	0	0	0.475	0	0	0	0.05	0.475
[min]$w_{21} + w_{22}$	0	0.425	0	0.05	0	0	0	0.475	0.05
[max]$w_{21} + w_{22}$	0	0	0	0.45	0.05	0	0	0	0.5
[min]$w_{21} + w_{22} + w_{23}$	0	0.425	0	0.05	0	0	0	0.475	0.05
[max]$w_{21} + w_{22} + w_{23}$	0	0	0	0	0.05	0.9	0	0	0.05
[min]w_{31}	0.3	0	0	0.3	0.05	0	0	0	0.35
[max]w_{31}	0	0	0	0	0.05	0	0.9	0	0.05
[min]$w_{31} + w_{32}$	0.3	0	0	0.3	0.05	0	0	0	0.35
[max]$w_{31} + w_{32}$	0	0	0	0	0.05	0	0.9	0	0.05
[min]$w_{31} + w_{32} + w_{33}$	0	0	0.9	0	0.05	0	0	0	0.05
[max]$w_{31} + w_{32} + w_{33}$	0	0	0	0	0.05	0	0.9	0	0.05
Barycenter[a]	**0.079**	**0.060**	**0.100**	**0.182**	**0.071**	**0.050**	**0.150**	**0.056**	**0.253**

[a]Average of the 18 solutions of the post-optimality analysis step

Subject to:

$$
\begin{aligned}
&-w_{11} + w_{21} + w_{22} \geq 0.05\\
&-w_{21} - w_{22} + w_{33} = 0\\
&-w_{12} + w_{22} + w_{32} \geq 0.05\\
&w_{11} + w_{12} + w_{13} + w_{21} + w_{22} + w_{23} + w_{31} + w_{32} + w_{33} = 1\\
&w_{11}, w_{12}, w_{13}, w_{21}, w_{22}, w_{23}, w_{31}, w_{32}, w_{33} \geq 0
\end{aligned}
$$

All the obtained solutions are listed in the rows 3–20 of Table 13.4, while the barycenter (average solution) appears in the last row (row 21). The maximum and minimum possible, and the barycentric marginal value functions are summarized in Fig. 13.3.

13.4.4 Bipolar Robustness Control

The implementation of the UTASTAR procedure reveals results of significantly low quality with regard to their robustness; no decision on the ranking of the ten job positions can therefore be supported at this current stage of the analysis.

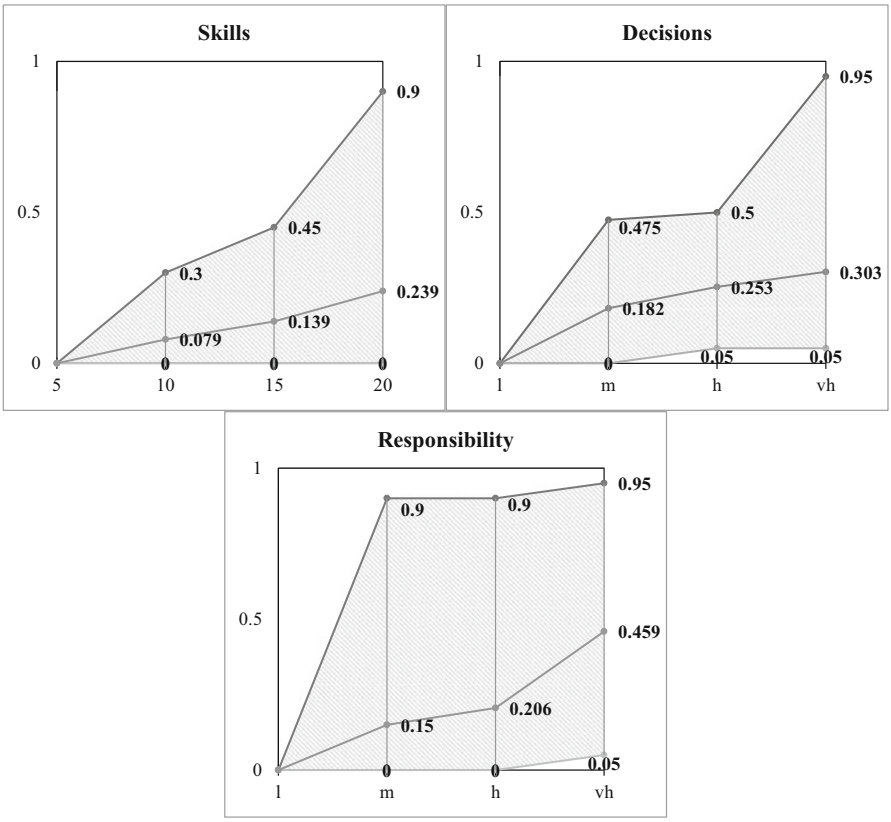


Fig. 13.3 Visualization of the variation of the three additive value functions (maximum, barycenter and minimum) from the application of the bipolar control iterations (1st iteration)

Specifically, the *ASI* index takes the value of 0.733, while the average range of the preferential parameters (*ARP*) is 0.686 (i.e., equals to 68.6% of their whole possible ranging area). In particular, certain parameters, such as w_{13}, w_{23}, and w_{31} can range from 0 to 0.9, being in essence uncontrollable.

Consequently, the bipolar robustness control procedure does not allow us to move to the aggregation pole (2nd pole of robustness control).

13.4.4.1 UTASTAR Re-Activation (2nd Iteration)

The analyst decides to ask new preference statements from the DM with a view to ameliorating the robustness of the results. In this procedure, care should be taken in order to ensure that the new preference information is consistent with the old preferential statements. The following dialogue excerpt is characteristic:

Analyst: It seems that the mathematical input required by the method is not sufficient for a good assessment of your preference model. Would you please insert to your ranking of Table 13.3 a highly qualified job (17 points), named Z_4, with a "very high" contribution to the decision making processes but without any important responsibility (limited)?

DM: I would rate this job fourth, between Z_2 and Z_3.

Accordingly, the analyst tests the compatibility of the fictitious job Z_4 with the DM's former ranking, before proceeding to the common calculations. Due to the increasing number of reference alternatives, the analyst decided to decrease and stabilize the value of δ to 0.01.

13.4.4.2 Step 1

$$\begin{aligned} u\left[\mathbf{g}\left(Z_4\right)\right] &= u_1(17) + u_2\left(\text{very high}\right) + u_3\left(\text{limited}\right) = \left(w_{11} + w_{12} + 0.4w_{13}\right) \\ &\quad + \left(w_{21} + w_{22} + w_{23}\right) + 0 \\ &= w_{11} + w_{12} + 0.4w_{13} + w_{21} + w_{22} + w_{23} \end{aligned}$$

It should be noted that $u_1(17)$ is calculated using linear interpolation in (15, 20). More specifically, applying formula (Eq. (13.9)) we have:

$$u_1(17) = \sum_{t=1}^{2} w_{1t} + \frac{17-15}{20-15} w_{13} = w_{11} + w_{12} + 0.4w_{13}$$

13.4.4.3 Step 2

$$\begin{aligned} &\Delta\left(Z_2, Z_4\right) = u\left[\mathbf{g}\left(Z_2\right)\right] - u\left[\mathbf{g}\left(Z_4\right)\right] + \sigma^+\left(Z_4\right) - \sigma^-\left(Z_4\right) \\ &= \left(w_{11} + w_{21} + w_{22} + w_{31} + w_{32}\right) - \left(w_{11} + w_{12} + 0.4w_{13} + w_{21} + w_{22} + w_{23}\right) \\ &\quad + \sigma^+\left(Z_4\right) - \sigma^-\left(Z_4\right) \\ &= -w_{12} - 0.4w_{13} - w_{23} + w_{31} + w_{32} + \sigma^+\left(Z_4\right) - \sigma^-\left(Z_4\right) \end{aligned}$$

$$\begin{aligned} &\Delta\left(Z_4, Z_3\right) = u\left[\mathbf{g}\left(Z_4\right)\right] - \sigma^+\left(Z_4\right) + \sigma^-\left(Z_4\right) - u\left[\mathbf{g}\left(Z_3\right)\right] \\ &= \left(w_{11} + w_{12} + 0.4w_{13} + w_{21} + w_{22} + w_{23}\right) - \sigma^+\left(Z_4\right) + \sigma^-\left(Z_4\right) \\ &\quad - \left(w_{11} + w_{12} + w_{21} + w_{31}\right) \\ &= 0.4w_{13} + w_{22} + w_{23} - w_{31} - \sigma^+\left(Z_4\right) + \sigma^-\left(Z_4\right) \end{aligned}$$

Table 13.5 Values of the four new fictitious jobs in the 3rd iteration

Reference job position	Criterion 1 (required qualifications and skills)	Criterion 2 (contribution to decision making)	Criterion 3 (responsibility)
Y_1	5	Medium	Limited
Y_2	?	Limited	Limited
Y_3	?	Limited	Limited
Y_4	5	Limited	Medium

13.4.4.4 Step 3

$$[\min] z' = \sigma^+ (Z_4) + \sigma^- (Z_4)$$

Subject to:

$$\begin{aligned}
&-w_{11} + w_{21} + w_{22} \geq 0.01\\
&-w_{21} - w_{22} + w_{33} = 0\\
&-w_{12} - 0.4w_{13} - w_{23} + w_{31} + w_{32} + \sigma^+ (Z_4) - \sigma^- (Z_4) \geq 0.01\\
&0.4w_{13} + w_{22} + w_{23} - w_{31} - \sigma^+ (Z_4) + \sigma^- (Z_4) \geq 0.01\\
&w_{11} + w_{12} + w_{13} + w_{21} + w_{22} + w_{23} + w_{31} + w_{32} + w_{33} = 1\\
&w_{11}, w_{12}, w_{13}, w_{21}, w_{22}, w_{23}, w_{31}, w_{32}, w_{33}, \sigma^+ (Z_4), \sigma^- (Z_4) \geq 0
\end{aligned}$$

The solution of the LP shows that the DM's ranking is cohesive. However, the min–max procedure that follows reveals that the *ASI* index is still very low (0.772), while *ARP* is 0.567, which equals to 56.7% of their whole possible ranging area (see the corresponding value functions in Fig. 13.4). The negligible increase of the *ASI* index, shows that the new alternative added to the ranking, reduced slightly the volume of the hyperpolyhedron of the feasible solutions. Consequently, the analyst is forced to return and ask the DM for additional preference information.

13.4.4.5 2nd Request for Feedback (3rd Iteration)

Analyst: It seems that your preference model is still not adjusted in a robust way. I would propose adding some additional preference information in a different way this time, according to the MAUT (Multiattribute Utility Theory) questioning paradigm. Suppose you have the four fictitious jobs, namely Y_1, Y_2, Y_3, and Y_4, as shown in Table 13.5. Comparing Y_1 and Y_2 what qualification degree is required for the job Y_2 to compensate exactly the difference "medium"–"limited" on decisions?

DM: I suppose 17.

Analyst: Ok. So, in this way, you are indifferent between Y_1 and Y_2. Right?

DM: Right.

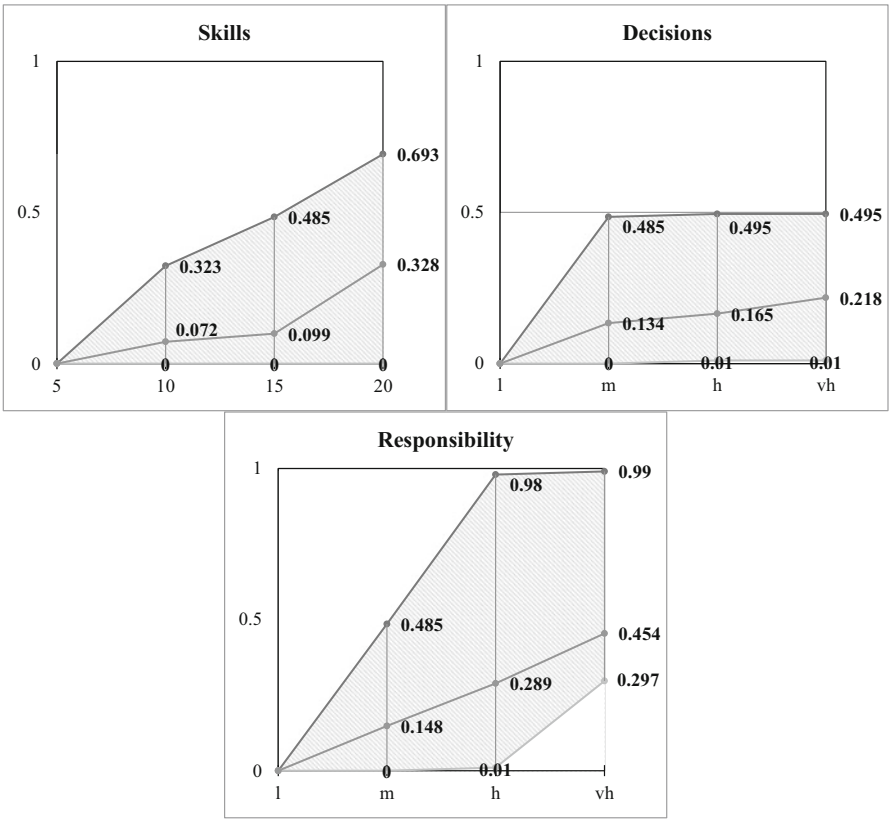


Fig. 13.4 Visualization of the variation of the three additive value functions (maximum, barycenter and minimum) from the application of the bipolar control iterations (2nd iteration)

Analyst: Let's compare now Y_3 and Y_4. What qualification degree is required for the job Y_3 to exactly compensate the difference "medium"–"limited" on responsibility?

DM: I would say 12 units.

Based on the previous, two separate indifference comparisons are created now: $Y_1 \sim Y_2$ and $Y_3 \sim Y_4$ or equivalently:

$$u\left[\mathbf{g}\left(Y_1\right)\right] = u\left[\mathbf{g}\left(Y_2\right)\right] \quad \text{and} \quad u\left[\mathbf{g}\left(Y_3\right)\right] = u\left[\mathbf{g}\left(Y_4\right)\right]$$

where

$$u\left[\mathbf{g}\left(Y_1\right)\right] = u_1(5) + u_2\,(\text{medium}) + u_3\,(\text{limited}) = 0 + w_{21} + 0 = w_{21}$$

$$u\left[\mathbf{g}\left(Y_2\right)\right] = u_1(17) + u_2\,(\text{limited}) + u_3\,(\text{limited})$$

$$= w_{11} + w_{12} + 0.4w_{13} + 0 + 0 = w_{11} + w_{12} + 0.4w_{13}$$

$$u\left[\mathbf{g}\left(\mathrm{Y}_3\right)\right] = u_1(12) + u_2\left(\text{limited}\right) + u_3\left(\text{limited}\right) = w_{11} + 0.4w_{12} + 0 + 0$$
$$= w_{11} + 0.4w_{12}$$

$$u\left[\mathbf{g}\left(\mathrm{Y}_4\right)\right] = u_1(5) + u_2\left(\text{limited}\right) + u_3\left(\text{medium}\right) = 0 + 0 + w_{31} = w_{31}$$

Consequently, the required constraints are:

$$w_{21} - w_{11} - w_{12} - 0.4w_{13} = 0$$

$$w_{11} + 0.4w_{12} - w_{31} = 0$$

Two couples of positive errors are introduced to these equations, before adding them to the previously constructed compatible system of four preference relations, as follows:

$$w_{21} - w_{11} - w_{12} - 0.4w_{13} - \lambda_1^+ + \lambda_1^- = 0$$

$$w_{11} + 0.4w_{12} - w_{31} - \lambda_2^+ + \lambda_2^- = 0$$

Therefore, the UTASTAR enforcement LP formulation becomes as follows:

$$[\min]\, z'' = \lambda_1^+ + \lambda_1^- + \lambda_2^+ + \lambda_2^-$$

Subject to:

$$\begin{aligned}
&-w_{11} + w_{21} + w_{22} \geq 0.01\\
&-w_{21} - w_{22} + w_{33} = 0\\
&-w_{12} - 0.4w_{13} - w_{23} + w_{31} + w_{32} \geq 0.01\\
&0.4w_{13} + w_{22} + w_{23} - w_{31} \geq 0.01\\
&w_{11} + w_{12} + w_{13} + w_{21} + w_{22} + w_{23} + w_{31} + w_{32} + w_{33} = 1\\
&w_{21} - w_{11} - w_{12} - 0.4w_{13} - \lambda_1^+ + \lambda_1^- = 0\\
&w_{11} + 0.4w_{12} - w_{31} - \lambda_2^+ + \lambda_2^- = 0\\
&w_{ij} \geq 0 \forall i,j,\ \ \lambda_1^+, \lambda_1^-, \lambda_2^+, \lambda_2^- \geq 0
\end{aligned}$$

The optimal values of the error variables are zero and the results obtained after following the min–max procedure are presented in Fig. 13.5. The *ASI* index again increases and receives the value of 0.795, while *ARP* is 0.427. Nevertheless, the analyst is not confident to proceed to the aggregation pole and asks again for DM's feedback.

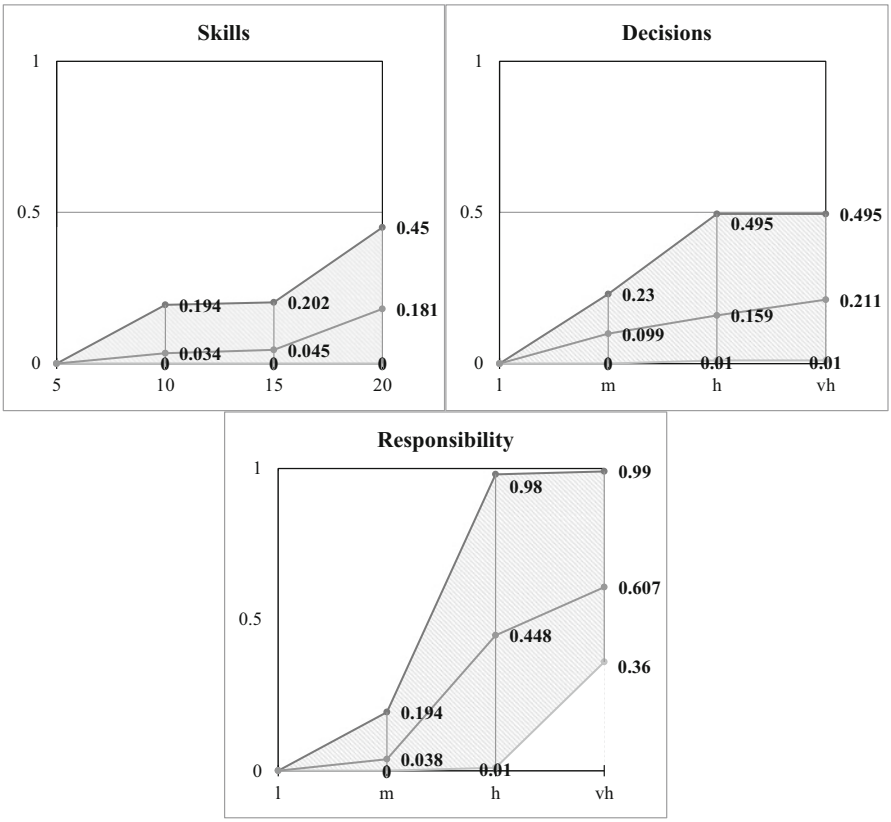


Fig. 13.5 Visualization of the variation of the three additive value functions (maximum, barycenter and minimum) from the application of the bipolar control iterations (3rd iteration)

Table 13.6 Values of the two new fictitious jobs in the 4th iteration

Reference job position	Criterion 1 (required qualifications and skills)	Criterion 2 (contribution to decision making)	Criterion 3 (responsibility)
Y_5	?	Medium	High
Y_6	5	High	Very high

13.4.4.6 3rd Request for Feedback (4th Iteration)

In the 3rd request for feedback, the analyst provides the DM with two additional reference job positions, as presented in Table 13.6. The following dialogue excerpt arises:

Analyst: Comparing Y_5 and Y_6 what qualification degree (if any) is required for the job Y_5 to compensate exactly the difference in the other two criteria?

DM: I would say the perfect score of 20.

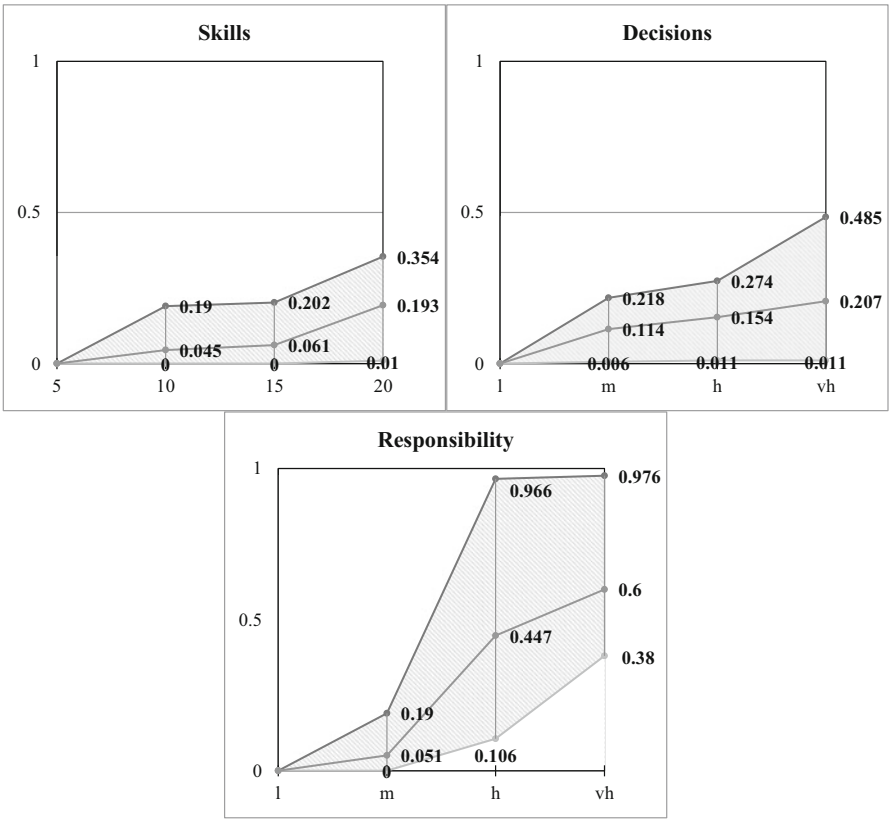


Fig. 13.6 Visualization of the variation of the three additive value functions (maximum, barycenter and minimum) from the application of the bipolar control iterations (4th iteration)

The DM's answer reveals the following new UTASTAR constraint in the form of an equation:

$$w_{11} + w_{12} + w_{13} - w_{22} - w_{33} = 0$$

The constraint is inserted to the previous set of preferential constraints, and the UTASTAR LP problem is validated for its cohesion (all over-under-estimation errors get zeroed). Next, the min–max procedure is executed and its results are presented in Fig. 13.6. The *ASI* index again increases and receives the value of 0.808, while *ARP* decreases significantly to 0.370.

At the current stage, the analyst decides to gain some insight on the robustness of the results that are obtained after the implementation of the additive value model. The extrapolation to the ten real job positions of Table 13.2 with the aid of the Extreme Ranking Analysis (ERA), is depicted in Fig. 13.7, which presents the

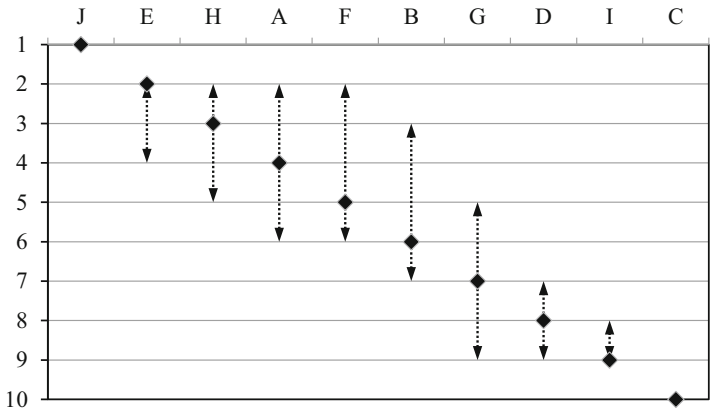


Fig. 13.7 Results of the Extreme Ranking Analysis in the 4th iteration

Table 13.7 Values of the two new fictitious jobs in the 5th iteration

Reference job position	Criterion 1 (required qualifications and skills)	Criterion 2 (contribution to decision making)	Criterion 3 (responsibility)
Y_7	15	High	?
Y_8	15	Very high	Medium

ranking of the ten real jobs in descending order (diamond dots), as well as the best and worst possible ranking position of each job, with the use of the two sided arrows.

The results obtained after the implementation of the ERA show that some significant instability exist, with regard to the ranking positions of the majority of the jobs, especially those that get ranked in the middle positions. Indicatively the *ARRI* is 3.4, while the *RARR* is 26.7%.

The analyst considers that the robustness of the ranking, obtained in the aggregation pole, is not acceptable and decides to revisit the disaggregation pole.

13.4.4.7 4th Request for Feedback (5th Iteration)

Analyst: Suppose you have the two new fictitious jobs Y_7 and Y_8 as shown in Table 13.7. What level of responsibility may compensate the gap "high"–"very high" on decision making?

DM: I think "high".

The next equation is therefore added to the previous set of constraints and the standard procedure is followed:

$$-w_{23} + w_{32} = 0$$

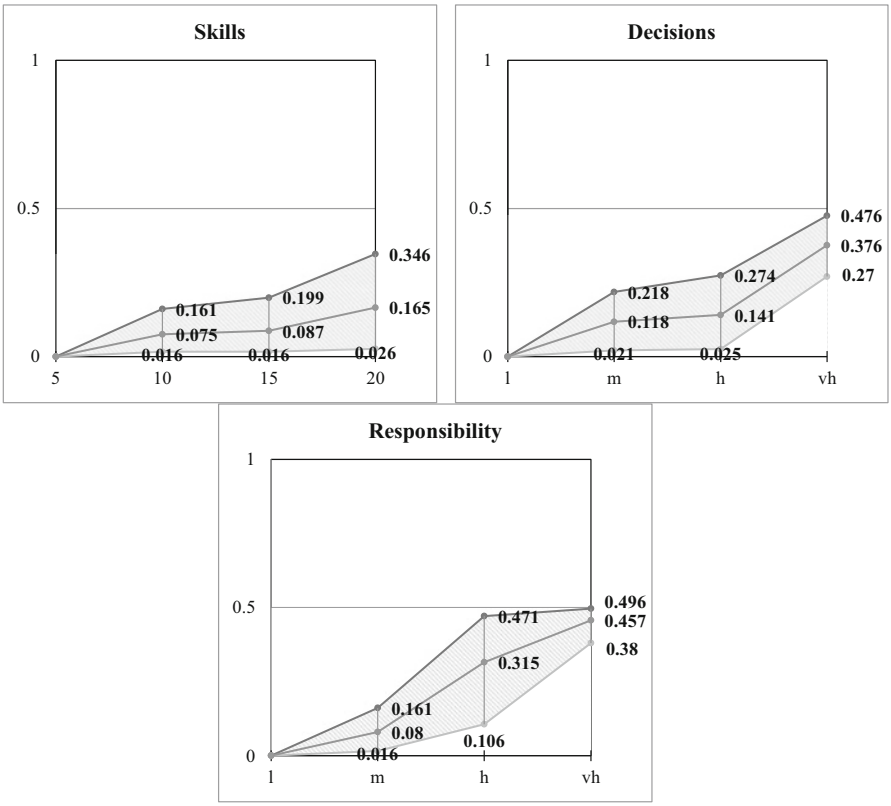


Fig. 13.8 Visualization of the variation of the three additive value functions (maximum, barycenter and minimum) from the application of the bipolar control iterations (5th iteration)

The *ASI* index again increases and receives the value of 0.898, while *ARP* is 0.214, which equals to 21.4% of their whole possible ranging area (see Fig. 13.8).

The analyst visits again the aggregation pole to rank the ten jobs (see Fig. 13.9). The obtained robustness indices after the implementation of the ERA are: $ARRI = 2.9$ and $RARR = 21.1\%$.

The ERA shows that the head and the tail of the ranking are stable, but significant instability remains regarding the middle positions in the ranking. Consequently, the analyst decides to make a last effort to increase robustness with regard to the jobs that belong to the in-between positions of the ranking.

13.4.4.8 5th Request for Feedback (6th Iteration)

The analyst discusses the previous ranking with the DM, in search of results that are not in convergence with the DM's viewpoints. As expected, the DM disagrees with the fact that job E can surpass job F, and he therefore demanded that this statement is reflected on a new constraint. Consequently:

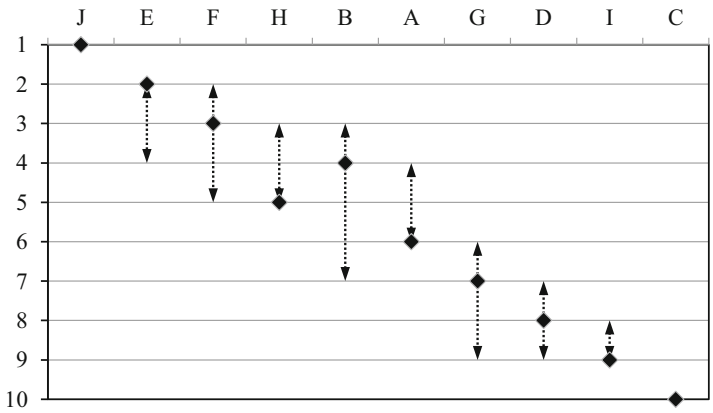


Fig. 13.9 Results of the Extreme Ranking Analysis in the 4th iteration

$$u\,[\mathbf{g}\,(\mathrm{F})] > u\,[\mathbf{g}\,(\mathrm{E})]$$

This preference provokes the following two inequalities, which replace the inequality $w_{11} + w_{21} + w_{22} \geq 0.01$, in order to assure the cohesion of the set of preference constraints:

$$w_{12} + 0.8w_{13} + w_{21} + w_{22} + w_{23} - w_{31} - w_{32} - w_{33} \geq 0.01$$

$$-w_{11} - w_{12} - 0.8w_{13} - w_{23} + w_{31} + w_{32} + w_{33} \geq 0.01$$

The new additive value model appears in Fig. 13.10. The *ASI* index increases again and receives the value of 0.909, while the *ARP* is 0.175, which equals to 17.5% of the whole possible ranging area. These results encourage the analyst to perform the ERA in the aggregation pole, the results of which are presented in Fig. 13.11.

After the visualization of the ERA, the DM endorses the adequacy of the results and decides to keep the final ranking as definitive. The ranking and the global values depicted at Table 13.8 have resulted after the implementation of the additive value model, using the barycentric value functions.

The decision support procedure, coupled with the bipolar robustness framework, ends at this stage. Table 13.9 depicts the evolution of the robustness indices, throughout the six iterations, along with their amelioration (in percentage) after each consecutive iteration.

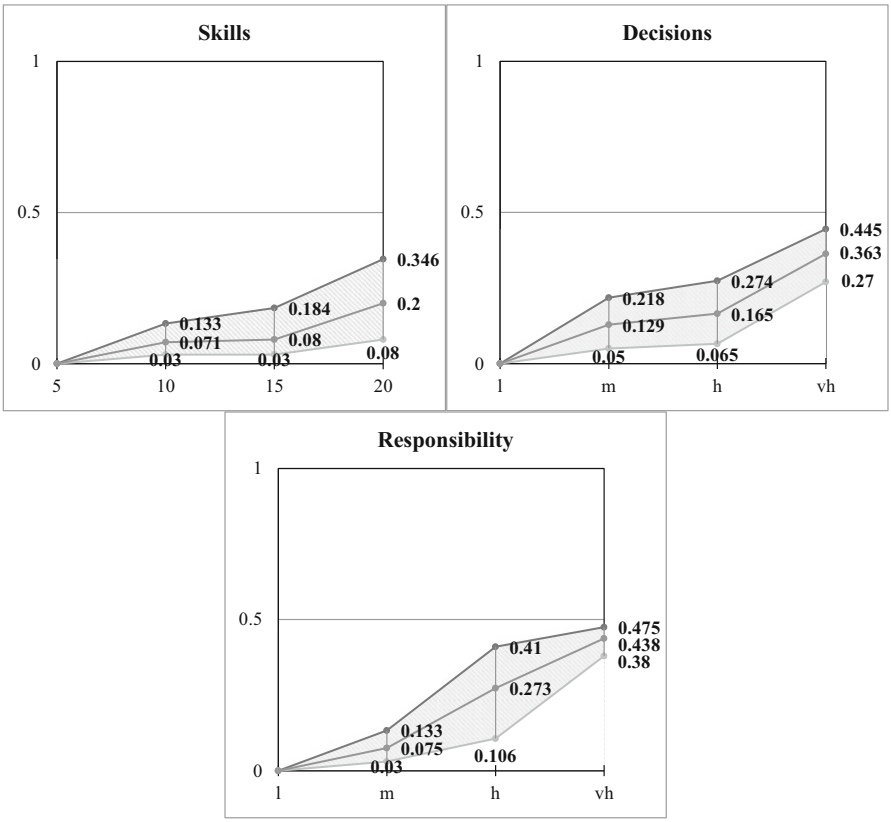


Fig. 13.10 Visualization of the variation of the three additive value functions (maximum, barycenter and minimum) from the application of the bipolar control iterations (6th iteration)

Table 13.8 Global evaluation of the ten job positions

Ranking	Job position	Global value
1	J	0.580
2	F	0.539
3	E	0.509
4	H	0.480
5	A	0.430
6	G	0.348
7	B	0.315
8	D	0.204
9	I	0.180
10	C	0.081

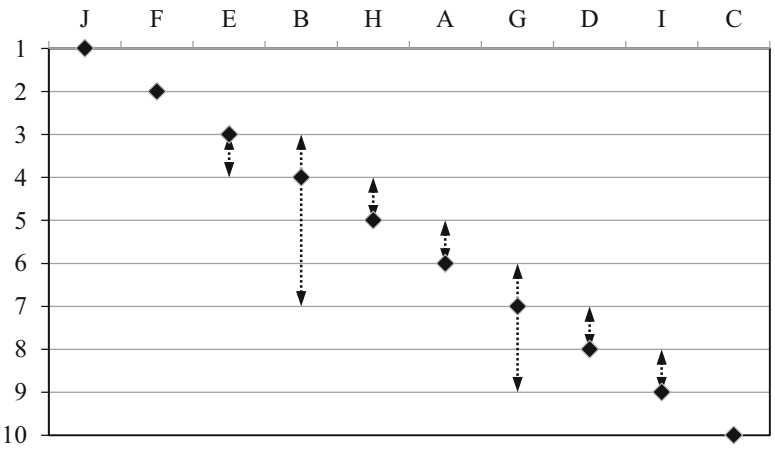


Fig. 13.11 Results of the Extreme Ranking Analysis in the 6th iteration

Table 13.9 Evolution of the robustness indices after each iteration (percentage amelioration in parentheses)

Iteration	ARP	ASI	ARRI	RARR
1	0.686	0.733	–	–
2	0.567 (17.3%)	0.772 (5.3%)	–	–
3	0.427 (24.8%)	0.795 (3.0%)	–	–
4	0.370 (13.3%)	0.808 (1.6%)	3.4	26.7%
5	0.214 (42.2%)	0.898 (11.1%)	2.9 (14.7%)	21.1% (21.0%)
6	0.175 (18.1%)	0.909 (1.2%)	2.2 (24.1%)	13.3% (37.0%)

13.5 Brief Overview of Existing Applications

The family of UTA-based methods is the main initiative and the most representative example of preference disaggregation theory. These methods cover different variants of the UTA approach (Siskos et al. 2016):

(a) Alternative optimality criteria (e.g., Kendall's tau between the estimated and the DM's ranking)
(b) Different forms of global preference (e.g., pairwise comparisons, intensity of DM's preferences, additional properties of the assessed value functions, construction of fuzzy outranking relations based on UTA's post-optimality analysis)
(c) Non-monotonic preferences (i.e., non-monotonic marginal value functions)
(d) Meta-UTA techniques (i.e., alternative post-optimality techniques)
(e) Stochastic UTA method, UTA II or UTA-type sorting approaches
(f) UTA-like multiobjective optimization approaches

UTA-based methods have been applied to several real-world decision-making problems, covering different fields:

- Job evaluation and human resources management (Spyridakos et al. 2000; González-Araya et al. 2002; Grigoroudis and Zopounidis 2012)
- Project evaluation (Jacquet-Lagrèze 1995; Beuthe et al. 2000)
- Environmental management (Siskos and Assimakopoulos 1989; Hatzinakos et al. 1991; Demesouka et al. 2013)
- Healthcare management (Manolitzas et al. 2013a, b; Doumpos et al. 2016)
- E-government evaluation (Siskos et al. 2013a, b; Giannakopoulos et al. 2010)
- Recommendation systems (Lakiotaki and Matsatsinis 2012; Delias et al. 2013a)
- Energy management (Diakoulaki et al. 1999; Androulaki and Psarras 2016; Angelopoulos et al. 2017)
- Education (Manouselis and Sampson 2002; Matsatsinis and Fortsas 2005; Krassadaki et al. 2015)
- Strategic management (Mastorakis and Siskos 2015)
- Marketing of agricultural products (Siskos and Matsatsinis 1993; Baourakis et al. 1993, 1996; Matsatsinis et al. 1999, 2000, 2007; Siskos et al. 2001; Matsatsinis and Siskos 2001, 2003)
- Consumer behavior (Siskos et al. 1995a, b; Baourakis et al. 1995; Kettani et al. 1998; Matsatsinis and Samaras 2000; Manouselis and Matsatsinis 2001; Matsatsinis 2002; Lakiotaki et al. 2009, 2011)
- Sales strategic management (Richard 1983; Siskos 1986)
- Portfolio management (Hurson and Zopounidis 1997; Zopounidis et al. 1999; Samaras et al. 2003; Hurson et al. 2012)
- Country risk assessment (Cosset et al. 1992; Oral et al. 1992; Zopounidis et al. 2000)
- Business financing (Siskos et al. 1994; Zopounidis et al. 1996; Zopounidis and Doumpos 1998; Zopounidis 2001)
- Maritime operations risk assessment (Stavrou et al. 2018)
- Venture capital evaluation (Siskos and Zopounidis 1987)
- Business failure prediction (Zopounidis 1987; Zopounidis and Doumpos 1999; Doumpos and Zopounidis 2002)

Variants of UTA have also been applied for conflict resolution in multi-actor decision situations (Jacquet-Lagrèze and Shakun 1984; Bui 1987; Matsatsinis and Samaras 2001) and have been extended in the case of multiple DMs (Matsatsinis et al. 2005; Siskos and Grigoroudis 2010; Delias and Matsatsinis 2013). Moreover, UTA-based approaches have been combined with multi-agent systems (Matsatsinis et al. 1999, 2000, 2001; Manouselis and Matsatsinis 2001; Matsatsinis and Delias 2003; Matsatsinis and Delias 2004; Delias and Matsatsinis 2013).

The bipolar robustness control procedure presented in this chapter may be applied in all of the aforementioned application domains in order to manage robustness in the disaggregation, as well as the aggregation stage of the UTA methods.

Finally, several real-world applications have been focused on the synergy between UTA methods and other MCDA approaches (Hurson et al. 2012; Lakiotaki and Matsatsinis 2012; Delias et al. 2013a, b; Krassadaki et al. 2015; Siskos et al. 2013a, b; Demesouka et al. 2013; Doumpos et al. 2016).

13.6 Conclusions

The interactive aggregation-disaggregation approach presented in this chapter aims to infer preference models using preferential structures, provided by the DM. In particular, the proposed approach may be considered as a new outlook on the UTA-family methods, devoted to the elicitation of values through the inference of multiple additive value models.

In general, robustness analysis should be considered as a tool of resistance of decision analysts against the phenomena of "vague approximations" and "ignorance zones". As noted by Roy (2010), robustness analysis should be considered differently from "sensitivity analysis", since the latter does not include all the forms of vague approximations and zones of ignorance that must be resisted or protected against (e.g., approximations due to simplifications, ill-defined data or arbitrary options and zones of ignorance due to imperfect knowledge about the complexity of the phenomena or the systems of values). Therefore, robustness analysis refers to a capacity for withstanding the aforementioned "vague approximations" and/or "zones of ignorance" in order to prevent undesirable impacts (see also Roy 2005). Moreover, robustness analysis should include the measurement of the robustness of a decision model, the development of appropriate robustness indicators, and the potential improvement of robustness (Siskos and Grigoroudis 2010; Siskos and Psarras 2016).

In this context, several robustness indicators are presented in this chapter, while the proposed bipolar robustness control procedure is able to take into account the different perspectives of robustness:

- Analyst's point of view during the 1st pole (disaggregation) that examines if a decision model is reliable.
- DM's point of view during the 2nd pole (aggregation) that examines if the results of a decision model are acceptable.

Based on the above, the interaction between the analyst and the DM is necessary during any robustness control procedure in UTA methods. In general, this interaction procedure may include:

(a) The consistency between the assessed preference model and the a priori preferences of the DM;
(b) The assessed values (e.g., values, weights, utilities);
(c) The overall evaluation of potential actions (extrapolation output).

Future research regarding the presented approach may examine additional interaction protocols between the analyst and the DM or explore additional robustness indicators. In this context, the additional preferential structures, given by the DM, during the robustness control procedure, can be further examined in order to analyze the required information (e.g., amount or consequences of information). Additional potential future research directions may include the further experimental evaluation of disaggregation-aggregation procedures.

References

Androulaki S, Psarras J (2016) Multicriteria decision support to evaluate potential long-term natural gas supply alternatives: the case of Greece. Eur J Oper Res 253(3):791–810

Angelopoulos D, Siskos Y, Psarras J (2017) Disaggregating time series on multiple criteria for robust forecasting: the case of long-term electricity demand in Greece. Paper presented at the 85th meeting of the EURO working group on MCDA, Padova, Italy, April 20–22, 2017

Baourakis G, Matsatsinis NF, Siskos Y (1993) Agricultural product design and development. In: Janssen J, Skiadas CH (eds) Applied stochastic models and data analysis. World Scientific, Singapore, pp 1108–1128

Baourakis G, Matsatsinis NF, Siskos Y (1995) Consumer behavioural analysis using multicriteria methods. In: Janssen J, Skiadas CH, Zopounidis C (eds) Advances in stochastic modelling and data analysis. Kluwer Academic, Dordrecht, pp 328–338

Baourakis G, Matsatsinis NF, Siskos Y (1996) Agricultural product development using multidimensional and multicriteria analyses: the case of wine. Eur J Oper Res 94(2):321–334

Beuthe M, Eeckhoudt L, Scanella G (2000) A practical multicriteria methodology for assessing risky public investments. Socio Econ Plan Sci 34(2):121–139

Bui TX (1987) Co-oP: a group decision support system for cooperative multiple criteria group decision making, Lecture notes in computer science, vol 290. Springer-Verlag, Berlin

Charnes A, Cooper W, Ferguson RO (1955) Optimal estimation of executive compensation by linear programming. Manag Sci 1(2):138–151

Ciomek K, Kadziński M, Tervonen T (2016) Heuristics for prioritizing pair-wise elicitation questions with additive multi-attribute value models. Omega 71:27–45

Cosset JC, Siskos Y, Zopounidis C (1992) Evaluating country risk: a decision support approach. Glob Finance J 3(1):79–95

Delias P, Kyriakaki G, Grigoroudis E, Matsatsinis N (2013a) Innovation management software exploiting multiple criteria analysis: the case of Innovation Centre of Crete. Int J Decis Support Syst Technol 4(1):30–42

Delias P, Manitsa P, Grigoroudis E, Matsatsinis N, Karasavvoglou A (2013b) Robustness-oriented group decision support: a case from ecology economics. Procedia Technol 8:285–291

Delias P, Matsatsinis N (2013) Multiple criteria decision aid and agents: supporting effective resource federation in virtual organizations. In: Doumpos M, Grigoroudis E (eds) Multicriteria decision aid and artificial intelligence. John Wiley & Sons, Ltd, Chichester, pp 275–284

Demesouka OE, Vavatsikos AP, Anagnostopoulos KP (2013) Spatial UTA (S-UTA): a new approach for raster-based GIS multicriteria suitability analysis and its use in implementing natural systems for wastewater treatment. J Environ Manag 125:41–54

Diakoulaki D, Zopounidis C, Mavrotas G, Doumpos M (1999) The use of a preference disaggregation method in energy analysis and policy making. Energy 24(2):157–166

Doumpos M, Zopounidis C (2002) Business failure prediction: a comparison of classification methods. Oper Res Int J 2(3):303–319

Doumpos M, Xidonas P, Xidonas S, Siskos Y (2016) Development of a robust multicriteria classification model for monitoring the postoperative behaviour of heart patients. J Multi-Criteria Decis Anal 23(1–2):15–27
Freed N, Glover G (1981) Simple but powerful goal programming models for discriminant problems. Eur J Oper Res 7:44–60
Giannakopoulos D, Manolitzas P, Spyridakos A (2010) Measuring e-government: a comparative study of the G2C online services progress using multi-criteria analysis. Int J Decis Support Syst Technol 2(4):1–12
González-Araya MC, Rangel LAD, Lins MPE, Gomes LFAM (2002) Building the additive utility functions for CAD-UFRJ evaluation staff criteria. Ann Oper Res 116(1–4):271–288
Greco S, Slowinski R, Figueira J, Mousseau V (2010) Robust ordinal regression. In: Ehrgott M, Greco S, Figueira J (eds) Trends in multiple criteria decision analysis. Springer, Berlin, pp 241–283
Grigoroudis E, Zopounidis C (2012) Developing an employee evaluation management system: the case of a healthcare organization. Oper Res Int J 12(1):83–106
Grigoroudis E, Siskos Y (2010) Customer satisfaction evaluation: methods for measuring and implementing service quality. Springer, New York, NY
Hatzinakos I, Yannacopoulos D, Faltsetas C, Ziourkas C (1991) Application of the MINORA decision support system to the evaluation of landslide favourability in Greece. Eur J Oper Res 50(1):60–75
Hurson C, Zopounidis C (1997) Gestion de portefeuille et analyse multicritère. Economica, Paris
Hurson C, Mastorakis K, Siskos Y (2012) Application of a synergy of MACBETH and MAUT multicriteria methods to portfolio selection in Athens stock exchange. Int J Multi-Criteria Decis Mak 2(2):113–127
Jacquet-Lagrèze E (1995) An application of the UTA discriminant model for the evaluation of R&D projects. In: Pardalos PM, Siskos Y, Zopounidis C (eds) Advances in multicriteria analysis. Kluwer Academic, Dordrecht, pp 203–211
Jacquet-Lagrèze E, Siskos J (1978) Une méthode de construction de fonctions d' utilité additives explicatives d' une préférence globale, Cahier du LAMSADE, 16, Université de Paris-Dauphine
Jacquet-Lagrèze E, Shakun MF (1984) Decision support systems for semistructured buying decisions. Eur J Oper Res 16:48–56
Jacquet-Lagrèze E, Siskos Y (1982) Assessing a set of additive utility functions for multicriteria decision making: the UTA method. Eur J Oper Res 10(2):151–164
Jacquet-Lagrèze E, Siskos Y (2001) Preference disaggregation: 20 years of MCDA experience. Eur J Oper Res 130(2):233–245
Kadziński M, Greco S, Słowiński R (2012) Extreme ranking analysis in robust ordinal regression. Omega 40(4):488–501
Karst OJ (1958) Linear curve fitting using least deviations. J Am Stat Assoc 53:118–132
Kelley JE (1958) An application of linear programming to curve fitting. J Ind Appl Math 6(1):15–22
Kettani O, Oral M, Siskos Y (1998) A multiple criteria analysis model for real estate evaluation. J Glob Optim 12(2):197–214
Krassadaki E, Lakiotaki K, Matsatsinis N, (2015) Students' behavior in peer assessment: a multi-criteria clustering approach. Eur J Eng Educ 39(3):233–246
Lakiotaki K, Matsatsinis NF (2012) Analyzing user behaviour in recommender systems. Int J Electron Bus 10(1):1–19
Lakiotaki K, Matsatsinis NF, Tsoukias A (2011) Multi-criteria user profiling in recommender systems. IEEE Intell Syst 26(2):64–76
Lakiotaki K, Delias P, Sakkalis V, Matsatsinis NF (2009) User profiling based on multi-criteria analysis: the role of utility functions. Oper Res Int J 9(1):3–16
Manas M, Nedoma J (1968) Finding all vertices of a convex polyhedron. Numer Math 14:226–229
Manolitzas P, Grigoroudis E, Matsatsinis N (2013a) MEDUTA: integrating simulation modeling and multiple criteria analysis to improve emergency department performance. Paper presented

at the 2nd international symposium and 24th national conference on operational research, National Technical University of Athens, Athens, September 26–28, 2013

Manolitzas P, Matsatsinis N, Grigoroudis E (2013b) Reforming the hospitals in Greece: an integrated framework for improving the health care services in an Emergency Department. Paper presented at the 6th Biennial Hellenic observatory PhD symposium on Contemporary Greece and Cyprus, London School of Economics (LSE), London, UK, June 6–7, 2013

Manouselis N, Sampson D (2002) Multi-criteria decision making for broker agents in eLearning environments. Oper Res Int J 2(3):347–361

Manouselis N, Matsatsinis NF (2001) Introducing a multi-agent, multi-criteria methodology for modeling electronic consumer's behavior: the case of internet radio. In: Klush M, Zambonelli F (eds) Lecture notes in artificial intelligence-cooperative information agents, vol 2182. Springer Verlag, Berlin, pp 190–195

Mastorakis K, Siskos E (2015) Value focused pharmaceutical strategy determination with multi-criteria decision analysis techniques. Omega 59A:84–96

Matsatsinis NF (2002) New agricultural product development using data mining techniques and multicriteria methods. In: Sideridis A (ed) Proceedings of the 1st Hellenic Association of information and communication technology in agriculture, food and environment (HAICTA's conference 2002), June 6–7, Athens, Greece

Matsatsinis NF, Samaras AP (2000) Brand choice model selection based on consumers' multicriteria preferences and experts' knowledge. Comput Oper Res 27(8):689–707

Matsatsinis NF, Samaras AP (2001) MCDA and preference disaggregation in group decision support systems. Eur J Oper Res 130(2):414–429

Matsatsinis NF, Delias P (2003) AgentAllocator: an agent-based multi-criteria decision support system for task allocation. In: Marik V, McFarlane D, Valckenaers P (eds) Holonic and multi-agent systems for manufacturing, Lecture notes in computer science, vol 2744. Springer Verlag, Berlin, pp 225–235

Matsatsinis NF, Delias P (2004) A multi-criteria protocol for multi-agent negotiations. In: Vouros GA, Panayiotopoulos T (eds) Methods and applications of artificial intelligence, Lectures notes in artificial intelligence, vol 3025. Springer-Verlag, Berlin Heidelberg, pp 103–111

Matsatsinis NF, Fortsas VC (2005) A multicriteria methodology for the assessment of distance education trainees. Oper Res Int J 5(3):419–434

Matsatsinis NF, Siskos Y (2001) DIMITRA: an intelligent decision support system for agricultural products development decisions. In: Sevila F (ed) Proceedings of the 3rd European conference of the european federation for information technology in agriculture, food and the environment (EFITA 2001), June 18–20, Montpellier, France

Matsatsinis NF, Siskos Y (2003) Intelligent support systems for marketing decision. Kluwer Academic, Dordrecht

Matsatsinis NF, Grigoroudis E, Samaras AP (2005) Aggregation and disaggregation of preferences for collective decision-making. Group Decis Negot 14(3):217–232

Matsatsinis NF, Grigoroudis E, Samaras AP (2007) Comparing distributors' judgements to buyers' preferences: a consumer value analysis in the Greek olive oil market. Int J Retail Distrib Manag 35(5):342–362

Matsatsinis NF, Moraitis P, Psomatakis V, Spanoudakis N (1999). Intelligent software agents for products penetration strategy selection. In: Garijo FG, Boman M (eds) Proceedings of the 9th European workshop on modelling autonomous agents in a multi-agent world 1999 (MAAMAW'99), June 30–July 2, Valencia, Spain (CD format)

Matsatsinis NF, Moraitis P, Psomatakis V, Spanoudakis N (2000) Multi-agent architecture for agricultural products development. In: Sciefer G, Helbig R, Rickert U (eds) Proceedings of the 2nd European conference of the European federation for information technology in agriculture, food and the environment (EFITA 2000), September 27–30, Bonn, Germany, pp 187–196

Matsatsinis NF, Moraitis P, Psomatakis V, Spanoudakis N (2001) An agent-based system for products penetration strategy selection. Appl Artif Intell J 17(10):901–925

Oral M, Kettani O, Cosset JC, Daouas M (1992) An estimation model for country risk rating. Int J Forecast 8(4):583–593

Richard JL (1983) Aide à la décision stratégique en P.M.E. In: Jacquet-Lagrèze E, Siskos Y (eds) Méthode de décision multicritère. Hommes et Techniques, Paris, pp 119–142

Roy B (2005) A propos de robustesse en recherche opérationnelle et aide à la decision. In: Billaut JC, Moukrim A, Sanlaville E (eds) Flexibilité et robustesse en ordonnancement. Lavoisier, Paris, pp 35–50

Roy B (2010) Robustness in operational research and decision aiding: a multi-faceted issue. Eur J Oper Res 200(3):629–638

Samaras GD, Matsatsinis NF, Zopounidis C (2003) A multicriteria DSS for a global stock evaluation. Oper Res Int J 3(3):281–306

Siskos E, Psarras J (2016) Bipolar robustness control methodology in disaggregation MCDA approaches: application to European e-government evaluation. Paper presented at the 28th European conference on operational research, Poznan, Poland, July 3–6, 2016

Siskos E, Askounis D, Psarras J (2013a) Robust e-government evaluation based on multiple criteria analysis techniques. Paper presented at the 77th meeting of the EURO working group on MCDA, Rouen, France, April 11–13, 2013

Siskos E, Malafekas M, Askounis D, Psarras J (2013b) E-government benchmarking in European Union: a multicriteria extreme ranking approach. In: Douligeris C, Polemi N, Karantias A, Lamersdorf W (eds) Collaborative, trusted and privacy-aware e/m-services. Springer, New York, NY, pp 338–348

Siskos J (1980) Comment modéliser les préférences au moyen de fonctions d'utilité additives. RAIRO Recherche Opérationelle 14:53–82

Siskos J (1983) Analyse de systèmes de décision multicritère en univers aléatoire. Foundations Control Eng 10(3–4):193–212

Siskos J, Zopounidis C (1987) The evaluation criteria of the venture capital investment activity: an interactive assessment. Eur J Oper Res 31(3):304–313

Siskos J, Assimakopoulos N (1989) Multicriteria highway planning: a case study. Math Comput Model 12(10–11):1401–1410

Siskos J, Matsatsinis NF (1993) A DSS for market analysis and new product design. J Decis Syst 2(1):35–60

Siskos Y (1986) Evaluating a system of furniture retail outlets using an interactive ordinal regression method. Eur J Oper Res 23:179–193

Siskos Y, Yannacopoulos D (1985) UTASTAR: an ordinal regression method for building additive value functions. Investigação Operacional 5(1):39–53

Siskos Y, Grigoroudis E (2010) New trends in aggregation-disaggregation approaches. In: Zopounidis C, Pardalos PM (eds) Handbook of multicriteria analysis. Springer, Heidelberg, pp 189–214

Siskos Y, Zopounidis C, Pouliezos A (1994) An integrated DSS for financing firms by an industrial development bank in Greece. Decis Support Syst 12(2):151–168

Siskos Y, Grigoroudis E, Matsatsinis NF (2016) UTA methods. In: Greco S, Ehrgott M, Figueira J (eds) Multiple criteria analysis: state of the art surveys, vol 1, 2nd edn. Springer, New York, NY, pp 315–362

Siskos Y, Grigoroudis E, Matsatsinis NF, Baourakis G (1995a) Preference disaggregation analysis in agricultural product consumer behaviour. In: Pardalos PM, Siskos Y, Zopounidis C (eds) Advances in multicriteria analysis. Kluwer Academic, Dordrecht, pp 185–202

Siskos Y, Grigoroudis E, Matsatsinis NF, Baourakis G, Neguez F (1995b) Comparative behavioural analysis of European olive oil consumer. In: Janssen J, Skiadas CH, Zopounidis C (eds) Advances in stochastic modelling and data analysis. Kluwer Academic, Dordrecht, pp 293–310

Siskos Y, Matsatsinis NF, Baourakis G (2001) Multicriteria analysis in agricultural marketing: the case of French olive oil market. Eur J Oper Res 130(2):315–331

Spyridakos A, Siskos Y, Yannakopoulos D, Skouris A (2000) Multicriteria job evaluation for large organisations. Eur J Oper Res 130(2):375–387

Srinivasan V, Shocker AD (1973) Linear programming techniques for multidimensional analysis of preferences. Psychometrika 38(3):337–396

Stavrou D, Siskos E, Ventikos N, Psarras J (2018) Robust evaluation of the risk of ship-to ship transfer operations: application in a multicriteria and stochastic environment. In: Lee PTW, Yang Z (eds) Multi-criteria decision making in maritime studies and logistics: applications and cases. Springer, Heidelberg. isbn:978-3-319-62336-8

Wagner HM (1959) Linear programming techniques for regression analysis. J Am Stat Assoc 54:206–212

Young FW, De Leeuw J, Takane Y (1976) Regression with qualitative and quantitative variables: an alternating least squares method with optimal scaling features. Phychometrica 41(4):505–529

Zopounidis C, Matsatsinis NF, Doumpos M (1996) Developing a multicriteria knowledge-based decision support system for the assessment of corporate performance and viability: the FINEVA system. Fuzzy Econ Rev 1(2):35–53

Zopounidis C (1987) A multicriteria decision-making methodology for the evaluation of the risk of failure and an application. Foundations Control Eng 12(1):45–67

Zopounidis C (2001) Preference disaggregation in financial modeling: basic features and some examples. Oper Res Int J 1(3):263–284

Zopounidis C, Doumpos M (1998) Developing a multicriteria decision support system for financial classification problems: the FINCLAS system. Optim Methods Software 8(3–4):277–304

Zopounidis C, Doumpos M (1999) Business failure prediction using UTADIS multicriteria analysis. J Oper Res Soc 50(11):1138–1148

Zopounidis C, Hurson C, Doumpos M (2000) Risque-Pays: evaluation des aspects economiques, sociaux et politiques. Economica, Paris

Zopounidis C, Doumpos M, Zanakis SH (1999) Stock evaluation using a preference disaggregation methodology. Decis Sci 30(2):313–336

Chapter 14
Eliciting Multi-Criteria Preferences: ELECTRE Models

Luis C. Dias and Vincent Mousseau

Abstract Outranking methods are a specific type of Multi-Criteria Decision Aiding methods. They are based on the construction of binary relations validating or invalidating, for any pair of alternatives (a, b), the assertion "*a outranks b*". This comparison is grounded on the evaluation vectors of both alternatives, and on additional information concerning the decision maker's preferences, typically accounting for two conditions: concordance and non-discordance. In decision processes using these methods, the analyst should interact with the decision maker in order to elicit values for the parameters that define a preference model. This can be done either directly or through a disaggregation procedure that infers parameter values from holistic judgements provided by the decision maker. In this chapter we discuss the elicitation of an outranking-based preference model, focusing on the valued outranking relation used in the ELECTRE III and ELECTRE TRI methods.

14.1 Introduction

As described in Chap. 12 in this book (Morton 2018), a common approach in the field of Multiple Criteria Decision Aiding (MCDA) is to aggregate the performances of an alternative being assessed on multiple criteria into a single number synthesizing its overall value (see also Keeney and Raiffa 1993). However, a different type of methods has been developed in parallel, which obtain a binary relation on the set of alternatives without aggregating multiple performances into a synthesis value. These methods are usually referred as outranking methods in the MCDA literature and have been, by and large, associated with the so-called European school of MCDA (see Roy and Vanderpooten 1996). This allows for decision aiding approaches able to model not only situations of preference or indifference between alternatives, but

L.C. Dias (✉)
Faculty of Economics, CeBER and INESC Coimbra, University of Coimbra,
Av. Dias da Silva 165, 3004-512, Coimbra, Portugal
e-mail: lmcdias@fe.uc.pt

V. Mousseau
Laboratoire Génie Industriel, CentraleSupélec, Châtenay-Malabry, France
e-mail: vincent.mousseau@centralesupelec.fr

L.C. Dias et al. (eds.), *Elicitation*, International Series in Operations Research & Management Science 261, https://doi.org/10.1007/978-3-319-65052-4_14

also situations in which alternatives are deemed to be incomparable in the light of the preference information elicited. Incomparability typically occurs when the strengths and weaknesses of two alternatives are so different that one cannot conclude that one is better than the other, but it is equally unwarranted to conclude that they are indifferent (i.e., similarly preferred).

Outranking methods ground the recommendations to the Decision Maker (DM) on the construction of one (or several) binary relation(s) representing the preference among pairs of alternatives (see Roy 1991; Roy and Bouyssou 1993). A simple binary relation is dominance: an alternative dominates another one if it is better on some criteria and it is not worse in any other criterion. It does not require any subjective parameters such as criteria weights, but the relation is usually poor (i.e., it applies to few pairs of alternatives). Outranking methods use additional inputs to enrich this relation. Examples of outranking methods include ELECTRE methods (Figueira et al. 2013), PROMETHEE methods (Brans and Vincke 1985; Majid Behzadian et al. 2010), RUBIS (Bisdorff et al. 2007), NAIADE (Munda 1995), and qualitative approaches (Martel and Matarazzo 2005). This chapter will focus on preference elicitation for ELECTRE methods, but analogous procedures can be applied for other outranking-based approaches.

Let us consider a decision situation involving a finite set of alternatives $A = \{a_1, a_2, \ldots, a_l\}$ evaluated on n criteria $g_1, g_2, \ldots, g_n$, ($F = \{1, 2, \ldots, n\}$ denotes the set of criteria indices).

The construction of an outranking relation S amounts at validating or invalidating, for any pair of alternatives $(a, b) \in A^2$, an assertion aSb, whose meaning is "a is at least as good as b" or, in other words, "a is not worse than b". This comparison is grounded on the evaluation vectors of both alternatives a and b, *i.e.*, $(g_1(a), g_2(a), \ldots, g_n(a))$ and $(g_1(b), g_2(b), \ldots, g_n(b))$, and on additional information concerning the DM's preferences. To validate a statement aSb, two basic conditions should be verified: concordance and non-discordance (or non-veto).

A criterion g_k is said to be concordant with the assertion aSb if a is at least as good as b with respect to criterion g_k. The concordance condition is fulfilled for the assertion aSb when the subset of criteria concordant with aSb is a "sufficient majority". A criterion g_k is said to veto the assertion aSb if a is so much worse than b on this criterion that the difference of evaluation $|g_k(b) - g_k(a)|$ becomes incompatible with the assertion aSb, whatever the evaluation on the other criteria. The non-discordance condition is fulfilled when no criterion opposes a veto to the assertion aSb.

Constructing an outranking relation S involves the elicitation of values for preference-related parameters, such as weights, majority thresholds and veto thresholds. The next section provides details about these parameters and how they shape a model of the DM's preferences. Sections 14.3 and 14.4 in this chapter discuss how to elicit parameter values. The elicitation of preference-related parameters can be done either in a direct way centered on parameters (discussed in Sect. 14.3) or indirectly through a disaggregation procedure centered on examples, that infers the parameters values from holistic preferences provided by the DM (see Jacquet-Lagrèze and Siskos 2001) (discussed in Sect. 14.4). Inference is usually performed

through an optimization program that accounts for the aggregation model and minimizes an "error function". This disaggregation approach has been largely used in additive models (e.g. see Jacquet-Lagrèze and Siskos 2001 and Chap. 13 in this book Matsatsinis et al. 2018). Section 14.5 discusses the elicitation process, namely focusing on the order parameters are elicited and how precise should the elicitation be. Section 14.6 closes the chapter summarizing the main takeaways and highlighting the research challenges that still lie ahead.

14.2 Preference Models with ELECTRE

This section briefly presents the ELECTRE preference model, namely describing how a valued outranking relation on the set of alternatives is built in methods such as ELECTRE III (see Roy 1978) and ELECTRE TRI (see Yu 1992a,b; Roy and Bouyssou 1993).

14.2.1 Outranking Relations for a Single Criterion

ELECTRE builds, for each criterion g_j, a valued outranking relation S_j modelling the comparison of alternatives on that single criterion. For any ordered pair $(a, b) \in A^2$, $S_j(a, b)$ is defined by (14.2) on the basis of $g_j(a)$, $g_j(b)$ and two thresholds: indifference q_j and preference p_j ($0 \leq q_j \leq p_j$). We consider the thresholds p_j and q_j as constant, although it is possible to consider them as affine functions (for such cases see Roy et al. 2014). For a more compact notation, we will write:

$$\Delta_j(b, a) = g_j(b) - g_j(a), \tag{14.1}$$

which for each pair $(a, b) \in A^2$ represents the advantage of b over a on the jth criterion. This assumes, without loss of generality, that the evaluations are coded in such a way that the higher the value, the better it is (if this is not the case, one simply considers that $\Delta_j(b, a) = g_j(a) - g_j(b)$).

$S_j(a, b)$ represents the degree to which alternative a outranks (is at least as good as) b, defined as (Fig. 14.1):

$$S_j(a, b) = \begin{cases} 0, & \text{if } \Delta_j(b, a) > p_j \\ \frac{p_j - \Delta_j(b,a)}{p_j - q_j}, & \text{if } q_j < \Delta_j(b, a) \leq p_j \\ 1, & \text{if } \Delta_j(b, a) \leq q_j \end{cases} \tag{14.2}$$

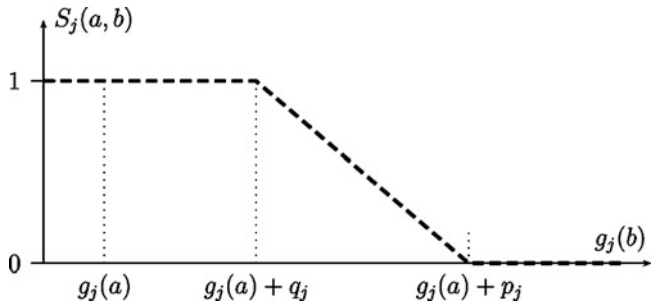


Fig. 14.1 Partial valued outranking relation

14.2.2 Concordance Relation

The valued concordance relation $C(a,b)$ aggregates the relations S_j ($j \in F$), and it represents the level of majority among the criteria in favor of the assertion "a is at least as good as b". When computing this majority level, each criterion g_j has a weight $w_j \geq 0$ representing its voting power. Without any loss of generality, we will consider $\sum_{j\in F} w_j = 1$. Therefore, $C(a,b)$ can be written as follows:

$$C(a,b) = \sum_{j\in F} w_j S_j(a,b) \tag{14.3}$$

14.2.3 Discordance Relations

ELECTRE also builds, for each criterion g_j, a valued discordance relation d_j restricted to that criterion. This relation $d_j(a,b)$ is traditionally defined by (14.4) on the basis of $g_j(a)$, $g_j(b)$, a veto threshold v_j and a preference threshold p_j ($p_j < v_j$; note we consider $p_j < v_j$, although ELECTRE also allows $p_j = v_j$) (see Fig. 14.2). We consider the thresholds v_j as constants (as we already did for p_j and q_j), although it is possible to consider them as affine functions.

$$d_j(a,b) = \begin{cases} 1, & \text{if } \Delta_j(b,a) \geq v_j \\ \frac{\Delta_j(b,a)-p_j}{v_j-p_j}, & \text{if } p_j < \Delta_j(b,a) < v_j \\ 0, & \text{if } \Delta_j(b,a) \leq p_j \end{cases} \tag{14.4}$$

The overall valued non-discordance relation $ND(a,b)$ as originally defined (Roy 1978) is grounded on $C(a,b)$ and on the relations d_j, $j \in F$; it represents the degree to which the minority criteria collectively oppose a veto to the assertion "a is at least as good as b". The classical way of defining $ND(a,b)$ is given in (14.5). $ND(a,b) = 0$ corresponds to a situation where some minority criteria are totally opposed to aSb

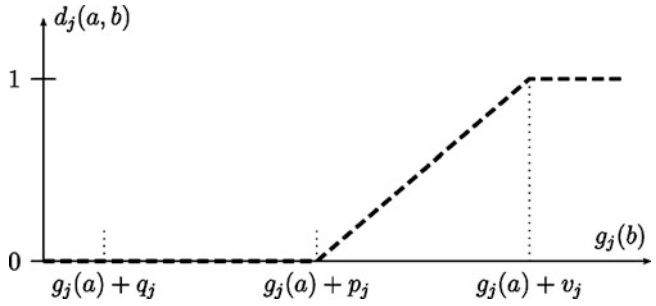


Fig. 14.2 Partial valued discordance relation

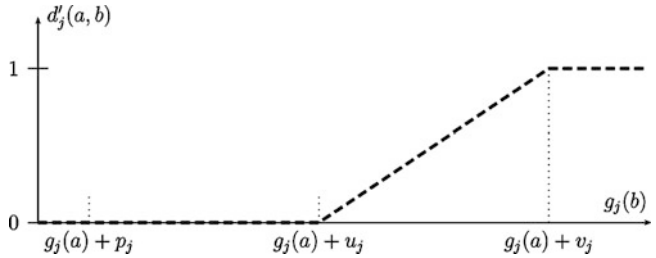


Fig. 14.3 Partial discordance relation $d'_j(a,b)$

whereas $ND(a,b) = 1$ means that none of the criteria oppose a veto to aSb.

$$ND(a,b) = \prod_{j \in \overline{F}} \frac{1 - d_j(a,b)}{1 - C(a,b)} \quad \text{where } \overline{F} = \{j \in F \text{ such that } d_j(a,b) > C(a,b)\} \tag{14.5}$$

This expression is equivalent to (14.6):

$$ND(a,b) = \prod_{j \in F} ND_j(a,b), \tag{14.6}$$

where:

$$ND_j(a,b) = Min\left\{1, \frac{1 - d_j(a,b)}{1 - C(a,b)}\right\}. \tag{14.7}$$

Mousseau and Dias (2004) have proposed an alternative valued non-discordance relation defined by (14.8)–(14.9), where $u_j \in [p_j, v_j[$ is a new parameter (discordance threshold) for the j-th criterion (Fig. 14.3):

$$ND'(a,b) = \prod_{j \in F} ND'_j(a,b) = \prod_{j \in F} \left(1 - d'_j(a,b)\right) \tag{14.8}$$

$$d'_j(a,b) = \begin{cases} 1 & \text{if } \Delta_j(b,a) \geq v_j \\ \frac{\Delta_j(b,a)-u_j}{v_j-u_j} & \text{if } u_j < \Delta_j(b,a) < v_j \\ 0 & \text{if } \Delta_j(b,a) \leq u_j \end{cases} \tag{14.9}$$

A second alternative to define a valued non-discordance relation is the following (see Mousseau and Dias 2004), which is simpler but only takes the highest discordance into account:

$$ND''(a,b) = Min_{j \in F} ND'_j(a,b) \tag{14.10}$$

Both definitions (14.8) and (14.10) follow ELECTRE's intention of allowing one minority criterion to veto the conclusion sustained by the majority of the criteria, if the performance difference is too large (and worse). These two definitions are mainly relevant when used in indirect elicitation (regression) processes, since they allow easier mathematical programming models to infer parameter values, especially variant $ND''(a,b)$.

14.2.4 Valued Outranking Relations

ELECTRE's valued outranking relation combines the concordance and non-discordance relations:

$$S(a,b) = C(a,b)\, ND(a,b), \tag{14.11}$$

or, according to the two alternative definitions of the discordance concept,

$$S'(a,b) = C(a,b)\, ND'(a,b) \tag{14.12}$$

$$S''(a,b) = C(a,b)\, ND''(a,b) \tag{14.13}$$

From a valued outranking relation such as $S(a,b)$, $S'(a,b)$ or $S''(a,b)$ it is possible to define a family of nested "crisp" outranking relations S_λ. These crisp relations correspond to λ-cuts of $S(a,b)$, where the cutting level $\lambda \in [0.5, 1]$ represents the minimum value for $S(a,b)$ so that $aS_\lambda b$ holds.

14.2.5 Exploitation of the Outranking Relation

Depending on the type of decision problem different ELECTRE methods can be applied. Roy (1996) identifies three main "problématiques" depending on the type of result sought:

- Selection (or choice): to identify the best alternative (or a predefined number of best alternatives) among a set of possibilities. Example: to select the best project among a set of possible variants.
- Ranking: to obtain a preference order among the alternatives, from best to worst. Example: a prioritization of projects defining the order by which they should be implemented.
- Sorting problems aim at assigning alternatives to categories, which are typically defined a priori and ordered. Example: sorting projects among the categories "not urgent", "urgent" and "very urgent".

Methods ELECTRE I and IS (Roy and Bouyssou 1993; Roy and Skalka 1984) have been proposed to deal with selection problems. Since the outranking relation is usually not transitive and not complete, often these methods are unable to identify a single winner. Their purpose is more modestly to identify a subset, named kernel, of candidates to be the most preferred alternative. The methods try to make this kernel as small as possible by excluding alternatives that are outranked. Alternatives in the kernel are incomparable, which typically means they are too different to be compared with the information requested by ELECTRE.

Methods ELECTRE II, III, and IV (Roy and Bertier 1973; Roy and Bouyssou 1993; Vallée and Zielniewicz 1994) have been proposed to deal with ranking problems. As in the case of choice, the lack of transitivity and incompleteness of the outranking relation hinder obtaining a clear-cut result. These methods yield a partial preorder as an output, i.e., an incompletely defined ranking allowing ties and in which some of the alternatives are incomparable.

Although ELECTRE methods for ranking and sorting have been used in many applications (Govindan and Jepsen 2016), the inconclusiveness of its results may disappoint some DMs and analysts. On the other hand, this inconclusiveness may be seen as a strength in that ELECTRE I-IV do not force the result to be more conclusive than warranted by the data and the preferences elicited. Another concern that has been much debated (e.g., Figueira and Roy 2009) is the fact that adding, removing, or modifying a possibly irrelevant alternative can change the relative position of the remaining alternatives. New ranking methods overcoming the latter issue have been proposed more recently (Rolland 2013).

Finally, ELECTRE TRI (Yu 1992b) and its variants are devoted to sorting problems. Since they do not compare the alternatives being evaluated against each other, adding, removing, or modifying an alternative has no effects of the results concerning the other alternatives. In ELECTRE TRI the alternatives are sorted based on how they compare to the profiles that define the available categories. These profiles are multidimensional preference vectors (each profile indicates one performance value for each criterion), which constitute new preference-related parameters to be elicited.

The original version of ELECTRE TRI (Yu 1992b) defined category profiles as bounds delimiting the categories: a profile b^1 separates category C^1 from category C^2 (b^1 can be considered a lower bound for C^2 and an upper bound for C^1); a profile

b^2 separates category C^2 from category C^3, and so on. This version is sometimes referred to as ELECTRE TRI-B.

A subsequent version of ELECTRE TRI is ELECTRE TRI-C (Almeida-Dias et al. 2010), which proposes to define profiles as central elements of the categories: a profile b^1 it the typical (characteristic) element of category C^1, a profile b^2 it the typical element of category C^2, and so on. Later, the extension ELECTRE TRI-nC was proposed to allow each category to be defined by more than one characteristic element (Almeida-Dias et al. 2012). By analogy, it is also possible to create an ELECTRE TRI-nB version (Fernández et al. 2017).

14.3 Direct Elicitation

14.3.1 Single-Criterion Concordance Parameters

It makes sense to start the process of eliciting an ELECTRE model by the single-criteria concordance parameters, since the parameters to be elicited afterwards are used in computations that refer to the relations S_j. Furthermore, the discussion about these parameters is not as cognitively demanding as for other parameters, and allows introducing the cornerstone concept of concordance in ELECTRE.

Given a pair of alternatives (a, b), $S_j(a, b)$ assesses the degree to which a outranks (is at least as good as) b according to the criterion g_j. According to (14.2), this depends on the advantage of a over b, denoted $\Delta_j(a, b) = -\Delta_j(b, a)$ and two parameters to be elicited: the indifference threshold q_j and the preference threshold p_j.

In the oldest ELECTRE methods (Roy 1968, 1971) the single-criterion concordance would be an absolute yes, i.e. $S_j(a, b) = 1$, if $\Delta_j(a, b) \geq 0$, or it would be an absolute no, i.e. $S_j(a, b) = 0$ otherwise. If a was worse than b on criterion g_j then there was no concordance at all, however small this difference might be. There are however some reasons why this model might be inadequate for some criteria:

- A small difference might be considered insignificant in relative terms concerning orders of magnitude. For instance, a difference of \$1 between two projects involving over \$1 million is not likely to be valued by any DM.
- Performances may be assessed in an imprecise way using measuring instruments or statistics. If the performance of a is 100 ± 5 and the performance of b is 101 ± 5, many DMs will be indifferent between one or the other because the performance difference is much lower than the acknowledged imprecision.
- Performance assessed may be just an imperfect indicator (or even a proxy) of real-world performance. For instance, the advertised fuel consumption of a car corresponds to its behavior in an idealized circuit (e.g., the New European Driving Cycle). If the performance of a is 5.0 l/100 km and the performance of b is 4.9 l/100 km, many DMs will be indifferent between them because they know that none of these values correspond to real-life performance. Likewise, when

recruiting a college graduate, the DM knows that the grade point average (GPA) of their degrees is just an imperfect indicator: a student with a GPA of 3.5 is not guaranteed to be more knowledgeable than another one with a GPA of 3.4.

The ELECTRE methods introduced later acknowledge these situations by allowing the DM to set an indifference threshold q_j, which is the largest difference such that the DM does not distinguish between two alternatives in terms of preference. A question addressing the need for such a threshold can be the following:

"If the difference between two alternatives on criterion j is not equal to zero then one of them must be preferred on that criterion, or can this difference be so small that you would not distinguish them in terms of preference?"

In the latter case, it is possible to ask for a limit to this indifference situation:

"How large can a performance difference be until you start hesitating about the indifference between two alternatives?"

The DM can reply in absolute terms, e.g., 2.0, or in relative terms, e.g., 2%. Although most ELECTRE software implementations allow modelling $q_j(g_j(a))$ as an affine function $\alpha_j + \beta_j g_j(a)$ for some parameters α_j and β_j, typically this option is not used and this function is either a constant value ($\beta_j = 0$) or a proportion of the performance ($\alpha_j = 0$). For simplicity, in the remainder of this text we assume it is a constant value. When thresholds are modelled as functions of the performance levels special care must be taken to ensure their consistency (Roy et al. 2014).

It is possible to ask verification questions and adjust the parameter by trial and error:

"If $g_j(a)$ has value x_a and $g_j(b)$ has value x_b (for some relatively close values x_a and x_b) would you say that on criterion j the two alternatives are indifferent, or would you have a clear preference?"

It is not uncommon that up to a difference δ_1 the DM feels clearly indifferent, for a difference larger than δ_2 ($\delta_2 > \delta_1$) the DM has a clear preference, and for differences in-between δ_1 and δ_2 the DM exhibits some hesitation in answering such a question. This allows setting $q_j = \delta_1$ and $p_j = \delta_2$, since the preference threshold p_j corresponds to the minimum difference such that the DM has a clear preference for one of the alternatives.

The elicitation of q_j and p_j can therefore result in one of these typical situations:

- $p_j > q_j > 0$, meaning that some differences are too small to warrant preference, and that up to a certain point there is a clear indifference, then some hesitation, and finally a clear preference as the difference in performance increases.
- $p_j > q_j = 0$, meaning that above a given threshold there is a clear preference, and below this threshold the DM hesitates if the alternatives are indifferent or one is better than the other.
- $p_j = q_j$, and possibly both are null, in some cases concerning a discrete scale (e.g., number of rooms in a house, or number of stars of an hotel).

14.3.2 Weights and Cutting Level

As presented in Sect. 14.2.2, criteria weights are used to aggregate the concordance of the different criteria concerning an outranking relation. Although they are used in a weighted sum of concordance indices, they should not be interpreted as trade-off weights. Unlike a typical additive aggregation model (e.g., Keeney and Raiffa 1993) weights are not scaling coefficients such that the ratio of two weights indicates the conversion rate between units of value (or utility) on two different criteria.

An adequate analogy for eliciting weights in ELECTRE is that of voting. Suppose for the moment that all indifference and preference thresholds are null, i.e., the single-criterion concordances are either 0 or 1. Suppose also that there is no discordance (veto thresholds are not set or are extremely high), so that $S(a,b) = C(a,b)$ for any pair of alternatives (a,b). Then, a outranks b if the weights of the coalition of criteria that add up to $C(a,b)$ reaches at least cutting level λ. Then, the cutting level λ can be interpreted as representing the required majority for establishing an outranking relation. Typical values for this parameter are 0.50 or 0.51 (a simple majority), 0.67 (requiring a 2/3 majority), etc., up to 1 (requiring unanimity). A direct elicitation question could be:

"How strong must the majority of the criteria that agree that a is at least as good as b be, in order to establish this conclusion, taking criteria weights into account? (in the absence of strong discordance)"

In a trial-and-error process tentative symbolic majority levels can be suggested, such as 1/2, 2/3 or 3/4. Otherwise, communicating in terms of percentages is preferable to decimal numbers (i.e., 60% communicates better the sense of a required majority than 0.60). The higher the majority level required, the less will the number of outranking relations be but the stronger is their justification. Often, a compromise is sought between the richness of the relation (number of pairs for which outranking holds) and the strength of the justification, by observing the effects of varying this parameter.

In the particular case of sorting problems with ELECTRE TRI (ELECTRE TRI-B) it may be more appropriate to inquire about the cutting level in a way that matches more directly its effects on the results:

"How strong must the majority of the criteria that agree that a is at least as good as the lower profile of a category be, taking criteria weights into account, to warrant that an alternative can be sorted on that category, if not better? (in the absence of strong discordance)"

Indeed, to be sorted in a given category (if not better) an alternative must outrank the category's lower profile. This parameter can be interpreted as denoting how much demanding the decision maker is. A high cutting level makes it more difficult for the alternatives to be classified in the best categories. Again, symbolic majority levels can be tentatively suggested.

Having the voting majority analogy in mind, then criteria weights simply reflect how much they count in the formation of such majorities. This means that weights

Table 14.1 Example of criteria weights

Criteria:	g_1	g_2	g_3	g_4	g_5
Weights (w_k):	0.15	0.20	0.15	0.15	0.35

match the analogy of weights in the physical world. A direct elicitation question could be:

"Considering that the support of all criteria for an outranking relation amounts to a 100% majority (unanimity), how much weight (or voting power) would you assign to criterion g_j alone?"

Confirmation questions can be asked concerning the elicited weights. Consider for instance the weights in Table 14.1. Since $w_1 < w_2$ one should confirm that having the support of the first criterion for an outranking relation is less important than having the support of the second criterion. Since $w_1 + w_2 = w_5$, one should confirm that the last criterion counts as much as the other two criteria. These are just two examples among many possible. Further confirmatory questions may interrelate the elicited weights and the cutting level. For instance, if $\lambda = 0.55$, one should confirm that:

- No criterion alone is strong enough to warrant an outranking relation.
- The only coalition of two criteria strong enough to warrant an outranking relation is g_2 together with g_5 (since $w_2 + w_5 = 0.55 = \lambda$)
- No coalition of three criteria is a sufficient majority unless g_5 is in it.
- Any coalition of four criteria is a sufficient majority (at the minimum, $w_1 + w_2 + w_3 + w_4 = 0.65$, which is larger than λ).

If indifference and preference thresholds are not null, the single-criterion concordances can be any value between 0 or 1, but this does not change the logic of the elicitation process. One simply has to reason that if, for instance, the performance of alternatives a and b is such that $S_j(a, b) = 0.50$, then criterion g_j contributes with half of its weight to the coalition supporting that a outranks b.

An alternative to directly asking for numerical criteria weights has been proposed by Simos (1990) and later revised by Figueira and Roy (2002). DMs can use cards with criteria names to indicate how they would rank the criteria by order of importance. Two or more cards can be placed together to indicate the respective criteria should have the same weight. In addition, DMs can place blank cards to indicate a higher difference in weights between ranks. For instance, DMs could indicate the following ranking: g_1 and g_2, g_3, (blank), g_4, (blank), (blank), g_5. This indicates that g_1 and g_2 are the two criteria with higher weight, followed by g_3, then g_4 and finally g_5. The blank cards in this example entail that one should have $w_4 - w_5 = 3*(w_2 - w_3)$ and that $w_3 - w_4 = 2*(w_2 - w_3)$. Since there are many weight vectors fulfilling these inequalities the revised Simos technique requires DMs to set a ratio between the first and the last ranked weights. The authors also propose a rounding technique if the resulting weights are required to have a predefined maximum number of decimal digits (for details see Roy and Figueira 2002).

14.3.3 Discordance Parameters

14.3.3.1 Parameters Defining $d_j(a,b)$

Being a noncompensatory preference model, ELECTRE allows specifying that a large disadvantage on one criterion may not be compensated by advantages on other criteria. Let us recall the way the non-discordance condition is implemented through $ND(a,b)$ as in Eq. (14.5). If $g_j(b) - g_j(a)$ exceeds v_j for at least one criterion then aSb is invalidated, *i.e.*, $\exists j \in F : d_j(a,b) = 1 \Rightarrow S(a,b) = 0$. This may happen even when the total concordance $C(a,b)$ is higher than the cutting level λ.

Traditionally, $ND(a,b)$ accounts both for the values of $d_j(a,b)$ and $C(a,b)$: the way $ND(a,b)$ accounts for $d_j(a,b)$ is amplified when $C(a,b)$ is low. This reflects the idea that a veto situation should be accentuated when the concordance relation is not firmly established. On the other hand, if $C(a,b)$ is high, then low values of $d_j(a,b)$ are not taken into account: the overall non-discordance relation defined in (14.5) considers the $d_j(a,b)$ only for criteria such that $d_j(a,b) > C(a,b)$.

The interplay between $d_j(a,b)$ and $C(a,b)$ in measuring discordance makes the process of eliciting veto thresholds v_j prone to misunderstandings. The typical question asked is often:

"What would be a performance difference in criterion j so large that it cannot be compensated, i.e., that would make this criterion oppose a veto to any concordant majority of other criteria?"

Suppose for instance that the previous steps of the elicitation process had let to set $p_j = 10$ and $k_j = 0.20$, for some $j \in F$. Suppose also that the answer to the previous question had led to set $v_j = 50$, possibly by "trial and error". The DM was found to have the opinion that if the performance difference is equal to 50 units or more, then there would be a veto, but if the difference was less than 50 then an outranking would be allowed. However, in this case any difference higher than 45 would necessarily veto an outranking relation:

From (14.4), $p_j = 10$, $v_j = 50$, and $\Delta_j(b,a) > 45$ imply $d_j(a,b) > 7/8$.

Even assuming that there is no other discordance and $C(a,b) = 1 - k_j = 0.8$, Eq. (14.7) together with $d_j(a,b) > 7/8$ yield $ND_j(a,b) < 0.625$.

Finally, Eqs. (14.6) and (14.11) yield $S(a,b) < 0.5$.

Since λ, the required majority, is at least 0.5, a cannot outrank b.

This means that the traditional question for eliciting a difference large enough to warrant a veto situation leads to an overestimation of this difference. A more rigorous way to question about the veto threshold, provided that criteria weights have been elicited, is the following (assuming the parameter values of this example):

"Suppose that j is the only discordant criterion, meaning that a coalition of 80% of the criteria weights agrees that aSb. What would be a performance difference in criterion j so large that it cannot be compensated, i.e., that would make this criterion oppose a veto to that coalition, even if λ was as low as 0.5?"

It the DM provided the same answer, 50, then to obtain the desired behavior it would be necessary to set:

$$v_j = p_j + \frac{C(a,b)(\Delta_j(b,a) - p_j)}{C(a,b) - 0.5(1 - C(a,b))} = 10 + \frac{0.8(50-10)}{0.8 - 0.5(1-0.8)} = 55.71429. \tag{14.14}$$

14.3.3.2 Parameters Defining $d'_j(a,b)$ for Relation $S'(a,b)$ or $S''(a,b)$

The indices $d'_j(a,b)$ are defined by (14.9) on the basis of $g_j(a)$, $g_j(b)$, a veto threshold v_j and an additional threshold u_j which we call *discordance threshold*. u_j represents the difference of evaluation $g_j(b) - g_j(a)$ above which the discordance condition starts to weaken concordance $C(a,b)$ in the definition of $S'(a,b)$. This discordance threshold u_j can be considered either:

- as an additional preferential parameter to be elicited through an interaction with the DM, or
- as a technical parameter (rather than a preference-related one), an option that should be used only when the DM does not wish to use the added flexibility offered by u_j, preferring to work with the thresholds v_j only. In such cases, a reasonable "rule-of-thumb" is to set $u_j = p_j + 0.75(v_j - p_j)$ (see Mousseau and Dias 2004).

In case the discordance threshold u_j is to be elicited, then the main difference in the use of relation $S'(a,b)$ rather than $S(a,b)$ is that criteria that intervene in the product are not restricted to those for which $d'_j(a,b) > C(a,b)$, *i.e.*, small values of $d'_j(a,b)$ will impact $ND'(a,b)$. Moreover, the concordance relation $C(a,b)$ does not intervene in the non-discordance implementation.

In model $S'(a,b)$, the discordance $d'_j(a,b)$ corresponds to a correction factor to the concordance of all other criteria taken together. One possibility is to ask two questions defining the performance differences that correspond to two distinct $d'_j(a,b)$ values, e.g., a 10% correction (decrease) and 25% correction. For the former case the question would be (the question pertaining the latter is similar):

"Suppose that j is the only discordant criterion, meaning that all other criteria agree that aSb. What would be a performance difference in criterion j that would warrant decreasing the weight of all concordant criteria by 10%?" (Note that unlike relation S there is no need to refer to the exact weight of the criteria).

If, for instance, the DM would state that $\Delta_j(b,a) = 40$ warrants decreasing the weight of all concordant criteria by 10% and $\Delta_j(b,a) = 50$ warrants decreasing the weight of all concordant criteria by 25% then, based on Eq. (14.9), solving the system

$$\begin{cases} \frac{40-u_j}{v_j-u_j} = 0.10 \\ \frac{50-u_j}{v_j-u_j} = 0.25 \end{cases} \tag{14.15}$$

leads to the solution $u_j = 100/3$ and $v_j = 100$.

It is also possible to ask only one of the above questions, and use a different question to elicit u_j:

"Suppose that j is the only discordant criterion, meaning that all other criteria agree that aSb. At what point (performance difference) would a veto effect start to occur, in that the weight of all concordant criteria would start to be decreased?"

If, for instance, the DM would reply that a veto effect gradually begins at a difference of 40, and that $\Delta_j(b,a) = 50$ warrants decreasing the weight of all concordant criteria by 25% then, based on Eq. (14.9),

$$\frac{50-40}{v_j-40} = 0.25 \text{ yields } v_j = 95. \tag{14.16}$$

14.3.4 Profiles in Sorting Problems

The elicitation of profiles in sorting problems in the framework of ELECTRE models must take into account their distinct nature in different variants of ELECTRE TRI: in the original version (ELECTRE TRI-B) the profiles are limits separating the consecutive categories, whereas in ELECTRE TRI-C the profiles are central elements of the categories.

Let us first address the original version (ELECTRE TRI-B). Here, a profile b^k separates category C^k from category C^{k+1} (it can be considered a lower bound for C^{k+1}). If there are n_{cat} categories, then $n_{cat}-1$ profiles need to be elicited. A lower bound for the first (worst) category, b^0, needs not be elicited by assuming that aSb^0 is true for every conceivable alternative a. Similarly, An upper bound for the last (best) category, $b^{n_{cat}}$, needs not be elicited by assuming that $aSb^{n_{cat}}$ is false and $b^{n_{cat}}Sa$ is true for every conceivable alternative a.

Considering the convention that C^1 is the worst category and $C^{n_{cat}}$ is the best category the following conditions should be ensured:

- Each profile dominates the profiles of lower categories: if $k' > k$ then $g_j(b^{k'}) \geq g_j(b^k)$ for criteria g_j to be maximized and $g_j(b^{k'}) \leq g_j(b^k)$ for criteria g_j to be maximized, with at least one of these inequalities being strict.
- Profiles should not be so close to each other that an alternative might be indifferent to both: for two different profiles $b^{k'}$ and b^k there is no alternative a such that $aSb^{k'}$ and $b^{k'}Sa$ and at the same time aSb^k and b^kSa.

The sorting of alternatives in ELECTRE TRI can be performed according to a pessimistic (pseudo-conjunctive) perspective or an optimistic (pseudo-disjunctive) perspective. Whenever the alternative to be sorted is incomparable to some profiles, the pessimistic perspective places it a lower category than the optimistic perspective; otherwise, both perspectives sort it in the same category. In this chapter we will consider the pessimistic perspective, according to which an alternative is sorted in a category C^k if it is good enough to outrank its lower bound but not good enough to outrank its upper bound:

$$a_i \text{ is sorted in } C^k \Leftrightarrow a_i S b^{k-1} \wedge \neg a_i S b^k \tag{14.17}$$

Elicitation of the profiles can be conducted by considering one criterion at a time. For each criterion g_j, the profiles for the successive categories can be asked in ascending order (starting from the worst one) or in descending order (starting from the best one). In descending order the performance value for $g_j(b^{n_{cat}-1})$ can be asked as follows:

"On criterion g_j, what level of performance is required for this criterion to vote in favor of sorting an alternative in the best category, $C^{n_{cat}}$?"

Then, the performance value for $g_j(b^{n_{cat}-2})$ can be asked as follows:

"On criterion g_j, what level of performance is required for this criterion to vote in favor of sorting an alternative in category $C^{n_{cat}-1}$? (if not better)"

Then, performances $g_j(b^{n_{cat}-3}),\ldots,g_j(b^1)$ would be elicited in the same way, before moving on to a different criterion. Focusing on one criterion at a time makes the task easier for decision makers, who are in this way invited to consider how each criterion would sort the alternatives, if there was not any other criterion.

As an alternative, the elicitation can focus on one category at a time, considering all the criteria, but often this task is harder. Decision makers would have to provide multi-criteria performances for a profile $b^{n_{cat}-1}$ such that all alternatives outranking it would be placed in the best category. Then, they would need to provide multi-criteria performances for a profile $b^{n_{cat}-2}$ such that all alternatives outranking it (but not outranking $b^{n_{cat}-1}$) would be placed in the second best category, and so on.

Let us now address the central profiles version ELECTRE TRI-C. Here, a profile b^k is the most representative (also called characteristic) element of category C^k. If there are n_{cat} categories, then n_{cat} profiles need to be elicited. For these profiles to be consistent, a profile for one category, say b^k, cannot be better than the profile b^{k+1} from a better category. At the minimum, $S(b^k, b^{k+1}) < 1$, but more stringent conditions such as $S(b^k, b^{k+1}) < 0.5$ or $S(b^k, b^{k+1}) < 0$ can be placed (Almeida-Dias et al. 2010). The basic idea of this method is to sort each alternative to the category such that the alternative outranks and is at the same time outranked by the profile as much as possible, i.e., with the largest $min\{S(a_i, b^k), S(b^k, a_i)\}$ (for details, see Almeida-Dias et al. 2010).

As in the case of ELECTRE TRI-B, elicitation of the profiles can be conducted by considering one criterion at a time. For each criterion g_j, the profiles for the different categories can be asked, in any order. The performance value for $g_j(b^k)$ can be asked as follows:

"On criterion g_j, what level of performance best characterizes an alternative in category, C^k?"

As an alternative, the elicitation can focus on one category at a time, considering all the criteria. In this case, a profile can be regarded as an ideal example characterizing the sort of performances the decision maker associates with each category. Method ELECTRE TRI-nC (Almeida-Dias et al. 2012), which extends ELECTRE TRI-C, even allows the decision maker to provide different examples of profiles to characterize each category.

14.4 Indirect Elicitation (Regression)

Assigning values to the parameters involved in the definition of an ELECTRE model might be a difficult task for the DM. The disaggregation approach (see Jacquet-Lagrèze and Siskos 2001) allows to infer parameter values from holistic preferences (i.e., global preferences rather than a criterion-by-criterion analysis). Holistic statements might be a ranking of a set of alternatives, comparisons of alternatives, or, in the case of sorting problems, the proposal of classification examples. The alternatives that are compared in a holistic manner might be a small subset of a much larger set of alternatives to be evaluated, or alternatives considered in past decision processes (possibly knowing how well they performed previously), or even examples constructed in a way that facilitates comparisons.

The disaggregation approach is usually performed using mathematical programs. Such inference programs can either be partial if only a subset of parameters is being inferred (the values of the other parameters being fixed), or global if all parameters are to be inferred. The inputs for the mathematical program are the holistic preference statements and the values of the parameters that are not being elicited. The decision variables are the parameters to be inferred. The objective function is to minimize an "error function" measuring how well the holistic preferences are reproduced by the inferred model. The constraints reflect the holistic preferences and also constraints that the method imposes on the model (e.g., weights are nonnegative and they add up to 1).

As described in Sect. 14.2.5, in ELECTRE methods the final choice set, or ranking, or sorting result is derived from the outranking relation. For ELECTRE TRI's pessimistic (or pseudo-conjunctive) variant, a statement in the form of a sorting example can be translated in two statements concerning outranking relations (Mousseau and Dias 2004). For instance, a statement "alternative a should be classified at least in the second category and at most in the third category" is translated into two outranking statements: "a outranks the lower profile of the second category" and "a does not outrank the lower profile of the fourth category". Unfortunately, statements based on ELECTRE methods devoted to choice or ranking problems, such as ELECTRE I-IV, do not have an easy direct translation into outranking statements. Therefore, the literature has concentrated on the cases of sorting problems or inferring parameters from outranking statements.

In order to elicit values for preference-related parameters (*i.e.*, w_j, $v_j(g_j)$, $p_j(g_j)$, $q_j(g_j)$, and limits of categories in ELECTRE TRI) it is possible to proceed using a disaggregation procedure that infers the parameters values from holistic preferences provided by the DM. Hence, it is necessary to formalize $S(a,b)$ through an optimization program that minimizes an "error function" that measures how much the values of the inferred parameters contradict the stated holistic preferences. However, $S(a,b)$ is rather "optimization unfriendly". Difficulties arise mainly from the way the non-discordance condition is implemented, *i.e.*, the way $ND(a,b)$ is defined.

More precisely, two features of the non-discordance relation are concerned. First, the subset of criteria $\overline{F}$ (see (14.5)) is difficult to integrate into an optimization program. Second, the fact that $C(a,b)$ intervenes in the definition of $ND(a,b)$ implies that the optimization program will necessarily be non-linear, even when all the parameters are fixed except the weights.

The problem of inferring the parameters of an ELECTRE method (ELECTRE TRI) based sorting examples translated into outranking statements was initially studied by Mousseau and Slowinski (1998). The resulting mathematical programming was nonlinear and would require global optimization techniques to find a solution. A simpler formulation was proposed to infer only the weights and the cutting level in situations without veto thresholds, in which $S(a,b) = C(a,b)$. In this case, an easy to solve linear programming formulation could be devised.

If veto thresholds are allowed, then the problem can no longer be solved by linear programming, even if the only parameters to be inferred are weights and the cutting level. Indeed, $S(a,b)$ is a non-differentiable and quasi-concave nonlinear function of the weights in the domain where it is strictly positive and therefore a constraint like $S(a,b) < \lambda$ (which reflects a holistic statement of the form $\neg aSb$) does not define a convex set (Dias and Climaco 1999). For this reason, Mousseau and Dias (2004) proposed variants for the outranking relation, S' and S'' (presented in Sect. 14.2.4) that allow using linear programming in such cases.

To provide an example of the mathematical programming approach to inference, the following section briefly recalls the inference of weights and cutting level for relation S'. The ensuing section overviews the literature on eliciting other subsets of parameters.

14.4.1 *Inferring Weights and Cutting Level from S' Outranking Statements*

Let us suppose that the DM is not able (or not willing) to assign directly values to the preference-related parameters involved in the outranking relation, but can state crisp statements about this relation for some specific pairs of alternatives (a,b), *i.e.*, either aSb (a outranks b) or $\neg aSb$ (a does not outrank b). Our purpose is to find criteria weights and a cutting level that restore the DM's statements.

Let A denote a set of alternatives. Let $S^+ = \{(a,b) \in A^2$ such that the DM stated $aSb\}$ and $S^- = \{(a,b) \in A^2$ such that the DM stated $\neg aSb\}$. Then, a combination of parameter values is able to restore the DM's request iff $S(a,b) \geq \lambda,\ \forall (a,b) \in S^+$ and $S(a,b) < \lambda,\ \forall (a,b) \in S^-$, which may be written as $S(a,b) - \lambda \geq 0,\ \forall (a,b) \in S^+$ and $\lambda - S(a,b) + \varepsilon \geq 0,\ \forall (a,b) \in S^-$ (ε being a small positive value).

The mathematical program given below (14.18)–(14.23) maximizes a common slack α for all these constraints, to obtain a relatively "central" combination of parameter values. Whenever the optimum value of α is negative, there is no combination of parameter values complying to all the constraints, *i.e.*, the DM

provided inconsistent information (a procedure to deal with such inconsistencies is proposed in Mousseau et al. 2003). Alternative objective functions can be considered (see Beuthe and Scannella 2001 and Mousseau and Slowinski 1998).

$$Max \ \alpha \tag{14.18}$$

$$s.t. \ \alpha \leq S(a,b) - \lambda, \quad \forall (a,b) \in S^{+} \tag{14.19}$$

$$\alpha \leq \lambda - S(a,b) + \varepsilon, \quad \forall (a,b) \in S^{-} \tag{14.20}$$

$$\lambda \in [0.5, 1] \tag{14.21}$$

$$v_j(g_j) > p_j(g_j) > q_j(g_j) \geq 0, \quad \forall j \in F \tag{14.22}$$

$$\sum_{j=1}^{n} w_j = 1; \quad w_j \geq \varepsilon, \ \forall j \in F. \tag{14.23}$$

Some additional constraints can be added to this program, in order to integrate explicit statements of the DM concerning the values of some parameters. From (14.5) and (14.11), it is obvious that this is a difficult nonlinear program if all the parameters were considered as variables. A solution to circumvent this difficulty is to formulate partial inference programs, where only a subset of the parameters are considered as variables, while the remaining ones are elicited by other means. Among the partial inference problems, previous research on related problems has focused mainly on inferring the weights and the cutting level (see Mousseau et al. 2000; Dias et al. 2002; Miettinen and Salminen 1999). This is an important partial inference problem because the weights and the cutting level are the only parameters involving inter-criteria judgements (the remaining parameters do not interrelate the criteria).

Let us consider the case where S' is used. In this case each product $\prod_{j \in F}(1 - d'_j(a,b)) = ND'(a,b)$ is a fixed constant $\forall (a,b)$. The following constraints concerning outranking statements are hence linear, since $C(a,b)$ is an affine function of the weights.

$$\alpha \leq C(a,b) \prod_{j \in F} \left(1 - d'_j(a,b)\right) - \lambda, \quad \forall (a,b) \in S^{+} \tag{14.24}$$

$$\alpha \leq \lambda - C(a,b) \prod_{j \in F} \left(1 - d'_j(a,b)\right) + \varepsilon, \quad \forall (a,b) \in S^{-} \tag{14.25}$$

Considering $S'(a,b)$ instead of $S(a,b)$, the weights and the cutting level can be inferred by solving a linear program whose variables are α, $w_1, \ldots, w_n$, and λ, where (14.24) and (14.25) appear as (14.27) and (14.28):

$$Max \ \alpha \tag{14.26}$$

$$s.t. \ \alpha \leq \sum_{j=1}^{n} w_j \, S_j(a,b) \, ND'(a,b) - \lambda, \quad \forall (a,b) \in S^{+} \tag{14.27}$$

$$\boldsymbol{\alpha} \leq \lambda - \sum_{j=1}^{n} w_j \, S_j(a,b) \, ND'(a,b) + \varepsilon, \quad \forall (a,b) \in S^- \tag{14.28}$$

$$\lambda \in [0.5, 1], \tag{14.29}$$

$$\sum_{j=1}^{n} w_j = 1 \quad w_j \geq \varepsilon, \ \forall j \in F \tag{14.30}$$

If the maximum value of α is positive, then the values of $w_1, \ldots, w_n$, and λ at the optimum are able to restore all the statements defining S^+ and S^-. Otherwise, the inferred values provide suggestions for changing those examples. The DM should ponder whether they want to change the sets S^+ and S^-, or to analyze the values of $ND'(a,b)$. Indeed, some of the differences among the current model and the DM's requests may stem from inadequate values for the veto and discordance thresholds. Considering $S''(a,b)$ instead of $S'(a,b)$ leads to a similar linear program.

As a particular case, the pessimistic procedure of ELECTRE TRI assigns alternative a to category C_h (b_{h-1} and b_h being the lower and upper profiles of C_h, respectively) iff $S(a, b_{h-1}) \geq \lambda$ and $S(a, b_h) < \lambda$ ($\lambda \in [0.5, 1]$ is the chosen cutting level).

Suppose the DM has specified a set of assignment examples, *i.e.*, a subset of $A^* \subset A$ such that each $a_k \in A^*$ is associated with $C^M(a_k)$ ($C^m(a_k)$, respectively) the maximum (minimum, respectively) category to which a should be assigned according to his/her holistic preferences. Hence $[C^m(a_k), C^M(a_k)]$ defines an interval of possible categories to which a_k can be assigned to. $C^m(a_k) = C^M(a_k) = C_{h_k}$ means that the DM wants a_k to be assigned to C_{h_k} precisely (we will note $a_k \rightarrow_{DM} C_{h_k}$ such statement), while $C^m(a_k) < C^M(a_k)$ corresponds to an imprecise statement ($a_k \rightarrow_{DM} [C^m(a_k), C^M(a_k)]$).

Inferring all ELECTRE TRI parameters is a difficult nonlinear program (Mousseau and Slowinski 1998). But if we consider $S'(a,b)$ instead of $S(a,b)$, the weights and the cutting level can be inferred by solving a linear program (all other parameters being given as inputs). The linear program for this partial inference problem is equal to (14.26)–(14.30) if we define:

$$S^+ = \left\{(a_k, b_{C^m(a_k)-1}) \in A^* \times B \ : \ a_k \rightarrow_{DM} [C^m(a_k), C^M(a_k)]\right\} \tag{14.31}$$

$$S^- = \left\{(a_k, b_{C^M(a_k)}) \in A^* \times B \ : \ a_k \rightarrow_{DM} [C^m(a_k), C^M(a_k)]\right\} \tag{14.32}$$

Considering $S''(a,b)$ instead of $S'(a,b)$ leads to a similar linear program.

14.4.2 Inferring Different Parameters for Sorting Problems

In recent years, several papers dealt with the learning of ELECTRE TRI parameters.

As mentioned previously, the first paper devoted to the learning of ELECTRE TRI parameters has been proposed by Mousseau and Slowinski (1998). The learning

algorithm takes as input a set of assignment examples and their associated vector of performances with respect to the problem criteria. The paper shows the difficulties to learn the parameters of ELECTRE TRI without veto. The main difficulty is the non-linearity of the partial concordance indices. Indeed, it makes the concordance index not differentiable which prevents the use of gradient optimization algorithms. In order to tackle this difficulty, Mousseau and Slowinski (1998) propose to approximate the partial concordance indices by sigmoid functions.

Learning all the parameters of an ELECTRE TRI model involves the determination of a lot of parameters. It requires a lot of cognitive effort from the user. Mousseau et al. (2001) consider the subproblem of finding the weights and the cutting threshold of an ELECTRE TRI model with fixed profiles and indifference and preference thresholds. In the paper, a linear program is proposed and some experiments are conducted. It shows that learning only a subpart of the ELECTRE TRI model simplifies the problem. Fewer assignment examples are required to obtain good results.

Ngo The and Mousseau (2002) proposed a mixed integer program in order to infer the profiles of an ELECTRE TRI model with fixed weights and thresholds. The mixed integer program presented in the paper finds the partial concordance indices in a first step. The second step consists in deducing the values of the profiles from the partial concordance indices. They propose to use this mixed integer program in combination with the linear program of Mousseau et al. (2001) in order to determine the whole set of parameters of an ELECTRE TRI model.

Mousseau and Slowinski (1998), Mousseau et al. (2001) and Ngo The and Mousseau (2002) consider only ELECTRE TRI models without veto. Dias and Mousseau (2006) present a manner to learn vetoes of an ELECTRE TRI model with fixed profiles, thresholds and weights. In the paper, two subproblems are treated. The first one considers the inference of veto parameters for a single criterion. The second considers the inference of all veto parameters for multiple criteria at the same time.

Doumpos et al. (2009) proposed a metaheuristic in order to learn all the parameters of an ELECTRE TRI model, including the veto thresholds. They developed a genetic algorithm in order to learn all the parameters of the model at the same time. The interest of this approach is that it allows to deal with larger data sets than mixed integer program based algorithms.

However ELECTRE TRI integrates a large number of preference parameters that are to be determined. MR-SORT is a simplified version of ELECTRE TRI which keeps the philosophy of ELECTRE TRI with the advantage of using less parameters (no veto thresholds and no discrimination thresholds are considered). Leroy et al. (2011) propose a mixed integer program in order to learn the parameters of such a model based on assignment examples. The experimental results presented in the paper show that the mixed integer program is able to find MR-SORT models which perform well in generalization. However, the experiments show the limitation of such an algorithm in terms of computing time. For a small problem involving five categories and three criteria, more than 100 s are required to restore all the parameters of a MR-SORT model on the basis of 100 assignment examples.

Damart et al. (2007) are the first to consider the problem of learning the parameters of an ELECTRE TRI model in the context of multiple decision makers. They propose an approach that aims at determining a set of fictitious alternatives that contain enough information to obtain a model that is satisfactory for all the DMs. The procedure is applied to an illustrative example.

Later, Cailloux et al. (2012) developed two mixed integer programs in order to learn the parameters of a MR-SORT model in the context of multiple DMs. The first mixed integer program aims at finding a set of profiles that is common to all the decision makers. The second mixed integer program learns a set of weights compatible with the preferences of each DM. The paper presents experimental results on real and fictitious applications.

Recently, Sobrie et al. (2013), Sobrie (2016) proposed an heuristic to efficiently infer MR-SORT parameters (weights and profiles) from large sets of assignment examples (over several thousands).

14.5 Elicitation Process

After reviewing elicitation techniques, we now focus on elicitation as a process that evolves in time, involving at least one DM and an analyst conducting the process. Two issues are discussed: elicitation sequence and numerical precision.

14.5.1 Elicitation Sequence

The elicitation sequence defines which parameters are elicited, in which order (or simultaneously), and using which technique.

All the parameters of an ELECTRE model should be discussed with the DMs, but not necessarily elicited from them. Indeed, there are at least three situations in which some parameters are not elicited:

- Indifference and preference thresholds, unlike preference-based parameters such as weights, may be considered technical parameters (Rogers and Bruen 1998; Roy et al. 2014) that can be set by the analyst, possibly with the help of experts on the domain that a criterion refers to. For instance, an analyst may set both thresholds equal to zero if a scale is ordinal, or an expert may set these thresholds based on considerations about the method that measures the performance of the alternatives on a cardinal scale, or a scientist may inform which differences in, say, noise levels, are negligible because a human cannot perceive them (Rogers and Bruen 1998).
- Veto thresholds may not be necessary, at least for all the criteria. The DMs may deem that no veto power is granted to some criteria, meaning that the discordance from those criteria is always null.

- The DMs may feel uncomfortable about setting criteria weights. In such cases, they may resort to ELECTRE IV, a method that does not ask for weights (Roy and Hugonnard 1982), or they may consider some freedom in setting the weights, as discussed in the following section. Many DMs may simply ask that all criteria have the same weight, but such a conclusion should result from (or be confirmed by) elicitation questions (Sect. 14.3.2).

There is no mandatory order by which parameters should be elicited. A possible sequence is the one followed by Sect. 14.3. Indifference and preference thresholds are clearly related and thus should be elicited simultaneously, one criterion at a time. Then, since the concordance part of the outranking relation is being addressed, the elicitation of weights may ensue. If the cutting level λ is communicated as a required majority level, then this parameter can be discussed simultaneously with weights, as described in Sect. 14.3.2. Finally, the possibility of veto is discussed, eliciting veto and non-discordance thresholds.

A different strategy is to initially focus on one criterion at a time and elicit indifference, preference, discordance and veto thresholds for each criterion. Then, criteria weights and the cutting level, which interrelate multiple criteria, would be elicited.

When an indirect elicitation (regression approach) is followed, multiple types of parameters can be inferred simultaneously, although that is a difficult optimization problem. Inferring only a subset of the parameters at a time allows overcoming this difficulty, and has an additional advantage. Since the DMs interactively revise the information they provide and observe the results of the mathematical program, partial inference problems allow them to focus their attention on a subset of parameters at a time and to better understand the consequences of modifying the examples they provide. We believe that inference programs should not be considered as a problem to be solved once, but rather as problems to be solved many times throughout an interactive learning process. Furthermore, it is possible to mix direct and indirect elicitation questions for different sets of parameters, and even for the same parameters (for confirmation purposes). Finally, the notion that parameters are elicited in a sequence does not mean that the elicitation process is linear. Often, the analyst may find out that the discussion concerning a subset of parameters puts into question the values elicited previously for another subset of parameters.

14.5.2 Numerical Precision

The issue of precision (and accuracy) arises in both direct and indirect elicitation. By precision we mean the freedom of variation one accepts for a parameter. For instance, setting the weight of the first criterion as $w_1 = 0.288$ is more precise than setting $w_1 \in [0.28, 0.29]$, which is more precise than setting or $w_1 \in [0.25, 0.30]$. The elicitation process is developed during a finite time window in which the DMs are available (and attentive!). Therefore, one has to accept the elicitation results

might not be "accurate" in the sense that they include the exact parameter values that would result from a much longer process. In a direct elicitation process, a DM would hardly state that $w_1 = 0.288$. Probably he or she would state 0.29 or 0.3 which are "rounder" numbers. Typically these inputs are accepted even knowing they might be slightly inaccurate: no analyst would ask if it should really be a value of 0.299 or 0.301 instead of 0.3. Analysts know that rounder numbers are more comfortable for the DMs and reckon it would not be worthwhile to trouble a DM for a degree of precision that might be irrelevant to the results of the analysis. These concerns can be addressed at the end by means of a sensitivity or a robustness analysis (Roy 1998).

In indirect elicitation processes the mathematical programs might admit many different solutions able to reproduce the examples provided by DMs. For instance, experimental studies have been developed (Mousseau et al. 2001) showing that to infer relevant values for w_j and λ, the cardinality of S^+ and S^- should be "sufficiently" large. On the other hand, accepting less precision leads to higher confidence that the elicitation results (a subset of the parameter space) contains the parameter vector that would result from an ideally long elicitation process.

There are two possible outcomes of an indirect elicitation process: a set of constraints defining a partial information set (a subset T of the parameter space) or a (precise) vector of parameter values $t^* \in T$ (the best fit found by a mathematical program). For instance, the IRIS implementation (Dias and Mousseau 2003) of an indirect elicitation process for ELECTRE TRI (Dias et al. 2002) infers a suggested parameter vector and displays the resulting sorting of the alternatives, but it always displays all other sorting possibilities that are compatible with examples and other constraints provided by DMs.

Often, precision is not required for a model to be requisite (as defined by Phillips 1984). The analyst can follow a strategy of progressively reducing the variation for the parameters by means of new questions depending on the observation of results that are robust relatively to information provided before Dias (2007). The process stops when the DMs feel the precision in the results is requisite for their purposes. As an example we can mention an application for sorting plots of land according to their suitability for photovoltaic plants (Sánchez-Lozano et al. 2014). A subset of 20 plots was considered as potential sorting examples. At the outset, an interval of weights was considered based on the maximum and minimum values indicated by a panel of stakeholders. Then, a DM observed the range of categories in which each plot could be sorted given their characteristics and the weight intervals considered. The DM then sorted a few of these plots according to his experience-based opinion, one at a time, and observed how the range of possible categories for each plot was reduced as a result of the new constraints associated with the example. After sorting the seventh plot the number of constraints collected defined a region in the parameter space that was sufficiently precise to be able to sort each one of the remaining 13 plots into a single category. The model was considered to be requisite, concluding the elicitation process.

Setting a precise figure for each parameter value may also be an elusive goal when seeking the agreement of multiple DMs, due to differences in their

preferences. It is easier for them to agree that $w_1 \in [0.25, 0.30]$ than to agree that $w_1 = 0.288$, and often conclusions are robust to vector variations within a subset of the parameter space. DMs may agree on a result although they would not be able to agree on precise values for the input parameters (Dias and Clímaco 2000). In such cases, DMs can start with little information and progressively constrain the subset of the parameter space they consider.

Avoiding eliciting precise figures is also a possibility to cope with situations in which DMs do not wish to set criteria weights, particularly in sensitive situations (e.g., impacts on the environment and on human health, or social impacts). Such DMs wish to treat criteria in a value-neutral way. An alterative to considering all criteria have the same weight is to consider that all criteria share a common interval of weights (for an example, see Domingues et al. 2015). This makes no distinction between the criteria importance, but does not entail they have the same weight. In this case, DMs would discuss the acceptable interval of weights for the criteria, discussing for instance that no criterion should weight more than all other criteria ($k_j < 0.5$), or defining a maximum acceptable ratio between any two weights (e.g., a criterion's weight cannot be more than α times greater than any other criterion's weight, Domingues et al. 2015).

14.6 Concluding Remarks

A large literature exists concerning the way by which ELECTRE methods can be implemented in practice and in particular with respect to the integration of the DM judgement in the preference model. Preference elicitation for ELECTRE methods have been largely developed and this chapter provides a synthesis of the corresponding literature.

However, there are still many challenges to be faced. An important one concerns the indirect elicitation of ELECTRE models for ranking problems: as ELECTRE methods are not invariant with respect to third a alternative, i.e, a DM can provide a statement "a is preferred to b", the inferred model will reproduce this comparison, but when applied to rank a larger set of alternatives, b can be better ranked than a.

Another challenge related to inference of ELECTRE model is related to the multiplicity of preference parameters. When eliciting indirectly these preference parameters, we usually can obtain a rather limited amount of preference statements (e.g. pairwise comparisons, or assignment examples). The contrast of the great flexibility of the preference models with the limited preference information makes it difficult to set the values of the preference parameters without some form of arbitrariness. In some applications, it might be relevant to consider some simplification of the original ELECTRE methods (avoiding some of the parameters). Another path is to collect la large amount of preference information, but this implies computational challenges related to the inference of ELECTRE models with large sets of preference statements.

A third challenge is to elicit and integrate "soft" requests, such as "I would like that criteria weights are not too different", or "I would like that more important criteria have greater veto power than the remaining ones" in direct and especially in indirect elicitation processes.

Finally, group decision making places many different challenges. A strategy to deal with lack of agreement is working with less precise information, as suggested in the previous section. But if the DMs wish to somehow aggregate their opinions assigning different weights for the DMs's requests (e.g. reflecting their expertise or past performance), then there is lack of research on how to take this into account in eliciting ELECTRE's parameter values.

References

Almeida-Dias J, Figueira JR, Roy B (2010) ELECTRE TRI-C: a multiple criteria sorting method based on characteristic reference actions. Eur J Oper Res 204(3):565–580

Almeida-Dias J, Figueira JR, Roy B (2012) A multiple criteria sorting method where each category is characterized by several reference actions: the Electre Tri-nC method. Eur J Oper Res 217(3):567–579

Behzadian M, Kazemzadeh RB, Albadvi A, Aghdasi M (2010) PROMETHEE: a comprehensive literature review on methodologies and applications. Eur J Oper Res 200(1):198–215

Beuthe M, Scannella G (2001) Comparative analysis of UTA multicriteria methods. Eur J Oper Res 130(2):246–262

Bisdorff R, Meyer P, Roubens M RUBIS: a bipolar-valued outranking method for the choice problem. 4OR 6(2):143–165 (2007)

Brans JP, Vincke P (1985) A preference ranking organization method. Manag Sci 31(6):647–656

Cailloux O, Meyer P, Mousseau V (2012) Eliciting ELECTRE TRI category limits for a group of decision makers. Eur J Oper Res 223(1):133–140

Damart S, Dias LC, Mousseau V (2007) Supporting groups in sorting decisions: Methodology and use of a multi-criteria aggregation/disaggregation DSS. Decis Support Syst 43(4):1464–1475

Dias LC (2007) A note on the role of robustness analysis in decision-aiding processes. In: Roy B, Ali Aloulou M, Kalaï R (eds) Robustness in OR-DA, Annales du LAMSADE, No. 7. Université-Paris Dauphine, Paris, pp 53–70

Dias LC, Climaco JN (1999) On computing ELECTRE's credibility indices under partial information. J Multi-Criteria Decis Anal 8(2):74–92

Dias LC, Clímaco JN (2000) ELECTRE TRI for groups with imprecise information on parameter values. Group Decis Negot 9(5):355–377

Dias LC, Mousseau V (2003) IRIS: a DSS for multiple criteria sorting problems. J Multi-Criteria Decis Anal 12(4-5):285–298

Dias L, Mousseau V (2006) Inferring ELECTRE's veto-related parameters from outranking examples. Eur J Oper Res 170(1):172–191

Dias LC, Mousseau V, Figueira J, Clímaco JN (2002) An aggregation/disaggregation approach to obtain robust conclusions with ELECTRE TRI. Eur J Oper Res 138(2):332–348

Domingues AR, Marques P, Garcia R, Freire F, Dias LC (2015) Applying multi-criteria decision analysis to the life-cycle assessment of vehicles. J Clean Prod 107:749–759

Doumpos M, Marinakis Y, Marinaki M, Zopounidis C (2009) An evolutionary approach to construction of outranking models for multicriteria classification: the case of the ELECTRE TRI method. Eur J Oper Res 199(2):496–505

Fernández E, Figueira JR, Navarro J, Roy B (2017) ELECTRE TRI-nB: a new multiple criteria ordinal classification method. Eur J Oper Res 263(1):214–224. DOI 10.1016/j.ejor.2017.04.048.

Figueira JR, Roy B (2008) A note on the paper, "Ranking irregularities when evaluating alternatives by using some ELECTRE methods", by Wang and Triantaphyllou. Omega 37(3):731–733

Figueira J, Greco S, Roy B, Slowinski R (2013) An overview of ELECTRE methods and their recent extensions. J Multi-Criteria Decis Anal 20:61–85

Govindan K, Jepsen MB (2016) ELECTRE: a comprehensive literature review on methodologies and applications. Eur J Oper Res 250(1):1–29

Jacquet-Lagrèze E, Siskos Y (2001) Preference disaggregation: 20 years of MCDA experience. Eur J Oper Res 130(2):233–245

Keeney RL, Raiffa H (1993) Decisions with multiple objectives-preferences and value tradeoffs. Cambridge University Press, Cambridge, New York

Leroy A, Mousseau V, Pirlot M (2011) Learning the parameters of a multiple criteria sorting method. In: Brafman R, Roberts F, Tsoukiàs A (eds) Algorithmic decision theory. Lecture notes in artificial intelligence, vol 6992. Springer, Berlin, pp 219–233

Martel J-M, Matarazzo B (2005) Other outranking approaches. In: Figueira J, Greco S, Ehrgott M (eds) Multiple criteria decision analysis: state of the art surveys. Springer, Berlin, pp 198–219

Matsatsinis NF, Grigoroudis E, Siskos E (2018) Disaggregation approach to value elicitation. In: Dias LC, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgement. Springer, New York

Miettinen K, Salminen P (1999) Decision-aid for discrete multiple criteria decision making problems with imprecise data. Eur J Oper Res 119(1):50–60

Morton A (2018) Multiattribute value elicitation. In: Dias LC, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgement. Springer, New York

Mousseau V, Dias LC (2004) Valued outranking relations in ELECTRE providing manageable disaggregation procedures. Eur J Oper Res 156(2):467–482

Mousseau V, Slowinski R (1998) Inferring an ELECTRE TRI model from assignment examples. J Glob Optim 12(2):157–174

Mousseau V, Slowinski R, Zielniewicz P (2000) A user-oriented implementation of the ELECTRE TRI method integrating preference elicitation support. Comput Oper Res 27(7-8):757–777

Mousseau V, Figueira J, Naux JP (2001) Using assignment examples to infer weights for ELECTRE TRI method: Some experimental results. Eur J Oper Res 130(2):263–275

Mousseau V, Dias LC, Figueira J, Gomes C, Clímaco JN (2003) Resolving inconsistencies among constraints on the parameters of an MCDA model. Eur J Oper Res 147(1):72–93

Munda G (1995) Multicriteria evaluation in a fuzzy environment. Physica Verlag, Heidelberg

Ngo The A, Mousseau V (2002) Using assignment examples to infer category limits for the ELECTRE TRI method. J Multi-Criteria Decis Anal 11(1):29–43

Phillips LD (1984) A theory of requisite decision models. Acta Psychol 56:29–48

Rogers M, Bruen M (1998) Choosing realistic values of indifference, preference and veto thresholds for use with environmental criteria within ELECTRE. Eur J Oper Res 107(3):542–551

Rolland A (2013) Reference-based preferences aggregation procedures in multi-criteria decision making. Eur J Oper Res 225(3):479–486

Roy B (1968) Classement et choix en présence de points de vue multiples: Le méthode ELECTRE. Revue Francaise d'Informatique et de Recherche Opérationnelle. 8:57–75

Roy B (1971) La méthode ELECTRE II. Technical report. METRA, Direction Scientifique, Note de Travail n. 142

Roy B (1978) ELECTRE III: Un algorithme de classements fondé sur une représentation floue des préférences en présence de critères multiples. Cahiers du CERO 20(1):3–24

Roy B (1991) The outranking approach and the foundations of ELECTRE methods. Theory Decis 31(1):49–73

Roy B (1996) Multicriteria methodology for decision aiding. Kluwer Academic, Dordrecht

Roy B (1998) A missing link in OR-DA, Robustness analysis. Found Control Eng 23(3):141–160

Roy B, Bertier P (1973) La méthode ELECTRE II - une application au média-planning. In: Ross M (ed) OR'72. North-Holland Publishing Company, Amsterdam, pp 291–302

Roy B, Bouyssou D (1993) Aide Multicritère à la Décision: Méthodes et Cas. Economica, Paris

Roy B, Figueira J (2002) Determining the weights of criteria in the ELECTRE type methods with a revised Simos' procedure. Eur J Oper Res 139:317–326

Roy B, Hugonnard JC (1982) Ranking of suburban line extension projects of the Paris metro system by a multicriteria method. Transp Res 16A:301–322

Roy B, Skalka JM (1984) ELECTRE IS : Aspects méthodologiques et guide d'utilisation. Technical report, Document du LAMSADE N^o 30, Université Paris-Dauphine, Paris

Roy B, Vanderpooten D (1996) The European school of MCDA: emergence, basic features and current works. J Multi-Criteria Decis Anal 5(1):22–37

Roy B, Figueira JR, Almeida-Dias J (2014) Discriminating thresholds as a tool to cope with imperfect knowledge in multiple criteria decision aiding: theoretical results and practical issues. Omega 43:9–20

Sánchez-Lozano JM, Antunes CH, García-Cascales MS, Dias LC (2014) GIS-based photovoltaic solar farms site selection using ELECTRE-TRI: Evaluating the case for Torre Pacheco, Murcia, Southeast of Spain. Renew Energy 66:478–494

Simos J (1990) Evaluer limpact sur lenvironnement: Une approche originale par l'analyse multicritère et la négociation. Presses Polytechniques et Universitaires Romandes, Lausanne

Sobrie O (2016) Learning preferences with multiple-criteria models. PhD thesis, UMONS and Université Paris Saclay

Sobrie O, Mousseau V, Pirlot M (2013) Learning a majority rule model from large sets of assignment examples. In: Perny P, Pirlot M, Tsoukiás A (eds) Algorithmic decision theory. Lecture notes in artificial intelligence, vol 8176. Springer, Brussels, pp 336–350

Vallée D, Zielniewicz P (1994) ELECTRE III-IV, version 3.x, Aspects Méthodologiques (tome 1), Guide d'utilisation (tome 2). Document du LAMSADE no. 85 et 85 bis, Université de Paris Dauphine, France

Yu W (1992a) Aide multicritère à la décision dans le cadre de la problématique du tri: méthodes et applications. PhD thesis, LAMSADE, Université Paris Dauphine, Paris

Yu W (1992b) ELECTRE TRI : Aspects méthodologiques et manuel d'utilisation. Document du lamsade no 74, Université Paris-Dauphine

Chapter 15
Individual and Group Biases in Value and Uncertainty Judgments

Gilberto Montibeller and Detlof von Winterfeldt

Abstract Behavioral decision research has demonstrated that value and uncertainty judgments of decision makers and experts are subject to numerous biases. Individual biases can be either cognitive, such as overconfidence, or motivational, such as wishful thinking. In addition, when making judgements in groups, decision makers and experts might be affected by group-level biases. These biases can create serious challenges to decision analysts, who need judgments as inputs to a decision or risk analysis model, because they can degrade the quality of the analysis. This chapter identifies individual and group biases relevant for decision and risk analysis and suggests tools for debiasing judgements for each type of bias.

15.1 Introduction

Behavioral decision research has identified a large number of behavioral individual biases in human judgment and decision making. Most of its findings address cognitive biases—faulty mental processes that lead judgments and decisions to violate commonly accepted normative principles. Equally important, but much less studied, are motivational biases, which include conscious or subconscious distortions of judgments and decisions because of self-interest, social pressures, or organizational context. Beyond these individual level biases, group biases are also relevant, whenever judgments involve teams of experts or groups of decision makers.

These biases create serious challenges to decision analysts, who want to use judgments of experts and preferences of decision makers as inputs to their analyses. This chapter identifies the individual and group biases relevant for decision and risk analysis and suggests debiasing tools. We start by exploring the relevant cognitive and individual biases, drawing from a recent literature review that we conducted

G. Montibeller (✉)
Loughborough University, Loughborough, UK
e-mail: G.Montibeller@lboro.ac.uk

D. von Winterfeldt
University of Southern California, Los Angeles, CA, USA

L.C. Dias et al. (eds.), *Elicitation*, International Series in Operations Research & Management Science 261, https://doi.org/10.1007/978-3-319-65052-4_15

on this topic (Montibeller and von Winterfeldt 2015). We then extend this review to group-level biases, suggesting which biases are relevant for decision and risk analysis. The chapter concludes with some directions for further.

15.2 Relevant Individual Biases

We distinguish two groups of individual biases which are relevant in the elicitation of value and uncertainty judgements: cognitive biases and motivational biases. For each relevant bias we will mention some common debiasing tools, drawing from our experience, best practices in decision and risk analysis, as well as from the limited literature on the topic (Arkes 1991; Larrick 2007; Milkman et al. 2009). Debiasing refers to attempts to eliminate, or at least reduce, cognitive or motivational biases.

15.2.1 Relevant Individual Cognitive Biases

A cognitive bias is a systematic discrepancy between the "correct" answer in a judgmental task, given by a formal normative rule, and the decision maker's or expert's actual answer to such a task (von Winterfeldt and Edwards 1986). There is a vast literature on cognitive biases and excellent compilations of papers are provided in Kahneman et al. (1982) and Gilovich et al. (2002).

We distinguish between cognitive biases that are relevant for decision and risk analysis and those that are less or not at all relevant. Relevant cognitive biases are difficult to correct in decision and risk analysis processes. Biases that are difficult to correct tend to be resistant to logic, decomposition, or the use of training and tools. Examples of these biases are the overconfidence bias (Lichtenstein et al. 1982; Lichtenstein and Fischhoff 1977), anchoring and insufficient adjustment (Tversky and Kahneman 1974), and the equalizing bias (Jacobi and Hobbs 2007).

In contrast there are biases that are less or not relevant to decision and risk analysis, because they are easy to correct by logic and decomposition. Examples of biases that are easy to correct are the conjunction fallacy (Tversky and Kahneman 1983), which can be corrected by demonstrating the probability logic, and the neglect of base rates (Bar-Hillel 1980; Kahneman and Tversky 1973), which can be eliminated by eliciting base rates and conditional probabilities separately.

We list each relevant bias below and suggest debiasing techniques (for details see Montibeller and von Winterfeldt 2015):

Anchoring The estimation of a numerical value is based on an initial value (anchor), which is then insufficiently adjusted to provide the final answer (Tversky and Kahneman 1974).

There is evidence of this bias occurring in several types of judgments, such as estimation tasks, pricing decisions and also in negotiations (Furnham and Boo 2011; Mussweiler and Strack 2001).

Debiasing includes avoiding anchors, providing multiple and counter-anchors, and using different experts who use different anchors.

Availability/Ease of Recall The probability of an event that is easily recalled is overstated (Bazerman and Moore 2013; Tversky and Kahneman 1973).

This bias occurs in judgments of simple frequency estimates (Tversky and Kahneman 1973; Wänke et al. 1995); frequency of lethal events (Lichtenstein et al. 1978); rare events that are anchored on recent examples.

Debiasing techniques include conducting probability training, providing counter examples, and providing statistics.

Certainty Effect Decision makers prefer sure things to gambles with similar expected utilities; they discount the utility of sure things dramatically, when they are no longer certain (Allais 1953; Kahneman and Tversky 1979).

This bias occurs in elicitation of judgments employing probability vs. certainty equivalent methods, which produce different results (Hershey and Schoemaker 1985; Schoemaker and Hershey 1992).

Debiasing techniques include avoiding sure things in utility elicitation, separating value and utility elicitation, and exploring relative risk attitude parametrically.

Equalizing Bias Decision makers allocate similar weights to all objectives (Jacobi and Hobbs 2007) or similar probabilities to all events (Fox et al. 2005; Fox and Clemen 2005).

This bias has been observed in the elicitation of probabilities in decision trees (Fox et al. 2005; Fox and Clemen 2005) and elicitation of weights in value trees (Jacobi and Hobbs 2007).

Debiasing techniques include ranking events or objectives first then assigning ratio weights and by eliciting weights or probabilities hierarchically.

Gain–Loss Bias Alternative descriptions of a choice and its outcomes (Tversky and Kahneman 1981), either as gains or as losses, may lead to different answers (Frisch 1993; Levin et al. 1998; Tversky and Kahneman 1981).

There are several types of judgments where this bias occurs, involving choices of risky options, evaluation of a single option on an attribute, and the way consequences are described to promote a choice (Kühberger 1998; Levin et al. 1998).

Debiasing techniques include clearly identifying the status quo (SQ), expressing values as marginal changes from SQ for value functions, eliciting utilities for gains and losses separately for utility functions and by cross checking utilities for mixed gambles to ensure consistency.

Myopic Problem Representation An oversimplified problem representation is adopted (Payne et al. 1999) based on an incomplete mental model of the decision problem (Legrenzi et al. 1993; Legrezni and Girotto 1996).

This bias leads to focus on a small number of alternatives (Eisenhardt 1989; Nutt 1998), a small number of objectives (Bond et al. 2008, 2010), or a single future state of the world (Russo and Schoemaker 1989). See also Payne et al. (1999) for further evidence.

Debiasing techniques include explicitly encouraging decision makers to think about more objectives, new alternatives, and other possible states of the future. In addition, it helps to involve multiple experts and stakeholders to improve the range of alternatives, objectives, and states of the world.

Omission of Important Variables An important variable is overlooked (Jargowsky 2005), for example, in the definition of objectives (Bond et al. 2008, 2010), identification of decision alternatives (Butler and Scherer 1997; Pitz et al. 1980), and in hypothesis generation (Fischhoff et al. 1978; Thomas et al. 2008).

Debiasing techniques include prompting for alternatives and objectives by providing specific guidance and categories, asking for extreme or unusual scenarios and by using group elicitation techniques. The use of multiple experts and stakeholder also helps.

Overconfidence Laypeople and experts provide estimates for a given parameter that are above the actual performance (overestimation) (Lichtenstein et al. 1982; Lichtenstein and Fischhoff 1977) or when the range of variation they provide is too narrow (over-precision) (Moore and Healy 2008).

This bias has been demonstrated in many quantitative estimates, such as in defense, legal, financial and engineering contexts (Lin and Bier 2008; Moore and Healy 2008). It is also present in judgments about the completeness of a hypothesis set (Fischhoff et al. 1978; Mehle 1982).

Debiasing techniques include providing probability training and demonstrations of overconfidence, by starting elicitations with extreme estimates (low and high), avoiding central tendency anchors, by using counterfactuals to challenge extremes and by using fixed value elicitations instead of fixed probability elicitations.

Splitting Biases The way the objectives are grouped affects the weights on objectives (Borcherding and von Winterfeldt 1988; Pöyhönen et al. 2001; Weber et al. 1988); or when the way events in a fault tree are grouped affects the event probabilities.

This bias occurs in the elicitation of weights in multi-criteria models (Borcherding and von Winterfeldt 1988; Fischer 1995; Pöyhönen et al. 2001; von Nitzsch and Weber 1993; Weber et al. 1993) and in the elicitation of probabilities in event and fault trees (Fischhoff et al. 1978; Ofir 2000).

Debiasing techniques include splitting objectives or events that receive high weights or probabilities and not splitting objectives or events that receive lower

weights or probabilities, using hierarchical estimation of weights or probabilities, and using ratio judgments instead of direct estimation or distribution of points.

Proxy Bias Proxy attributes receive larger weights than the respective fundamental objectives (Fischer et al. 1987).

This bias has been reported in the elicitation of weights in multiattribute utility and value measurement (Fischer et al. 1987).

Debiasing techniques include avoiding proxy attributes or building models relating proxies and fundamental objectives and providing weights for fundamental objectives.

Range Insensitivity Bias The weights of objectives are not properly adjusted to changes in the range of attributes (Gabrielli and von Winterfeldt 1978; von Nitzsch and Weber 1993).

This bias occurs in the elicitation of weights in multiattribute utility and value measurement (Gabrielli and von Winterfeldt 1978; von Nitzsch and Weber 1993).

Debiasing techniques include making attribute ranges explicit and using swing weighting procedures and by using trade-off or pricing out procedures.

Scaling Biases A family of stimulus-response biases (Poulton 1982, 1989) which comprises: *contraction bias* (underestimating large sizes/differences and overestimating small/size differences); *logarithmic response bias* (using step-changes in the number of digits used in the response, which fit a log scale); *range equalizing bias* (using most of the range of response whatever is the size of the range of the stimuli); *centering bias* (producing a symmetric distribution of responses centered on the midpoint of the range of stimuli); and *equal frequency bias* (using equally all parts of the response scale).

These biases have been reported in the assessment of judgments related to physical and social measurements of various kinds (Poulton 1982, 1989).

Debiasing techniques include developing scales that match stimuli and responses and by choosing appropriate scaling techniques for the measurement required.

15.2.2 Relevant Individual Motivational Biases

We define motivational biases as those in which judgments are influenced by the desirability or undesirability of events, consequences, outcomes, or choices (see also Kunda 1990, von Winterfeldt 1999 and Molden and Higgins 2012). An example of a motivational bias is the deliberate attempt of experts to provide optimistic forecasts for a preferred outcome. Another example is the underestimation of the costs of a project to provide more competitive bids.

Motivational biases do not always have to be conscious. For example, estimates of the time it takes to complete a software project are often overly optimistic

(Connolly and Dean 1997) even when there is no outside pressure or value in misrepresenting the actual time. We focus here on outcome-motivated biases, as they matter in several modeling steps, but recognize that lack of motivation to provide accurate judgments is also an issue in the elicitation of judgments (Molden and Higgins 2012). Contrary to cognitive biases, all motivational biases are hard to correct, thus relevant to decision and risk analysis.

We list each motivational bias below and suggest debiasing techniques against the bias, when eliciting value and uncertainty judgments (for details see Montibeller and von Winterfeldt 2015):

Affect-Influenced Bias An emotional predisposition for, or against, a specific outcome or option taints judgments (Finucane et al. 2000; Slovic et al. 2004).

Several studies have reported this bias, assessing the role of affect causing an inverse perceived relationship between positive and negative consequences related to climate change, pandemics, consumer products, technologies, and human-caused hazards (Siegrist and Sütterlin 2014). There is also evidence that affect influences the estimation of probabilities of events (Rottenstreich and Hsee 2001).

Debiasing techniques include avoiding loaded descriptions of consequences in the attributes, cross-checking judgments with alternative elicitation protocols when eliciting value functions, weights and probabilities, and by using multiple experts with alternative points of view.

Confirmation Bias The desire to confirm one's belief, leading to unconscious selectivity in the acquisition and use of evidence (Nickerson 1998).

This bias has been reported in several experimental settings, such as in information gathering, selection tasks, evidence updating, and own-judgment evaluation (Klayman 1995; Nickerson 1998). Also in real-world contexts, such as medical diagnostics, judicial reasoning, and scientific thinking (Nickerson 1998).

Debiasing techniques include using multiple experts with different points of view about hypotheses, challenging probability assessments with counterfactuals, and by probing for evidence for alternative hypotheses.

Desirability of a Positive Event or Consequence The desirability of an outcome leads to an increase in the extent to which it is expected to occur (Krizan and Windschitl 2007, p. 96). It is also called "wishful thinking" (Seybert and Bloomfield 2009) or "optimism bias" (Weinstein 1980).

This bias occurs in the prediction of outcomes in games of chance (Krizan and Windschitl 2007); impact on estimates of probabilities of future outcomes in expert foresight (Ecken et al. 2011; Tichy 2004), estimates of costs (Dillon et al. 2002) and duration (Connolly and Dean 1997) in projects, as well as some possible effect in sport tournaments (Bar-Hillel et al. 2008).

Debiasing techniques include using multiple experts with alternative points of view, using scoring rule and place hypothetical bets against the desired event

or consequence, and by using decomposition and realistic assessment of partial probabilities to estimate the event probability.

Undesirability of a Negative Event or Consequence The desire to be cautious, prudent, or conservative in estimates that may be related to harmful consequences (Chapin 2001; Dolinski et al. 1987).

Most evidence about this bias is related to probabilities of life events (Chapin 2001; Dolinski et al. 1987); but also in long-term estimated of future events in expert foresight (Tichy 2004) and estimates of risks and benefits about risky technologies (Marks and von Winterfeldt 1984); some risk assessments that are intentionally biased towards "conservative" estimates in each step (as discussed in the recent report by the Institute of Medicine 2013) involve this bias.

Debiasing techniques include using multiple experts with alternatives points of view, using scoring rules and place hypothetical bets in favor of the undesired event or consequence, and by using decomposition and realistic assessment of partial probabilities to estimate the event probability.

15.3 Relevant Group Biases

Most organizational decisions are made in groups, in which members have to express their preferences for different outcomes and indicate their value trade-offs (Keeney 2002). In addition, behavioral aggregation of expert judgments is often employed when eliciting parameters about uncertainties.

There are many benefits of engaging with groups instead of individuals (Kerr and Tindale 2004, 2011). Expert groups providing judgements concerning future events may benefit from an increase of accuracy (from the pooling of information and perspectives, from error checking as well as from motivation gains). There are also benefits associated with social goals, such as procedural fairness and satisfaction/enjoyment. In tasks involving group preferences about different decision alternatives or outcomes there is no accuracy goal, as there are no true values in such cases. Yet, these groups also benefit from the pooling of information and perspectives, from error checking as well as from motivation gains. In addition, to procedural fairness and satisfaction/enjoyment, such groups can benefit from a sense of common purpose and agreement on the way forward (Phillips 2007).

However, in both types of tasks, group biases might affect the quality of the preference statements and judgments. Behavioral decision research has shown that groups may increase or attenuate individual biases, depending on the type of group decision/judgment process, the type and strength of the bias and the individual preferences among members of the group (Kerr and Tindale 2004).

Facilitated modelling (Franco and Montibeller 2010), in which a decision analyst works with the group eliciting their preferences, such as the ones employed in

decision conferences (Phillips 2007), may alleviate group biases and increase the effectiveness of group decision processes. In the same way, well designed elicitation protocols for teams of experts, such as the Delphi technique, can increase the quality of their judgments while reducing group biases.

We list each relevant group bias below and suggest debiasing techniques, when eliciting preferences about decision outcomes and judgments about uncertainties. We base these lists of biases on the comprehensive reviews on group biases for forecasting elicitation (Kerr and Tindale 2011) and, in particular, on team-based decision making (Jones and Roelofsma 2000). Because the debiasing techniques are similar for all group biases, we discuss them after the presentation of each bias.

False Consensus The individual group participant overestimates the similarities between his/her judgements and the others, while viewing alternative perspectives as uncommon or deviant (Ross et al. 1977). This may lead to judgments in which individual members base their decision on incorrect assumptions about other members of the team, anchoring their judgments about others on their own perspective, even if they are aware about their information deficiencies (Jones and Roelofsma 2000). The evidence about this bias comes from studies in social psychology and we are not aware about evidence directly related to decision making or group expert judgment.

Groupthink Members in very cohesive groups that are focus on getting consensus, no matter how it was formed, to the detriment of realistically appraising other courses of action (Janis 1983). In groups affected by this bias, there is a strong pressure to conform and dysfunctional shared representations (Kerr and Tindale 2011). It affects several decision making tasks, such as an incomplete search for alternatives, the consideration of too few objectives, and limited information search (Jones and Roelofsma 2000).

Group Polarization The group discussions enhance the position/opinion that was initially held by the majority of its members (Lamm 1988). When a group is affected by this bias, if group members are initially in favor of a given, further group discussions will increase such favorability for most individuals. This also affects the group's risk attitude, which may become more risk averse than the original risk aversion of individual members or, conversely, may become more risk seeking than the original risk seeking attitude of each member (see Isenberg 1986 for details).

Group Escalation of Commitment Groups continue to support a course of action that is clearly failing, presenting negative outcomes. This is related to the sunk-cost bias (Arkes and Blumer 1985) and also influenced by the loss-gains bias (Tversky and Kahneman 1981) at individual level but exacerbated in groups, by groupthink and group polarization (Jones and Roelofsma 2000). Group think prevents dissenters of confronting the majority in challenging the sunk-cost bias. Group polarization makes groups take riskier choices, already influenced by the risk prone attitude created by a large loss, as predicted by Prospect Theory (Tversky and Kahneman 1992).

Group Overconfidence Groups affected are more confident in the accuracy of their judgments than the individual overconfident members (Plous 1995), particularly when the decision task is complex (Sniezek 1990). There are several possible causes for this bias, such as the use of a limited amount of shared information, the trust placed by the group in the accuracy of its judgements, the convergence in preferences generated by the group, or the social validation promoted by reaching consensus (Kerr and Tindale 2011). This bias affects all judgment task affected by individual overconfidence, as described previously.

The debiasing techniques against group biases are similar for each bias. They encompass using multiple experts with alternatives points of view from different organizations, encouraging different perspectives, using structured elicitation procedures and facilitated decision processes.

While face-to-face meetings have many content and social benefits, as described previously, they are also more prone of group biases. Thus the crucial importance of a well-trained facilitator and carefully designed elicitation protocols, which maximizes their benefits while reducing the occurrence of group biases.

Table 15.1 summarizes the cognitive, motivational, and group biases affecting value and uncertainty judgments for each type of analyses: risk analysis for uncertainty modeling, multi-criteria analysis for decisions with conflicting objectives, and decision tree analysis for decision making under uncertainty. We also identify which

Table 15.1 Bias affecting value and uncertainty judgements in decision and risk analysis

Type of analysis	Biases
Risk analysis	
Modeling tasks: UM1: Definition of target variable and events; UM2: Assessment of probabilities; UM3: Aggregation of probabilities.	• Affect influenced bias (M) [UM2] • Anchoring bias (C) [UM2, UM3] • Availability bias (C) [UM1, UM2] • Confirmation bias (M) [UM1] • Desirability biases (M) [UM2, UM3] • Equalizing bias (C) [UM2] • Myopic problem representation • Omission bias (C) [UM1] • Overconfidence bias (C) [UM1, UM2, UM3] • Scaling biases (C) [UM2] bias (C) [UM1] • False consensus (G) [UM1, UM2, UM3] • Groupthink (G) [UM1, UM2, UM3] • Group polarization (G) [UM1, UM2, UM3] • Group escalation of commitment (G) [UM1, UM2, UM3] • Group overconfidence (G) [UM1, UM2, UM3]

(continued)

Table 1 (continued)

Type of analysis	Biases
Multi-criteria analysis	
Modeling tasks: VM1: Definition of objectives; VM2: Definition of attributes; VM3: Elicitation of value or utility functions; VM4: Elicitation of attribute weights.	• Affect influenced bias (M) [VM3, VM4] • Anchoring bias (C) [VM3] • Availability bias (C) [VM1] • Certainty effect bias (C) [VM3] • Desirability of options bias (M) [VM3, VM4] • Equalizing bias (C) [VM4] • Gain-loss bias (C) [VM2, VM3, VM 4] • Myopic problem representation bias [VM1] • Omission bias (C) [VM1] • Proxy bias (C) [VM2, VM4] • Range insensitivity bias (C) [VM4] • Scaling biases (C) [VM2, VM4] • Splitting bias (C) [VM4] • False consensus (G) [VM1, VM2, VM3, VM4] • Groupthink (G) [VM1, VM2, VM3, VM4] • Group polarization (G) [VM1, VM2, VM3, VM4] • Group escalation of commitment (G) [VM1, VM2, VM3, VM4] • Group overconfidence (G) [VM1, VM2, VM3, VM4]
Decision tree analysis	
Modeling tasks: CM1: Identification of alternatives; CM2: Identification of events and outcomes; CM3: Assessment of probabilities; CM4: Estimation of consequences	• Affect influenced bias (M) [CM1, CM3, CM4] • Anchoring bias (C) [CM1, CM3, CM4] • Availability bias (C) [CM1, CM2, CM3] • Confirmation bias (M) [CM2, CM3] • Desirability biases (M) [CM3, CM4] • Desirability of options bias (M) [CM1] • Equalizing bias (C) [CM3] • Gain–loss bias (C) [CM3] • Myopic problem representation bias (C) [CM1, CM2] • Omission bias (C) [CM1, CM2] • Overconfidence bias (C) [CM2, CM3, CM4] • Scaling biases (C) [CM4] • Splitting bias (C) [CM3, CM4] • False consensus (G) [CM1, CM2, CM3, CM4]

(continued)

Table 1 (continued)

Type of analysis	Biases
Decision tree analysis	
	• Groupthink (G) [CM1, CM2, CM3, CM4] • Group polarization (G) [CM1, CM2, CM3, CM4] • Group escalation of commitment (G) [CM1, CM2, CM3, CM4] • Group overconfidence (G) [CM1, CM2, CM3, CM4]

Key: *C* cognitive bias, *G* group bias, *M* motivational bias

specific modeling task is affected by every bias in brackets, see Montibeller and von Winterfeldt (2015) for details.

15.4 Conclusions

The elicitation of values and uncertainties about decision outcomes is a key feature, and a major strength, of decision and risk analysis. It supports decision makers in thinking clearly about the tough decisions that they have to make. It enables analysts to employ expert judgments in problems where levels of complexity and uncertainty are too high just to extrapolate past trends. However there are numerous biases in value and uncertainty judgments that can affect the quality of decision or risk analyses. The very large number of biases[1] poses a challenge for decision analysts who want to make sure that they are eliciting unbiased judgments from decision makers and experts. The question they face is: Which of these biases should I worry about and how can I correct them, if they occur?

This chapter provides a road map for decision analysts to navigate into the extensive and rather fragmented literature on biases in judgments and decision making. We focused on biases that are relevant for decision and risk analysis. We also classified such biases by their underlying cause: as cognitive or as motivational. In addition we extended this coverage, which was originally focused on individual biases, to include group-level biases.

Some cognitive biases are not relevant to decision analysis, because they are easy to correct; however all motivational biases are relevant. When elicitation processes involve groups, all group biases are relevant. Regarding the latter, the trend of using facilitated decision modeling (Franco and Montibeller 2010) to support group decision making in complex societal decisions (e.g. Morton et al. 2009; Del Rio Vilas et al. 2013; Ferretti and Montibeller 2016) means that more attention must be devoted in understanding group-level biases and how to minimize them in decision conferencing workshops.

[1]For a compiled list see: https://en.wikipedia.org/wiki/List_of_cognitive_biases.

There is a very limited literature on debiasing and few attempts of assessing their effectiveness in improving judgments and the statement of preferences (Montibeller and von Winterfeldt 2015). Our recent research efforts are in this direction, as exemplified by our study on how to debias over-precision in the elicitation of cumulative distribution functions techniques (Ferretti et al. 2016). We hope that future research will be able to match debiasing techniques with the underlying cause of a bias and thoroughly evaluate their effectiveness in mitigating the bias. It is an exciting and important research endeavor for decision and risk analysis and anyone interested in improving the quality of judgment and preference elicitation processes.

References

Allais M (1953) Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'Ecole Americaine. Econometrica 21:503–546. doi:10.2307/1907921

Arkes HR (1991) Costs and benefits of judgment errors: implications for debiasing. Psychol Bull 110:486–498. doi:10.1037/0033-2909.110.3.486

Arkes HR, Blumer C (1985) The psychology of sunk cost. Organ Behav Hum Decis Process 35:124–140. doi:10.1016/0749-5978(85)90049-4

Bar-Hillel M (1980) The base-rate fallacy in probability judgments. Acta Psychol 44:211–233. doi:10.1016/0001-6918(80)90046-3

Bar-Hillel M, Budescu DV, Amar M (2008) Predicting World Cup results: do goals seem more likely when they pay off? Psychon Bull Rev 15:278–283. doi:10.3758/PBR.15.2.278

Bazerman MH, Moore DA (2013) Judgment in managerial decision making, 8th edn. Wiley, Hoboken, NJ

Bond SD, Carlson KA, Keeney RL (2010) Improving the generation of decision objectives. Decis Anal 7:238–255

Bond SD, Carlson KA, Keeney RL (2008) Generating objectives: can decision makers articulate what they want? Manag Sci 54:56–70. doi:10.1287/mnsc.1070.0754

Borcherding K, von Winterfeldt D (1988) The effect of varying value trees on multiattribute evaluations. Acta Psychol 68:153–170. doi:10.1016/0001-6918(88)90052-2

Butler AB, Scherer LL (1997) The effects of elicitation aids, knowledge, and problem content on option quantity and quality. Organ Behav Hum Decis Process 72:184–202. doi:10.1006/obhd.1997.2737

Chapin J (2001) Self-protective pessimism: optimistic bias in reverse. N Am J Psychol 3:253–262

Connolly T, Dean D (1997) Decomposed versus holistic estimates of effort required for software writing tasks. Manag Sci 43:1029–1045

Del Rio Vilas VJ, Voller F, Montibeller G, Franco LA, Sribhashyam S, Watson E, Hartley M, Gibbens JC (2013) An integrated process and management tools for ranking multiple emerging threats to animal health. Prev Vet Med 108:94–102. doi:10.1016/j.prevetmed.2012.08.007

Dillon RL, John R, von Winterfeldt D (2002) Assessment of cost uncertainties for large technology projects: a methodology and an application. Interfaces 32:52–66. doi:10.1287/inte.32.4.52.56

Dolinski D, Gromski W, Zawisza E (1987) Unrealistic pessimism. J Soc Psychol 127:511–516. doi:10.1080/00224545.1987.9713735

Ecken P, Gnatzy T, von der Gracht HA (2011) Desirability bias in foresight: Consequences for decision quality based on Delphi results. Technol Forecast Soc Chang 78:1654–1670. doi:10.1016/j.techfore.2011.05.006

Eisenhardt KM (1989) Making fast strategic decisions in high-velocity environments. Acad Manag J 32:543–576. doi:10.2307/256434

Ferretti V, Guney S, Montibeller G, von Winterfeldt D (2016) Testing best practices to reduce the overconfidence bias in multi-criteria decision analysis. IEEE:1547–1555. doi:10.1109/HICSS.2016.195

Ferretti V, Montibeller G (2016) Key challenges and meta-choices in designing and applying multi-criteria spatial decision support systems. Decis Support Syst 84:41–52. doi:10.1016/j.dss.2016.01.005

Finucane ML, Alhakami A, Slovic P, Johnson SM (2000) The affect heuristic in judgments of risks and benefits. J Behav Decis Mak 13:1–17. doi:10.1002/(SICI)1099-0771(200001/03)13:1<1::AID-BDM333>3.0.CO;2-S

Fischer GW (1995) Range sensitivity of attribute weights in multiattribute value models. Organ Behav Hum Decis Process 62:252–266. doi:10.1006/obhd.1995.1048

Fischer GW, Damodaran N, Laskey KB, Lincoln D (1987) Preferences for proxy attributes. Manag Sci 33:198–214

Fischhoff B, Slovic P, Lichtenstein S (1978) Fault trees: sensitivity of estimated failure probabilities to problem representation. J Exp Psychol Hum Percept Perform 4:330–344. doi:10.1037/0096-1523.4.2.330

Fox CR, Bardolet D, Lieb D (2005) Partition dependence in decision analysis, resource allocation, and consumer choice. In: Zwick R, Rapoport A (eds) Experimental business research, Marketing, accounting, and cognitive perspectives, vol III. Springer, Dordrecht, pp 229–251

Fox CR, Clemen RT (2005) Subjective probability assessment in decision analysis: partition dependence and bias toward the ignorance prior. Manag Sci 51:1417–1432

Franco LA, Montibeller G (2010) Facilitated modelling in operational research. Eur J Oper Res 205:489–500. doi:10.1016/j.ejor.2009.09.030

Frisch D (1993) Reasons for framing effects. Organ Behav Hum Decis Process 54:399–429. doi:10.1006/obhd.1993.1017

Furnham A, Boo HC (2011) A literature review of the anchoring effect. J Socio-Econ 40:35–42. doi:10.1016/j.socec.2010.10.008

Gabrielli WF, von Winterfeldt D (1978) Are importance weights sensitive to the range of alternatives in multiattribute utility measurement? (No. 78–6), Research Report. Los Angeles, CA, Research Report 78-6, Social Science Research Institute, University of Southern California

Gilovich T, Griffin DW, Kahneman D (2002) Heuristics and biases: the psychology of intuitive judgement. Cambridge University Press, Cambridge

Hershey JC, Schoemaker PJH (1985) Probability versus certainty equivalence methods in utility measurement: are they equivalent? Manag Sci 31:1213–1231

Institute of Medicine (ed) (2013) Environmental decisions in the face of uncertainty. The National Academies, Washington, DC

Isenberg DJ (1986) Group polarization: a critical review and meta-analysis. J Pers Soc Psychol 50:1141–1151. doi:10.1037/0022-3514.50.6.1141

Jacobi SK, Hobbs BF (2007) Quantifying and mitigating the splitting bias and other value tree-induced weighting biases. Decis Anal 4:194–210. doi:10.1287/deca.1070.0100

Janis IL (1983) Groupthink: psychological studies of policy decisions and fiascoes. Houghton Mifflin, Boston, MA

Jargowsky PA (2005) Omitted variable bias. In: Encyclopedia of social measurement, vol 2. Elsevier, New York, pp 919–924

Jones PE, Roelofsma PHMP (2000) The potential for social contextual and group biases in team decision-making: biases, conditions and psychological mechanisms. Ergonomics 43:1129–1152. doi:10.1080/00140130050084914

Kahneman D, Slovic P, Tversky A (1982) Judgment under uncertainty: heuristics and biases. Cambridge University Press, Cambridge

Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. Econometrica 47:263–291. doi:10.2307/1914185

Kahneman D, Tversky A (1973) On the psychology of prediction. Psychol Rev 80:237–251. doi:10.1037/h0034747

Keeney RL (2002) Common mistakes in making value trade-offs. Oper Res 50:935–945

Kerr NL, Tindale RS (2011) Group-based forecasting?: A social psychological analysis. Int J Forecast 27:14–40. doi:10.1016/j.ijforecast.2010.02.001
Kerr NL, Tindale RS (2004) Group performance and decision making. Annu Rev Psychol 55:623–655. doi:10.1146/annurev.psych.55.090902.142009
Klayman J (1995) Varieties of confirmation bias. In: Busemeyer J, Hastie R, Medin D (eds) Psychology of learning and motivation, Decision making from a cognitive perspective, vol 32. Academic, New York, NY, pp 365–418
Krizan Z, Windschitl PD (2007) The influence of outcome desirability on optimism. Psychol Bull 133:95–121. doi:10.1037/0033-2909.133.1.95
Kühberger A (1998) The influence of framing on risky decisions: a meta-analysis. Organ Behav Hum Decis Process 75:23–55. doi:10.1006/obhd.1998.2781
Kunda Z (1990) The case for motivated reasoning. Psychol Bull 108:480–498. doi:10.1037/0033-2909.108.3.480
Lamm H (1988) A review of our research on group polarization: eleven experiments on the effects of group discussion on risk acceptance, probability estimation, and negotiation positions. Psychol Rep 62:807–813. doi:10.2466/pr0.1988.62.3.807
Larrick RP (2007) Debiasing. In: Koehler DJ, Harvey N (eds) Blackwell handbook of judgment and decision making. Blackwell, Malden, MA, pp 316–338
Legrenzi P, Girotto V, Johnson-Laird PN (1993) Focussing in reasoning and decision making. Cognition 49:37–66. doi:10.1016/0010-0277(93)90035-T
Legrezni P, Girotto V (1996) Mental models in reasoning and decision making. In: Oakhill J, Garnham A (eds) Mental models in cognitive science: essays in honour of Phil Johnson-Laird. Psychology, Hove, pp 95–118
Levin IP, Schneider SL, Gaeth GJ (1998) All frames are not created equal: a typology and critical analysis of framing effects. Organ Behav Hum Decis Process 76:149–188
Lichtenstein S, Fischhoff B (1977) Do those who know more also know more about how much they know? Organ Behav Hum Perform 20:159–183. doi:10.1016/0030-5073(77)90001-0
Lichtenstein S, Fischhoff B, Phillips LD (1982) In: Kahneman D, Slovic P, Tversky A (eds) Calibration of probabilities: the state of the art to 1980. Cambridge University Press, Cambridge, pp 306–334
Lichtenstein S, Slovic P, Fischhoff B, Layman M, Combs B (1978) Judged frequency of lethal events. J Exp Psychol Hum Learn Mem 4:551–578. doi:10.1037/0278-7393.4.6.551
Lin S-W, Bier VM (2008) A study of expert overconfidence. Reliab Eng Syst Saf 93:711–721. doi:10.1016/j.ress.2007.03.014
Marks G, von Winterfeldt D (1984) "Not in my back yard": influence of motivational concerns on judgments about a risky technology. J Appl Psychol 69:408–415. doi:10.1037/0021-9010.69.3.408
Mehle T (1982) Hypothesis generation in an automobile malfunction inference task. Acta Psychol 52:87–106. doi:10.1016/0001-6918(82)90028-2
Milkman KL, Chugh D, Bazerman MH (2009) How can decision making be improved? Perspect Psychol Sci 4:379–383. doi:10.1111/j.1745-6924.2009.01142.x
Molden DC, Higgins ET (2012) Motivated thinking. In: Holyoak KJ, Morrison RG (eds) The Oxford handbook of thinking and reasoning. Oxford University Press, New York, pp 390–409
Montibeller G, von Winterfeldt D (2015) Cognitive and motivational biases in decision and risk analysis. Risk Anal 35:1230–1251. doi:10.1111/risa.12360
Moore DA, Healy PJ (2008) The trouble with overconfidence. Psychol Rev 115:502–517. doi:10.1037/0033-295X.115.2.502
Morton A, Airoldi M, Phillips LD (2009) Nuclear risk management on stage: a decision analysis perspective on the UK's Committee on Radioactive Waste Management. Risk Anal 29:764–779. doi:10.1111/j.1539-6924.2008.01192.x
Mussweiler T, Strack F (2001) The semantics of anchoring. Organ Behav Hum Decis Process 86:234–255. doi:10.1006/obhd.2001.2954
Nickerson RS (1998) Confirmation bias: a ubiquitous phenomenon in many guises. Rev Gen Psychol 2:175–220. doi:10.1037/1089-2680.2.2.175

Nutt PC (1998) How decision makers evaluate alternatives and the influence of complexity. Manag Sci 44:1148–1166
Ofir C (2000) Ease of recall vs recalled evidence in judgment: experts vs laymen. Organ Behav Hum Decis Process 81:28–42. doi:10.1006/obhd.1999.2864
Payne JW, Bettman JR, Schkade DA, Schwarz N, Gregory R (1999) Measuring constructed preferences: towards a building code. J Risk Uncertain 19:243–270
Phillips LD (2007) Decision conferencing. In: Edwards W, Miles RF, von Winterfeldt D (eds) Advances in decision analysis: from foundations to applications. Cambridge University Press, New York, NY, pp 375–399
Pitz GF, Sachs NJ, Heerboth J (1980) Procedures for eliciting choices in the analysis of individual decisions. Org Behav Hum Perform 26:396–408. doi:10.1016/0030-5073(80)90075-6
Plous S (1995) A comparison of strategies for reducing interval overconfidence in group judgments. J Appl Psychol 80:443–454. doi:10.1037/0021-9010.80.4.443
Poulton EC (1989) Bias in quantifying judgements. Erlbaum, Hove
Poulton EC (1982) Biases in quantitative judgements. Appl Ergon 13:31–42. doi:10.1016/0003-6870(82)90129-6
Pöyhönen M, Vrolijk H, Hämäläinen RP (2001) Behavioral and procedural consequences of structural variation in value trees. Eur J Oper Res 134:216–227. doi:10.1016/S0377-2217(00)00255-1
Ross L, Greene D, House P (1977) The "false consensus effect": an egocentric bias in social perception and attribution processes. J Exp Soc Psychol 13:279–301. doi:10.1016/0022-1031(77)90049-X
Rottenstreich Y, Hsee CK (2001) Money, kisses, and electric shocks: on the affective psychology of risk. Psychol Sci 12:185–190. doi:10.1111/1467-9280.00334
Russo JE, Schoemaker PJH (1989) Decision traps: ten barriers to brilliant decision-making and now to overcome them. Doubleday, New York, NY
Schoemaker PJH, Hershey JC (1992) Utility measurement: signal, noise, and bias. Organ Behav Hum Decis Process 52:397–424. doi:10.1016/0749-5978(92)90027-5
Seybert N, Bloomfield R (2009) Contagion of wishful thinking in markets. Manag Sci 55:738–751. doi:10.1287/mnsc.1080.0973
Siegrist M, Sütterlin B (2014) Human and nature-caused hazards: the affect heuristic causes biased decisions. Risk Anal 34:1482–1494. doi:10.1111/risa.12179
Slovic P, Finucane ML, Peters E, MacGregor DG (2004) Risk as analysis and risk as feelings: some thoughts about affect, reason, risk, and rationality. Risk Anal 24:311–322. doi:10.1111/j.0272-4332.2004.00433.x
Sniezek JA (1990) A comparison of techniques for judgmental forecasting by groups with common information. Group Organ Stud 15:5–19. doi:10.1177/105960119001500102
Thomas RP, Dougherty MRP, Sprenger AM, Harbison JI (2008) Diagnostic hypothesis generation and human judgment. Psychol Rev 115:155–185. doi:10.1037/0033-295X.115.1.155
Tichy G (2004) The over-optimism among experts in assessment and foresight. Technol Forecast Soc Chang 71:341–363. doi:10.1016/j.techfore.2004.01.003
Tversky A, Kahneman D (1992) Advances in prospect theory: cumulative representation of uncertainty. J Risk Uncertain 5:297–323. doi:10.1007/BF00122574
Tversky A, Kahneman D (1983) Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. Psychol Rev 90:293–315. doi:10.1037/0033-295X.90.4.293
Tversky A, Kahneman D (1981) The framing of decisions and the psychology of choice. Science 211:453–458. doi:10.1126/science.7455683
Tversky A, Kahneman D (1974) Judgment under uncertainty: heuristics and biases. Science 185:1124–1131. doi:10.1126/science.185.4157.1124
Tversky A, Kahneman D (1973) Availability: a heuristic for judging frequency and probability. Cogn Psychol 5:207–232. doi:10.1016/0010-0285(73)90033-9
von Nitzsch R, Weber M (1993) The effect of attribute ranges on weights in multiattribute utility measurements. Manag Sci 39:937–943

von Winterfeldt D (1999) On the relevance of behavioral decision research for decision analysis. In: Shanteau J, Mellers BA, Schum DA (eds) Decision science and technology: reflections on the contributions of ward edwards. Kluwer, Norwell, pp 133–154

von Winterfeldt D, Edwards W (1986) Decision analysis and behavioral research. Cambridge University Press, New York, NY

Wänke M, Schwarz N, Bless H (1995) The availability heuristic revisited: experienced ease of retrieval in mundane frequency estimates. Acta Psychol 89:83–90. doi:10.1016/0001-6918(93)E0072-A

Weber EU, Böckenholt U, Hilton DJ, Wallace B (1993) Determinants of diagnostic hypothesis generation: effects of information, base rates, and experience. J Exp Psychol Learn Mem Cogn 19:1151–1164. doi:10.1037/0278-7393.19.5.1151

Weber M, Eisenfuhr F, von Winterfeldt D (1988) The effects of splitting attributes on weights in multiattribute utility measurement. Manag Sci 34:431–445

Weinstein ND (1980) Unrealistic optimism about future life events. J Pers Soc Psychol 39:806–820. doi:10.1037/0022-3514.39.5.806

Chapter 16
The Selection of Experts for (Probabilistic) Expert Knowledge Elicitation

Fergus Bolger

Abstract Several different EKE protocols are reviewed in this volume, each with their pros and cons, but any is only as good as the quality of the experts and their judgments. In this chapter a structured approach to the selection of experts for EKE is presented that is grounded in psychological research.

In Part I various definitions of expertise are considered, and indicators and measures that can be used for the selection of experts are identified. Next, some ways of making judgements of uncertain quantities are discussed, as are factors influencing judgment quality.

In Part II expert selection is considered within an overall policy-making process. Following the analysis of Part I, two new instruments are presented that can help guide the selection process: expert profiles provide structure to the initial search, while a questionnaire permits matching of experts to the profiles, and assessment of training needs. Issues of expert retention and documentation are also discussed.

It is concluded that although the analysis offered in this chapter constitutes a starting point there are many questions still to be answered to maximize EKE's contribution. A promising direction is research that focusses on the interaction between experts and the tasks they perform.

16.1 Introduction

Sound decision and policy making depend upon quantitative analysis but hard data for analysis is not always available or of good quality. There is consequently a need to supplement (or substitute for) empirical evidence with expert judgement. Unfortunately expert judgement is also often poor due to the operation of psychological and social factors that lead to error and bias: this is particularly true when it comes to the assessment of uncertainty surrounding judgements. For this reason structured methods have been developed—referred to as expert knowledge elicitation (EKE) techniques—to help improve the quality of expert judgement: these methods are

F. Bolger (✉)
Strathclyde Business School, Strategy & Organisation, 199 Cathedral Street, G4 0QU, UK
e-mail: fergus.bolger@strath.ac.uk

L.C. Dias et al. (eds.), *Elicitation*, International Series in Operations Research & Management Science 261, https://doi.org/10.1007/978-3-319-65052-4_16

based on both psychological research, and statistical and logical principles. EKE techniques are becoming increasingly used in some areas (e.g. risk analysis) and are starting to be applied in forecasting and foresight (Bolger and Wright 2017).

In this chapter I consider the application of EKE to judgement of quantities (with particular reference to the assessment of the uncertainty surrounding these quantities). Although the debiasing methods incorporated in EKE are very important for promoting judgement quality, perhaps even more important is the selection of sufficient and appropriate experts in the first instance. For this reason, rather than concentrating on the EKE techniques themselves—which are described elsewhere in this volume—I wish to focus on the identification and selection of experts for such (probabilistic) EKE's.

If experts are considered sources of data then their selection, and subsequent use can all be discussed in terms of their effectiveness for the maximization of the reliability and validity of the experts' judgmental inputs: as such I will discuss the measurement of expertise for screening, weighting and establishment of training needs, as well as initial identification, selection and recruitment. Management of experts—particularly, motivation and retention, and provision of feedback—are also relevant to maintaining the quality of expert judgment, however, these will be only touched upon because they depend more on the characteristics of specific protocols, which are dealt with in other chapters in this book.

In the first part of this paper I will discuss some general issues regarding defining, identifying, and measuring expertise. In the second part I will work through in detail—using a case study—a two stage expert recruitment strategy whereby, in the first stage a long list of potential experts is created with the help of an instrument called the 'expert profile matrix', and in the second stage this long list is reduced to a short list using a second instrument called the 'Expert-Skills Questionnaire' (E-SQ).

16.2 Part I: Defining, Identifying and Measuring Expertise

16.2.1 Defining Expertise

In order to select experts for EKE you must first have some idea of who might be considered expert in the domain of interest, in other words, who is likely to make the best estimates of the target quantities. To this end it is worth considering how expertise is commonly defined.

16.2.1.1 Expertise as Superior Knowledge and/or Ability

Probably the feature that is most associated with expertise is superior knowledge; thus, an expert is: "... anyone especially knowledgeable in the field ..."

(Meyer and Booker 1991, p. 85). However, there may be more to expertise than simply a large body of domain knowledge as experience is not just about learning facts and rules, but about recognising how to apply this knowledge appropriately (and also how to acquire more knowledge). Hence the Nobel Prize-winning physicist Niels Bohr described an expert as: "A person that has made every possible mistake within his or her field". Thus, experience of the practical use of knowledge is important because it provides a 'reality check': knowledge can be modified in the light of feedback about when it does and does not apply. This is in contrast to 'textbook learning' or 'armchair philosophising', where knowledge is acquired or elaborated without any verification against what is true in the world or works in practice. A related component of practical expertise is the ability to solve problems by applying knowledge to new situations that have not previously been encountered, and having strategies for acquiring knowledge when it is found to be lacking (e.g. scientific research methods or how to use data resources).

16.2.1.2 Socially Defined Expertise

Although we expect experts to be more knowledgeable in their field than non-experts, expertise is often ascribed on the basis of role (and symbols of that role such as the scientist's or doctor's white coat). Meanwhile those whom we know well, and see as being like us, are less likely to be ascribed expert status than strangers—the comedian Will Rogers summed this up in his comic but astute definition of an expert as: "A man fifty miles from home with a briefcase". This 'social' expertise must be treated with caution because the correlation between social rank and skill or knowledge is often weak due to the many ways of gaining rank other than by knowing a lot (e.g. 'old boy' networks, being a 'squeaky wheel', appearing to work hard, coming from a wealthy family, being in 'the right place at the right time' etc.).

16.2.1.3 Properties of Experts

Two basic views of expertise underlie the definitions above respectively. In the first, expertise is seen as a property of individuals, mainly as a consequence of extensive practice, but also partly as a function of characteristics thought to be innate (e.g. personality and intelligence). In the second, expertise is regarded as an emergent property of 'communities of practice' (Lave and Wenger 1991; Wenger 1998) such that the practices, indicators and standards of expert performance are defined by consensus within a particular group, for example a professional group such as doctors or lawyers.

Although seemingly of less relevance to the goal of selecting experts for elicitation, the view that expertise is socially constructed should not be ignored because it impacts on who is considered to be 'expert' and thus put into the pool

of people to be potentially approached for an elicitation. Professions, trades and other groups formed to provide some specific good or service that is perceived by the general public—or sold to them—as requiring knowledge or skills beyond what an average person could achieve without training, usually have a set of 'good practices' that define their activities. For example, academics will have certain standards regarding teaching (e.g. dealing with student queries, providing feedback on work, and use of audio-visual aids) and research (e.g. citing and referencing, ethical procedures and responding to requests to peer review articles). Some of these practices may be formalised (e.g. in handbooks, guidelines and employment contracts) and others may not. Conformity with these practices is part of what identifies an academic as an academic and distinguishes him or her from other similar individuals (e.g. teachers, industrial scientists).

MacIntyre's notion of a practice as applied to such activities may be of relevance here. Practitioners, in MacIntyre's sense, engage collectively in a "coherent and complex form of socially established cooperative activity" in which they seek to achieve "those standards of excellence which are appropriate to, and partially definitive of, that form of activity" (MacIntyre 2007, p. 187; see also Moore and Beadle 2006).

In many cases there will be some peer or professional accreditation of standards of practice (e.g. society membership, awards, sinecures, etc.) to reinforce 'agreed' good practices, but what might be considered good practice by peers and professional bodies may not necessarily be the criteria applied by the public or even managers (e.g. good pedagogical practice might not be evaluated highly by students). The point relevant to our current concerns is that those who are considered experts by their peers may often be so because of perceived conformity to good practices whereas different criteria (e.g. confidence, fame, how arcane it appears) might be used by outsiders—neither the former nor the latter criteria for social expertise are necessarily well correlated with knowledge or skill-based expertise: I will return to this later.

16.2.1.4 Expertise Continuum

Clearly there is a continuum of expertise from the 'naive' or 'lay' person, who has no specialist knowledge or experience of the task domain, to the novice, who is just starting to acquire skills in the domain, to the intermediate whose knowledge and skills are yet to plateau, to 'grand master' who is unlikely to learn significantly more. Grand master level might not always be the most desirable for elicitation purposes as knowledge and skills often become 'compiled' with experience (i.e. move from deliberate conscious strategies to automatic unconscious ones) and so less accessible to introspection (see e.g. Bargh 1994; Dror 2011). Thus, if the aim is to model decision processes then an intermediate or even a novice might be more useful. However, for the purpose of eliciting quantitative estimates, and associated uncertainty, we would normally wish to recruit experts with as much

relevant experience as possible; there are a couple of exceptions, though. First, if there has been some 'structural change' in the world, then an expert who has many years of experience, but mostly before the change point, may be less useful than an expert who has fewer years of experience in total, but more of these have been acquired after the change point. Second, if there is reason to believe that greater experience leads to entrenched thinking or biases in probability judgement such as overconfidence or risk aversion.

With regard to the former, it may seem at first sight that such structural change would be very rare. However, there are actually many reasons why such change might occur; for instance, there may be new technological or scientific developments, there may be revisions of legal or regulatory frameworks, or experts may simply move from one country to another. The development of entrenched or biased thinking is perhaps a more pervasive problem, though, and difficult to spot. It was observed many years ago that experts are often insensitive to the differential diagnosticity of information, such that giving them more information simply leads to an increase in confidence but no improvement in performance (e.g. Oskamp 1965). Another example is the institutionalisation of risk aversion among social-workers and doctors (Dalgleish 1988)—a notable illustration of this is the Cleveland child abuse case in the UK, where instances of abuse were hugely overdiagnosed, presumably because the costs of missing an instance were greater than the costs of false positives.

16.2.1.5 Granularity and Scope of Expertise

As well as amount of knowledge held by the expert we need also to consider its level of specificity: "... the individual should not be considered expert unless he or she is knowledgeable at the level of detail being elicited..." (Meyer and Booker 1991, p. 85). For example, an expert entomologist might be less able to estimate the risk to a crop of a particular sort of insect than an expert who specialises in that type of insect. However, more specific of fine-grained expertise is not necessarily better than more general, coarser-grained expertise. In the example just considered, an even better choice of expert might be someone who has studied a variety of threats to crops—including the insect in question—particularly if they have local contextual knowledge: specific domain knowledge coupled with a broad perspective will therefore often be the best choice.

16.2.1.6 Types of Expertise

Procedural Versus Declarative

Procedural knowledge is about how to do things (e.g. drive a car) whereas declarative knowledge is about rules and facts (e.g. the Highway Code). The latter

may be, but is not necessarily, easier to express—hence the label 'declarative' (i.e. it can be declared). With a great deal of practice how we do things becomes automatic in many domains (e.g. Anderson 1982), and not available to consciousness; instead we just see the results of expertise. A consequence of this is sometimes that the more expert an expert is, the harder it is for him or her to teach others about it.[1] In the current context—assessing uncertain quantities—it may appear that the manner in which these assessments are arrived at is unimportant, and thus that the problem of automaticity, and consequent lack of access, is not a problem. However, as I will argue later, this is not the case: the manner in which quantitative judgements and, in particular, assessments of uncertainty, are made has important consequences for the quality of these judgements, the method of EKE that is best used, and the training that experts might need.

The distinction between procedural and declarative knowledge might be blurred, though. For instance, probability distributions are unlikely to be stored in experts' heads, and thus may not be the same as facts and rules. Rather, the expert may have to construct the distribution de novo during the elicitation exercise. This being the case, it may be advantageous for the expert to be able to state the reasoning processes whereby probabilities are derived to the elicitor: there may therefore be arguments for using less experienced experts for whom reasoning processes have not become implicit. At very least it may be useful from the perspective of improving assessment of uncertain quantities to examine those with less experience in order to better understand the judgement processes involved. As I will argue later, such understanding might assist in designing more effective EKE procedures.

A concern related to the procedural-declarative distinction is whether knowledge—in the broadest sense (i.e. including strategic knowledge)—is available to consciousness. As I have just implied, procedural knowledge is less likely to be available to consciousness than declarative, but this is not necessarily the case. For instance, facts and rules could be applied to judgement unconsciously (e.g. well-learned mental arithmetic using number facts and mathematical rules) whereas some heuristic procedures could be applied deliberately and consciously (e.g. judging likelihood of class membership of an exemplar by its similarity to the prototypic member of the class). Further, much of what appears to be conscious and deliberate has been shown to be actually unconscious and governed by processes over which we have little control or awareness (see e.g. Nisbett and Wilson 1977).

With regard to the assessment of uncertainty, probabilities could be assessed either consciously or unconsciously, or by a combination of these. For instance, in some models of probability judgement (e.g. Support Theory; Tversky and Koehler 1994), uncertainty is assessed by weighing evidence for and against, say, the

[1] This is a manifestation of the 'paradox of expertise' (e.g. Dror 2011), which is that experts become worse in some respects as they become more expert e.g. less flexible, creative and responsive; more biased etc. This is because knowledge and reasoning become 'fossilized': less amenable to inspection, change and communication.

occurrence of a target event: this evaluation may or may not be conscious. In other models (e.g. Probabilistic Mental Models; Gigerenzer et al. 1991) probabilities are derived from implicit (i.e. unconscious) knowledge of how well cues predict criteria. In yet other models (e.g. the Decision-Variable Partition model; Ferrell and McGoey 1980), there is a two-stage process: first a feeling of uncertainty is arrived at (usually, but not necessarily, unconscious); second that feeling is mapped onto an external scale (usually, but not necessarily, conscious). An important point here is that—until we know more about the processes of uncertainty judgement—we should be circumspect about judging the quality of this judgement on the basis of whether or not it is conscious.

Theoretical Versus Practical

This distinction is often related to the previous distinction, but it is not exactly the same. Theoretical knowledge is about general principles, while practical knowledge is about how to *apply* the principles in specific cases (e.g. statistician vs. actuary): the former often being declarative and the latter often procedural, but not necessarily. In the context of eliciting judgements of quantities and probabilities we are mostly concerned with practical expertise (i.e. the application of expertise to predicting a particular target variable of interest) but possibly interested in theoretical knowledge too (e.g. to formally document the elicitation process, or if reasons for a probability judgement are required as, for instance, in a Delphi procedure).[2]

Substantive Versus Normative

Again this is related, but not identical, to the previous distinctions. Substantive knowledge concerns particular domains and is the type of knowledge most commonly associated with expertise. In contrast, normative expertise refers to formal aspects of a knowledge domain such as commonly agreed units of measurement,

[2]In the Delphi procedure, 'groups' of experts—who never meet or interact directly, and are anonymous to each other (all to reduce sources of social bias)—are polled for their opinions. These opinions are usually point estimates or forecasts of event occurrence (see e.g. Chap. 5 for a discussion of the elicitation and evaluation of such judgments) but can also be judgments of uncertain quantities expressed as probability intervals or distributions (see Bolger et al. 2014): reasons for judgments are also often elicited. Once experts have individually expressed their opinions they are collated by a facilitator and fed back to the expert panel (most normally quantitative estimates are averaged in some manner, and qualitative responses summarized, although individual responses may also be fed back if the group is not too large). The experts are then invited to revise their opinions in the light of the feedback and resubmit them to the facilitator. The process continues through a number of iterations usually until opinion change ceases. Normally the aggregated judgements from the final round are the output although partially aggregated or disaggregated judgements can be submitted if the process fails to lead to consensus.

general rules for evaluating and analyzing data, standards of assessment, procedures of 'best practice' and so on. For instance, an 'expert' in experimental psychology would know: procedures for running experiments so as to maximize internal validity (a.k.a. experimental control), minimize sampling bias (e.g. random vs. convenience sampling), and maintain ethical standards (e.g. principles of informed consent and anonymity); the existence of relevant measurement instruments (e.g. tests of memory, intelligence, mood, reaction times etc.) and their units and their properties (e.g. reliability and validity); how to apply statistical tests to analyze data (ANOVA, multiple regression, non-parametric tests), and the interpretation of the outputs of such tests (e.g. F-ratios, standardized regression coefficients, chi-squared values); and much more. None of this is directly related to their substantive knowledge regarding the theories and empirical findings in their area of research, although it is necessary for the conduct of original research and the interpretation of the research of others.

While substantive expertise is what we are chiefly after, normative expertise will usually assist in the elicitation of reliable and communicable judgements *with uncertainty estimates that are both realistic and in conformance with the laws of probability*. I have emphasized the latter part of the preceding sentence because this is crucial to achieving good outcomes from an EKE, but is often difficult to achieve in practice since available experts may lack normative expertise with regard to the expression of uncertainty as probabilities. The distinction between substantive and normative aspect of expertise will accordingly be reprised as a central component of the following sections of Part I.

16.2.2 Identifying Expertise

In the previous section I identified some properties of experts and dimensions of their expertise. For the practical purpose of selection and recruitment for an EKE it is necessary to translate these properties and dimensions into concrete indicators and measures of expert judgement in the target domain. Alvarado-Valencia et al. (2017) distinguish between a priori indicators of expertise (e.g. things that can be gleaned from a CV) and on-task measures (i.e. performance data). For the purposes of this review, I will make a similar distinction: I will use 'indicators' to refer to things that can potentially be found out about experts prior to directly approaching them (e.g. biographical or bibliographical information, or peer/employer recommendations), that can be used for the identification of experts to be added a long list of potential candidates for EKE, and 'measures' to signify tests of expertise—including self-assessments and reports—and any other performance data obtained once potential experts have been long-listed, that can be used for screening, short-listing and weighting. Of course, some indicators can also be used as measures and thus will be discussed—in somewhat different ways—both here and in the next section on measuring expertise.

16.2.2.1 Substantive Expertise

Potentially good indicators here are those that are based on experts actually knowing more about the domain in question, or having more experience of making judgements in this domain, or both. Examples include: formal qualifications, proof of completion of training courses, years of on-the-job experience, awards and published papers. To my knowledge, with the exception of publications, the reliability and validity of these metrics as indicators of expert performance has not been systematically investigated and remains a topic to be investigated in future research. It is worth noting, though, that these indicators are potential enemies of heterogeneity in groups as they will tend to identify experts with similar characteristics (e.g. white, middle-aged, male academics).

With regard to publications, it is not only their quantity but their quality that should probably be taken into account thus, for example, peer-reviewed papers could be given more weight than those which are not, while within peer-reviewed articles the source journals could be evaluated in terms of their impact ratings and other such metrics, rankings by professional bodies, and/or peer opinion. Attempts to weight experts on the basis of evaluations of outputs alone have not proved to be particularly successful (e.g. Burgman et al. 2011; Cooke et al. 2008) so I suggest that outputs are only used for identifying experts in conjunction with other indicators: this may preclude the use of some automated methods for finding experts (see e.g. Moreira and Wichert 2013), at least as the sole methodology. Further, it may often be the case that experts outside academia are required, as they will have valuable practical knowledge. Such experts will be less likely to publish in peer-reviewed journals, thus it will be necessary to look at trade publications, technical reports, conference proceedings etc. However, due to confidentiality issues much of the output of industry experts may not be in the public domain at all, so it may be necessary to find other types of output that indicate expertise such as oral presentations, media appearances, or patents. Outputs like papers and presentations (including teaching experience) are not only evidence of domain-knowledge but also indicators of the ability to communicate expertise, which can be useful in an EKE (e.g. when giving rationales for judgements).

16.2.2.2 Normative Expertise

Recall that normative expertise refers to formal, abstract methods for expressing domain knowledge. For instance, for a weather forecaster, normative knowledge might be ways in which precipitation is measured, or how to express uncertainty in forecasts on a probability scale; this is in contrast to a weather forecaster's substantive knowledge regarding factors affecting likelihood of precipitation. Clearly much normative knowledge is domain specific and, as such, this sort of expertise may be indicated in the same way as domain-specific substantive knowledge: through examination of an expert's CV, publications, and references from colleagues.

However, in a probabilistic EKE all experts should ideally be able to express uncertainty probabilistically.[3] This kind of normative knowledge tends to be more generic. Although some domains of expert activity may be associated with greater experience of working with probability it could be difficult to establish which are which *a priori*, and authorship of papers with statistical content are not necessarily evidence of statistical expertise, particularly since most papers are multi-authored. Qualifications and attendance of relevant training courses may be better evidence of expertise in probability assessment, but could be difficult to appraise (e.g. Was the syllabus appropriate? What exactly was the level of attainment?). Thus in the main I propose that normative expertise will usually need to be appraised by specially designed tests after long-listing.

16.2.2.3 Social Expertise

Many indicators of expertise that could be (have been) used are bad in the sense that they are weakly, or unreliably, associated with expertise. These indicators often reflect the social aspect of expertise, such as job title or position. As I already mentioned, the problem is that title or position can be attained for numerous reasons unrelated to expertise in the field in question, for example, nepotism, 'old-boy' networks, willingness to take on management or other administrative roles, or simply being in the right place at the right time. Similarly, the reputation of the organisation where the expert is stationed may be only loosely related to ability. However, it should be noted that sometimes a big name from an elite institution might be a useful asset on a project, for instance lending it credibility and thereby facilitating the recruitment of other experts. Further, the principle of selecting the 'best' experts might be violated in order to have balanced representation of different interested parties on the expert panel and/or to demonstrate transparency and openness of the elicitation process. Thus, experts may sometimes be selected for reasons other than the quality of their knowledge.

Another poor indicator of expertise is confidence. It has been shown that people tend to ascribe greater expertise to those perceived as being more confident (evidence that people use what has been named as the 'confidence heuristic', Price and Stone 2004). However, like most heuristics, while there is a certain amount of truth to it, the evidence is that the relationship between confidence and performance is actually rather weak (e.g. Gibbons et al. 2003; Phillips 1999; Rowe and Wright 1996; Rowe et al. 2005) so caution should be exercised in using peer assessments that might be influenced by perceptions of the target's confidence or other personality traits related to confidence, such as charisma, extroversion,

[3] Although potentially non-probabilistic modes of expressing uncertainty, such as natural-language terms, could be used these have not been found to be easily converted to the probabilities usually required for policy and decision-making (see e.g. Dhami and Wallsten 2005; Wallsten and Budescu 1995).

drive, ambitiousness, and self-assuredness (possible effects on the quality of *group* judgments of selecting judges who are high in such characteristics are discussed later).

Giving numerous presentations and/or being a prolific writer may also not be associated well with a high degree of substantive knowledge. For example a popular speaker might receive frequent invitations to speak, but have a limited repertoire, while numerous publications may reflect status as head of a laboratory rather than up-to-date knowledge of the field. Similarly, having a large media presence is no guarantee of expertise since this may be more indicative of eloquence, photogeneticity, and contacts than anything more substantial (although there are undoubtedly many scientists with high media profiles who are also very knowledgeable the presence of criteria for media inclusion other than knowledge will weaken the relationship). However, again, these 'high-profile' figures can be useful for attracting other, perhaps more genuine, experts to an EKE.

Another method for identifying expertise that it is often used in peer recommendation. While references from other experts might be usefully employed to establish a long-list (e.g. by means of 'snowballing', see Part II), or to establish that a potential expert has credibility within his or her field, uses beyond this—such as for screening or weighting experts—must, however, be considered carefully. Reasons for this include problems of establishing the expertise of recommenders in the first place, and difficulty in controlling the basis of peer-assessments: peers may well be using weak criteria such as confidence and charisma when evaluating expertise levels. I will return to evaluate peer-assessment of expertise in more detail in the next section on the measurement of expertise.

16.2.3 Measuring Expertise

When measuring things we have two main concerns: the reliability and validity of those measures.

16.2.3.1 Reliability and Validity of Measurement

A reliable measure will give fairly consistent readings of the true underlying quantity; 'fairly' because even if the quantities being measured are physical ones, there will be some error in the measurement. For example, the length of a steel ruler will vary slightly with temperature. When the quantity to be measured is a psychological one—such as a degree of belief, or extent of knowledge—then this measurement error can be quite large.

Validity refers to the extent to which a measure is actually measuring the target quantity. Again validity of measurements tends to be higher for physical than psychological quantities but it is still an issue for both. Measurement of the speed of light, for instance, will need to be done under ideal conditions—a perfect vacuum

away from gravitational fields et cetera—these ideal conditions may be difficult or impossible to create. In psychology, we may seek to elicit 'true' beliefs but there are a number of reasons why expressed beliefs may not correspond perfectly with beliefs actually held: people may not have access to their beliefs, or if they do might not be motivated to express them accurately, or beliefs may not be fully formed in the first place, and thus may be created at the time of testing, potentially influenced by the test instrument. The same holds for the measurement of any psychological quantity although the problems are perhaps greatest for self-report data, particularly when there are strong social conventions and/or the need for ego-protection. For example, it is difficult to elicit true attitudes via self-report towards people of different races, or giving to charity, so researchers prefer to observe physiological responses to appropriate stimuli, that is not necessarily under conscious control, or actual behaviour where 'actions speak louder than words'. Getting experts to comply with such procedures would be a challenge, though!

Other tactics to improve the validity of measurement of psychological variables are to use multiple measurements, perhaps of different types (self-report, behaviour, or physiological responses), and paying people to ensure high degrees of motivation, perhaps in combination with 'proper scoring-rules' that are designed to reward truthful responses. Ensuring the subjects of research also trust the researcher is also important to ensure validity of responses.

Ecological and Face Validity

A number of different types of validity are distinguished but of particular relevance here are 'face' and 'ecological' validity. In the current context, face validity refers to experts *perceiving* measures of their expertise as actually measuring their expertise, while ecological validity refers to measures of expertise measuring an aspect of expertise as it is actually practiced by experts in their everyday job. Usually these two types of validity will be related such that a measure low in ecological validity will also often lack face validity: a test of skills not used by experts in performing their job will tend not to be seen as valid by those experts. Lack of face validity does not necessarily imply lack of ecological validity, though: a test might be good one of experts' everyday skills but just not appear to be so. I will argue below that conclusions of poor performance in experts may sometimes be due to lack of ecological validity of the measures of expertise used, however, it is worth noting that even if there is ecological validity, conclusions regarding the quality of expert judgement may also be degraded as a result of poor motivation to respond by experts who perceive a lack of face validity.

The Relationship Between Reliability and Validity

As I have previously commented (Bolger and Wright 1993), if measurement is unreliable then its validity is impacted too. Take, for instance, a metre ruler that

changes length by plus or minus a centimetre then, for any single measurement, this ruler cannot be considered an accurate (valid) measure of length (if a centimetre either way is critical). If the change in length is random then taking several measurements with the ruler will allow us to improve accuracy: such repeated measurement is common practice, for example, in medicine. If the change in length is not-random, in other words the ruler has a *bias* towards measuring too high or too low, then that can also be accommodated once the direction and extent of the bias is discovered, again by repeating measurement many times. In the extreme, our ruler might have bias but no variation, for instance, it always measures 1 cm too high or too low. This being the case, the ruler is now a reliable measure, but not valid, unless the bias is perfectly understood, in which case the ruler becomes both reliable and valid (once adjusted for bias).

The Reliability and Validity of Measures Versus Judgements

I have thus far been discussing the reliability and validity of *measures* of expertise (i.e. experts' beliefs, knowledge, judgements etc.), however, what we are really interested in is the reliability and validity of the expertise itself. Of course, we cannot directly observe expertise, only its effects (on what the experts do and say) thus we must evaluate expertise on the basis of measuring its effects. Thus our measures of expertise are a proxy for the real thing, which means at times we may be tempted to talk about a measure as if it *is* the real thing (thereby implying that the measure is perfectly reliable and valid): I will try to avoid doing this.

I have spent time on this discussion of aspects of reliability and validity because they are important not only for the initial selection of experts for EKE but also for the evaluation of expertise during and after the process, in particular, for screening and weighting of experts. Accordingly, I will now consider issues of reliability and validity in the assessment of expert judgements of uncertain quantities from the perspective of the experts' substantive and normative expertise in turn.

16.2.3.2 Measuring Substantive Expertise

Returning to my definitions of expertise above it is clear that for an expert to be expert he or she should demonstrate good performance in the domain in question where 'good' can be defined either in absolute or relative terms. In the context of judging uncertain quantities we would expect that an expert's judgments would be reasonably close to the actual value of the target quantities (i.e. good in absolute terms) and better than the judgments of a non-expert (i.e. good in relative terms). In both cases, performance advantages will be manifest on average and in the long run, as there will inevitably be a degree of error in judgments leading in accuracy varying from one occasion to the next. Of course, judgment domains vary in difficulty from

impossible[4] to easy, so relative assessments, that control for variation in difficulty, are often to be preferred (although expressions of confidence associated with a judgment allow experts to indicate the difficulty level they perceive—see next section on normative expertise).

Tests of Judgment Accuracy

Ideally an elicitor wishes to assess each expert on many judgments of the (exact same) target then produce an average error score from the differences between each judgment and the true answer. This error score can then be compared to some benchmark for satisfactory performance (i.e. absolute accuracy) or to the performance of others (such as non-experts or other experts, e.g. for screening or weighting). However, this ideal procedure is not usually possible because the very reason for performing EKE is that the value of the target variable is unknown. For relatively short-term forecasting problems the true value of the target variable (sometimes referred to as its 'realization') will become available within a useable time-frame, examples being short-term forecasts of precipitation, sales or stock prices. In other situations it will be necessary to use judgments for similar variables to the target rather than the target itself, but where the answer is already known (sometimes referred to as 'seed variables'—and in sufficient numbers to reliably measure expert performance.[5]

More generally, there is not very much or persuasive evidence that past performance predicts future performance (even when target is the same as the test e.g. Genre et al. 2013), although, recent longitudinal studies of geopolitical forecasting show some promise in this respect (see e.g. Tetlock and Gardner 2016; Hanea et al. Chap. 5). Reasons for a poor relationship between past and future judgment performance may include regression to the mean and lack of generalizability of skill (e.g. Solomon et al. 1985): the latter exacerbated by the fact that expertise is associated with specialism. Heterogeneity, and/or over generality, of test items may

[4]For instance, stock price movements have been characterized as random (e.g. Fama 1965). Although more recent research suggests that stock markets are, in fact, predictable in the long term (e.g. De Bondt and Thaler 1989) it is still agreed that it is not in the short-term, contrary to the beliefs of 'day-traders'. It may often be the case that 'experts' believe there to be predictability where there is not, or it is rather low. In such situations, there can, of course, be little or no expertise (see e.g. forecasting of GDP growth, Budescu and Chen 2015) nor variation in performance. Further, perceived ability where there is none is another name for 'overconfidence' (more generally, insensitivity to task difficulty will lead to miscalibration).

[5]Bolger and Rowe (2015a) identify a number of problems with this approach, including finding a sufficient number of suitable seeds—ones that draw on the same expert knowledge as the target. They also comment that this 'Classical Method'? is atheoretical, in that it is not founded on any particular conceptualization of expert knowledge, and propose a cue-based approach that would provide a reasoned basis for the selection of similar seeds (i.e. those that are related by the cues used to judge them). This cue-based approach is outlined in Sections "Cue-Based Judgements" and 16.2.4.2 below.

also play a role: the ability to make judgements about one sub-area of a knowledge domain may not predict ability in another sub-area, while generic test items may not predict performance well in the more specialist target sub-area. There are also practical problems such as getting busy experts, who often hold themselves in high regard, to take a test, and not cheat (e.g. if the test is completed remotely). For all these reasons, in most cases other measures of expertise are used, such as social indicators or self- or peer-assessments.

Social Indicators of Expertise

Perhaps the most commonly used indicators of expertise are social markers such as job title or role—which can be used to identify potential experts, as discussed above—and metrics such as years of professional experience, and number of awards, citations, patents and publications. All of these can and have been used by automated systems to select experts (e.g. Moreira and Wichert 2013) so could be a useful tool, particularly if EKE becomes routine. Burgman et al. (2011) argue that the use of such indicators of expertise is justified by what they refer to as the 'social expectation hypothesis': society, which includes experts themselves, expects that more experienced, better regarded, and more formally qualified individuals have privileged access to knowledge through specialist training, and therefore perform better. They go on to test this hypothesis by examining the correlations between measures of years of experience, number of publications, extent of professional or academic qualifications; and performance measured by peer- and self-assessments, and questions designed to test substantive expertise (in topics such as animal and plant biosecurity, weed ecology, and public health). Moderate to high correlations between peer assessments, experience, publications and qualifications—and between self- and peer-assessments—support the structure of the social expectation hypothesis. However, weak (and not statistically significant) correlations of performance as measured by peer-assessments or knowledge-test respectively, suggest that the social indicators are not good measures of true substantive expertise.[6]

Peer-Assessed Expertise

In a different approach to peer assessment than collecting references, Germain (Germain 2006; Germain and Tejeda 2012) has developed the Generalized Expertise Measure (GEM), a questionnaire which seeks to capture aspects of the knowledge and person-based views of expertise that can be used by someone to assess the

[6]However it must be stressed that this is just one study (which fails to report all the potentially relevant correlations). Further, we do not know the extent to which the tests of substantive expertise are good measures of *actual* expert performance on, for example, a real-world risk-assessment or forecasting task.

expertise of another. The GEM's 18-item scale contains six 'objective' measures—largely social indicators (e.g. education, training and qualifications)—and 12 more subjective items (e.g. self-assurance, potential for self-improvement and intuition). I have already discussed above the possible limitations of social indicators and there is reason to believe that the subjective items in GEM will have even less validity (e.g. confidence has been found to be poorly related to performance, see e.g. Gibbons et al. 2003; Phillips 1999; Rowe and Wright 1996; Rowe et al. 2005). It should be noted that since there are twice as many subjective items as objective, the former are implicitly given more weight than the latter: this reflects Germain and Tejeda's finding that individuals place more weight on the subjective than objective items when assessing the expertise of others. However, the subjective items could have a use for identifying potential sources of bias in interacting groups (see e.g. Bolger and Wright 2011). Further to this, it should be stressed that the GEM scale is designed to measure *perceptions* of expertise rather than actual possession of expertise, and, as such, Germain and Tejeda did not attempt to validate it against measures of 'true expertise' (i.e. either past or subsequent job performance in the target role): the authors do show that GEM has good internal consistency, though.

To my knowledge GEM has so far only been used once to try to differentiate levels of true expertise. Alvarado-Valencia et al. (2017) sorted experts in demand forecasting into high and low expertise groups on basis of GEM. The scale, particularly the 'objective' knowledge sub-scale, had high reliability while high-scoring experts made more accurate forecasts—and more useful adjustments to forecasts—than low thereby demonstrating validity of GEM as a tool for identifying expertise. The authors also commented that selection might be further improved if personality characteristics associated with good forecasting could be identified and a questionnaire tailored to specific domain knowledge requirements was developed. I concur with the these points and suggest that, contingent on further research into GEM's reliability and validity—and into the potential for adding further items such as creativity and the ability to argue and communicate (see e.g. Ivlev et al. 2015)—GEM might best be used in conjunction with other measures of expertise that are filled out by the experts themselves. These other measures should be ones that are more readily completed in an impartial manner than GEM, such as the Expert-Skills Questionnaire to be described in Part II. I will return to the potential of the GEM for identifying expertise later in the section on selecting experts for the long list in Part II, meanwhile I will briefly turn to consider assessment of expertise by the experts themselves.

Self-Assessed Expertise

Excluding direct measures of performance/knowledge (already discussed above) possibilities include:

- Personality. As already mentioned, personality traits may be associated with being a 'good expert' in an EKE (Alvarado-Valencia et al. 2017). Certain characteristics may help experts integrate different information sources and

perspectives, keep them motivated and on task, or smooth the interaction with the elicitor and other experts. For instance, of the 'Big 5' personality traits: openness to experience; conscientiousness; and agreeableness would seem particularly relevant to being successfully elicited.[7]

- Cognitive skills: IQ, working memory, creativity (the latter suggested by Ivlev et al. 2015)—are all likely to be associated with substantive expertise. For instance, we have found that working-memory capacity is related to the ability to learn the relationships between cues and criterion in a multiple-cue probability learning (MCPL) task (Bayindir et al. 2017).
- Susceptibility to biases: tests of rationality (e.g. choosing according to maximization of expected utility, updating beliefs in line with Bayes' Theorem), use of intuition (e.g. as measured by the 'Cognitive Reflection Test'—Frederick 2005), and heuristics (e.g. anchoring or availability) when making judgements. All of which could also be relevant to normative expertise, for example, tendency to be overconfident.

It is arguable, however, whether such generic qualities really constitute substantive expertise itself but potentially could support or enable acquisition of substantive expertise (as in our MCPL study) and/or be helpful in the EKE process, such as the personality traits identified above.[8] This is all speculation at present, though, and warrants some research attention, however, practical utility might be compromised by expert reluctance to take such generic tests (i.e. they may question their relevance): tests of specific domain knowledge might therefore be easier to 'sell' to experts.

A compromise position between using generic instruments, or specific knowledge tests, is to try to determine details of the judgment task at hand and the skills and qualities each expert brings to bear as a result of their experience at performing that particular task, and any relevant training they have received. This is the approach I take in my E-SQ described in Part II and, although relevant to measuring substantive expertise, is based on an analysis of studies of the realism of probability judgment, hence I will defer further discussion of this issue to after I discuss how realism of probability judgment is measured.

Effects of Self- and Peer-Assessed Expertise on Group Judgements

Generally in EKE we wish to garner the judgments and other opinions of several experts in order to try and maximize the information base, and reduce error and bias

[7]The other two traits being extroversion and neuroticism.

[8]It is possible that the advantages of some personality traits are protocol-dependent. For example, conscientiousness might be good for a remotely administered elicitation such as is often the case in Delphi, while agreeableness might be particularly helpful in protocols that require face-to-face interaction. Openness to experience is probably a useful characteristic in both protocols as it should assist opinion change towards the true value.

(through aggregation). However, in interacting groups there are social processes that can undermine the goal of improving the outcome by means of adding more experts. For example, dominant individuals can seriously skew the opinions of the group as a whole as well as some other undesirable effects, including: 'premature closure' (i.e. coming to a judgement before all the evidence has been considered), and suppression of minority opinion—both features of 'Groupthink' (Janis 1982). Measures are taken in EKE to reduce such effects, such as having a strong facilitator in the Sheffield Method or not permitting experts to interact directly in the Delphi and Classical methods. Despite such measures, experts' opinions of their own and other panellists' expertise could still affect the quality of an EKE's outcome. A particular case in point is confidence. If experts vary in confidence in their own knowledge then those higher in confidence may be more likely to stick to their original views, while those who are less confident might be likely to shift their opinions towards the more confident. This tendency is likely to be exacerbated when the experts can see each and hear each other (i.e. perceive the differences in confidence between panel members), although, it could also occur in Delphi where it is common to feedback assessments of confidence. This would all be fine if confidence was clearly related to expertise, but as I have already discussed, there is no evidence that this is the case.[9]

16.2.3.3 Measuring Normative Expertise

Many aspects of normative expertise, such as the ability to use 'tools of the trade' (measures, procedures, software, tests, common models etc.) can be assessed in the same way as for substantive expertise, for example, from CV's, publications and peer-recommendations. As is the case when measuring substantive expertise, provisos regarding the reliability and validity of these measures apply: for instance, CV's are 'sexed-up', the contribution of an expert to publications may be obscure, the appropriateness and quality of training courses may be largely unknown, and peers may be influenced by salient surface qualities (e.g. confidence and charisma) rather than more objective qualities (e.g. evidence of ability to analyse data) that may be difficult to observe: the latter is a bias that has been proposed as a cause of overconfidence, for instance, in personnel selection (see Griffin and Tversky 1992). Further, these sorts of measures are pretty blunt instruments that often do not capture the nuances of a particular elicitation. In view of these considerations,

[9]Indeed, the 'Theory of Errors' (Dalkey 1975; Parenté and Anderson-Parenté 1987), which is the leading account for why the Delphi technique works, assumes those who stick are on average closer to the truth than those who shift—Bolger and Wright (2011) propose that in order to achieve 'virtuous opinion change'—i.e. opinion change towards the truth—rationales for opinions should be fed back between Delphi rounds rather than confidence as the former will be better indicators that the expert is knowledgeable about the topic than the latter.

it may be desirable to use a specially designed questionnaire, such as that described in Part II, to measure aspects of normative expertise that are particularly relevant to the assessment task in hand.

For a probabilistic EKE there is a necessity to express uncertainty probabilistically: this is often the stumbling block for experts, as substantive and normative aspects of expertise frequently diverge when it comes to probability judgement. This is because, although many experts might be required to express uncertainty surrounding judgements, quantification of this uncertainty does not form part of most experts' *modus operandi*. Verbal uncertainty statements are generally the preferred way of expressing likelihood, however, the problem with verbal probability statements is that they are unreliable—they are not consistent either between or within experts (see e.g., Dhami and Wallsten 2005; Wallsten and Budescu 1995)—and by dint of this also have low validity. At least we assume low validity as we cannot properly test it—I describe how we test the quality of probability judgements next.

There are two basic ways of assessing the quality of probability judgement: these are known as 'coherence' and 'calibration' and can be considered as reliability and validity criteria respectively (Bolger and Wright 1993). Coherence refers to the consistency of probability judgements and their conformity to the laws, or rules, of probability. Consistency and conformity criteria are related in that they both apply across a number of probability judgements, rather than to a single judgement, and conformity with probability theory is the logical and mathematical basis for consistency. An example of 'incoherence' is that, logically, the judged probabilities of a set of mutually exclusive and exhaustive events must sum to certainty, however, the judgements of experts often do not (Tversky and Koehler 1994): this is a phenomenon referred to as 'sub-additivity'. Another commonly observed example of incoherence is so-called 'conservatism', where people, experts included, fail to update their probability beliefs sufficiently in the light of new information, where sufficiency is determined by comparison with the degree of updating indicated by Bayes' theorem (Edwards 1968).

Calibration refers to the correspondence between subjective and objective probabilities. Calibration is also sometimes referred to as 'realism' because a good correspondence between probabilities in the head and in the world implies that the judged probabilities are realistic assessments of uncertainty. As with coherence, calibration is measured over a set of judgements rather than a single one. For example, experts may be asked to make a series of judgements about the range of uncertain quantities for given probabilities such as "what are the highest and lowest judged values of x such that the true value of x will fall within this range on 90% of occasions it is observed?" Over a number of such judgements—say ten—a realistic ('well-calibrated') expert will have nine true values (sometimes referred to as 'realizations') falling within his or her given ranges. In contrast, an 'overconfident' expert will have fewer than nine values falling within their ranges (i.e. they tend to give ranges that are too narrow); this is what is typically observed (e.g. Griffin and Brenner 2004; Lichtenstein et al. 1982; Lin and Bier 2008).

It should be clear that in order to calculate calibration in this manner you need to know what the true values are. The consequences of this limitation is that it is difficult—and sometimes impossible—to assess calibration for most variables that we wish to estimate using expert judgement because there is little or no historic data directly pertinent to the target variables of interest to use as realizations (and, if there were, then we may not need to use expert judgement in the first place). The way around this is to calibrate experts on similar variables to the target where the answer is known: this is what is done in the Classical Method (see Chap. 2), where the calibration variables are known as 'seed variables'.

As we pointed out in Bolger and Rowe (2015a, b), the assumption that good calibration on the seed variables (as opposed to accuracy, which I discussed above in the section on measuring substantive expertise) generalizes to the target variable is a big one that has rarely properly been put to the test. This is because, since the realizations of real-world target variables are often not available for quite some time, the differential weighting of experts using calibration has typically been evaluated using 'cross-validation', whereby calibration across a set of seed variables is used to weight experts' estimates of another seed variable. Although some of this cross-validation research produces an advantage for the Classical Method relative to equally weighting expert opinion (Cooke 2014; Eggstaff et al. 2014) other cross-validation research does not (Clemen 2008; Lin and Cheng 2009), it depends on the cross-validation procedure that is used. We argue (Bolger and Rowe 2015b) that controlled experiments and simulations are required to settle the issue, but we are sceptical that significant practical advantages for the Classical Method will be found due to problems of low reliability and validity of calibration measures (Bolger and Rowe 2015a).

At the start of the section on measuring expertise I differentiated absolute and relative assessments. The Classical Method weights experts in terms of their absolute performance (mostly normative—calibration—tempered by some consideration of substantive expertise) in the sense that each expert's independent performance on the (same) seed variables directly affects an expertise score that is subsequently used for weighting relative to other experts on the panel. In contrast, Budescu and Chen (2015) propose assessing each expert's performance relative to the performance of the group (or 'crowd' as they call it) and use this as the basis of weighting. More specifically, Budescu and Chen's Contribution Weighted Model (CWM) assesses how much each expert contributes to the overall crowd performance, increasing weights for positive contributions and reducing them for negative. Although CWM could potentially be applied to judgments made for seed variables it has not been so far, rather it has been tested on longitudinal data from forecasting geopolitical events and economic variables (inflation levels and GDP growth). In these instances, CWM produced significantly better-calibrated probabilistic forecasts than unweighted crowds (i.e. including all experts' forecast on an equal basis) in most knowledge domains and also performed better than using absolute performance weights. This latter advantage can be credited to the fact that CWM always gives higher weights

to those who do better relative to the crowd and thus takes into account variations in the difficulty of the items. For absolute scoring an expert could get a high weight if they just get easy items right (i.e. where the crowd majority also get it right) whereas CWM particularly rewards correct answers to difficult questions (i.e. those experts who make correct forecasts against the majority crowd predictions). CWM requires more testing, though, and also suffers from the same primary limitation as the approach taken by the Classical Method, namely, the need for sufficient and appropriate data to measure performance in the first place (with the added problem of also needing sufficiently large numbers of experts whose judgments can be compared—although the application assumed by Budescu and Chen (2015) is a 'crowdsourcing' one i.e. lots of relatively inexpert forecasters).

A general conclusion from decades of research into the quality of probability judgement of both experts and 'laymen' is that it is commonly unreliable and biased. Typical findings are, of both incoherence and miscalibration, most usually overconfidence, as I indicated above. This conclusion that probability judgement is typically unreliable and biased has been questioned, for instance, in terms of the 'ecological validity' of tasks and judges used in much of the research this conclusion is based on (i.e. tasks are artificial and judges are inexperienced at performing them; see Bolger and Wright 1994), and as a result of problems in measuring uncertainty beliefs (e.g. regression effects due to measurement error combined with the requirements to map beliefs onto a fixed scale; see e.g. Olsson 2014). Despite this, it seems clear that many people, including those regarded as 'expert' in some domain, have difficulty with expressing uncertainty probabilistically with potentially negative consequences for the quality of the probabilities derived from a probabilistic EKE.

Lack of normative expertise may be a reason why expert probability judgements are sometimes little or no better than lay judgements, and demonstrate the same biases. Koriat et al. (1980) propose that probability judgement progresses in three stages. First, memory is searched for relevant information. Second, evidence is assessed to arrive at a feeling of uncertainty. Third, the feeling has to be mapped onto a conventional metric—if experts are unfamiliar with performing this mapping then the quality of the resulting judgement of uncertainty may be poor. Thus, with regard to this third, mapping stage, lack of experience at expressing uncertainty in the form of numeric probabilities may lead to a corresponding lack of reliability, and/or incoherence, in statements of probability, even if the underlying uncertainty assessment processes (stages one and two) are sound. Further, as I argued above, there will be knock-on effects of mapping failures that lead to incoherence because reliable (i.e. coherent) probability judgements may be a prerequisite for valid (i.e. realistic or 'well-calibrated') judgements (Wright et al. 1994). For these reasons I (and others, e.g. Phillips 1987) propose that experts are trained in expressing uncertainty as probabilities: determination of the need for such training is something that perhaps can most usefully be assessed during the selection process rather

than after it (e.g. it can help in assessment of the timeline, and selection of EKE protocol).[10] Accordingly I propose the identification of training needs a part of my E-SQ described in Part II.

16.2.4 The Nature of Expertise in Judgement of Uncertain Quantities

Although problems in mapping subjective assessments of likelihood onto numeric probability scales are undoubtedly a major determinant of the final quality of uncertainty judgements there is reason to suspect that errors and biases in judgement might also occur during the first two of Koriat et al.'s (1980) stages. For this reason I argue that training, while necessary, is not in itself sufficient to ensure the quality of expert judgement in probabilistic EKE. To better understand the potential sources of difficulties faced by expert judges—and thus to best assist them in producing the highest quality judgments possible—I believe that it is useful to take a step back and consider how an intelligent system, in general, might go about making judgements of uncertain quantities across a range of different judgment tasks. Once we understand this better we can then proceed to develop measures of expertise that allow the selection of the best experts *for a particular judgement task* and to improve the fit between experts and EKE through training.

In general, approaches to measuring expertise to date have not been based on any deep theory or analysis of expert judgment, and consequently offer limited possibility for validation. Further, research has tended to focus on the outputs of judgment—how accurate or well-calibrated they are—rather than the psychological processes leading to the judgments. This research still needs to be done but I offer the following analysis as a starting point.

So how are judgements of uncertain quantities made? In order to answer this question we need to break it down into two smaller questions, namely 'how does one go about making judgements of quantities?' and 'how does one estimate the uncertainty in those judgements?'

16.2.4.1 Judging Quantities

At its base, judgment involves applying a mental model of the world to some data. For example, if I need to judge how long it will take to get to the nearest airport on a Friday afternoon I might use a mental model composed of a real or cognitive map to estimate the distance, combined with knowledge of any current construction

[10]There is some empirical support for the suggestion that relative frequency is a more natural way of representing uncertainty than probability (Gigerenzer and Hoffrage 1995) thus posing questions as relative frequencies rather than probabilities might be an alternative to training.

work (from the internet, memory, or asking colleagues who have recently made the trip), plus an understanding of peak travel times (e.g. extended rush hour on Fridays due to people going away for the weekend) and so on. This model is created on the basis of observed or inferred regularities, which may have a known—or hypothesized—causal basis (i.e. theoretical underpinning), or may be known rather from manifest associations (i.e. empirical underpinning), or both. Models can differ in the reliability and validity with which judgements can be made, which is in part due to the inherent predictability of the domain, and in part due to the processes by which models are developed, acquired and applied.

Our knowledge for models comes from one or other, or both, of two sources: observation, or being taught about them by others (a distinction sometimes referred to as 'learning from experience versus learning from description' e.g. Barron and Erev 2003; Hertwig et al. 2004; Rakow and Newell 2010). The reliability and validity of the models are therefore contingent upon how well they have been observed (determined, for instance, by the quantity and quality of data available, and how systematic the observation is) and/or how well the model has been communicated (determined, for instance, by the coherence of the theory, and expertise of the teacher). Acquisition by either route is further influenced by model complexity, both in terms of the number of variables, and the nature of the relationships between them (e.g. direct, mediated, hierarchical etc.): more complex models will generally be harder to learn, and be more subject to error in their application.

The models held by experts that they use to make judgements of quantities may be fairly explicit and/or formal, particularly those that they have acquired in their professional training (i.e. learned from description), or they may be implicit and/or informal, particularly those acquired during their professional experience. In some cases models might be a hybrid of both explicit-formal models and implicit-informal ones. Experts' judgement models might also either be well-established, or created de novo for a particular circumstance, or again something of both.

From the viewpoint of selecting and preparing experts for a probabilistic EKE then it could be useful to find out what explicit/formal models they use for assessing uncertain quantities, and also the extent and circumstances of any judgemental adjustment to the outputs of formal models. I propose that a questionnaire such as that described in Part II is used for this purpose. If face-to-face EKE is performed then this offers another opportunity to elicit information about the use of formal models.

The same methods can be used to find out about the implicit/informal models held, but it may be difficult to get at them since—as I discussed in the section on Procedural versus Declarative types of expertise—expert knowledge tends to move from explicit to implicit with experience, and this implicit knowledge may not be represented in a way that allows easy access to its owner. So, although there is no harm in asking—hence such questions included in my questionnaire described in Part II—we can by no means guarantee that we will get at the 'truth'. Face-to-face we may do a better job, as we can apply specialist elicitation techniques such as card-sorting, repertory grids, and verbal protocols (see e.g. Bolger et al. 1989)

to reveal the underlying structures and processes of the experts, or use methods as cognitive mapping and influence diagrams to elicit causal models directly (e.g. Eden 1988; Howard and Matheson 2005; Oliver and Smith 1990).[11] All this is, however, beyond the scope of the current paper because such methods have rarely been applied in the context of probabilistic EKE,[12] although they are more commonly used in the development of expert systems (e.g. Bolger et al. 1989); the noted problem of potential lack of accessibility of expert knowledge—known as the 'knowledge acquisition bottleneck'—remains a serious barrier, though. The purpose of the questions included in my E-SQ are not to elicit the actual implicit models of experts but, amongst other things, to determine the balance and interaction of use of 'intuition' and formal models so that we can better determine the potential quality of both quantitative assessments and uncertainty estimation.

I have identified five different ways in which quantitative judgments can be made: rules; problem solving; evidence accumulation or argumentation; pattern matching; cues. I will now outline these in turn as each way has implications for the accuracy of the judged quantities as well as how uncertainty might be evaluated, and thus the quality of probability judgment.

Rule-Based Judgement

Starting in the 1970s, 'expert systems' began to be developed based on 'production systems': expertise was modelled by if-then rules applied to a knowledge-base, for instance: 'if x is true (i.e. matched in the knowledge-base) then do (say) y' (Hayes-Roth et al. 1983; Klahr et al. 1987; Waterman and Hayes-Roth 1978). For quantitative judgments this could be: 'if x is true, or attains a certain value, then y must have the value z'. Broadly, the rules used be experts can be categorized as either algorithms or heuristics. In the former case, given an input a rule produces the correct answer (e.g. a formula for converting Celsius to Fahrenheit), possibly with some error (e.g. from misapplying or misremembering the algorithm). In the latter case, heuristics produce approximate answers from inputs (e.g. clarity of an image can be used as a guide to its distance from the viewer).

Problem Solving

Expertise is often seen as a facility to solve problems by, for instance, making connections, or distinctions, that most people do not see. Some early artificial intelligence (AI) systems were attempts to capture this aspect of expertise (e.g.

[11] In principle, many of these things could be done remotely but I am not aware of any existing protocols or software to support this.

[12] However, Quigley and his colleagues use maps when eliciting priors with engineers assessing the reliability of new systems (e.g., Hodge et al. 2001; Walls et al. 2006—with the aerospace industry).

Newell et al. 1959; Newell and Simon 1972) and focused on extracting rules that permitted them to solve problems such as the Missionaries and Cannibals and Tower of Hanoi problems. Quantitative estimation could be characterized as a decompose-recompose problem. For example, take estimating the number of letters posted each day in US (MacGregor et al. 1988, 1991): first, estimate the population of the US; next, make a correction so as to calculate the adult population; then reflect on how many letters one personally sends in a year; divide this by 365 to get a daily figure; finally multiply this daily figure by the estimate of the adult population.

Pattern Matching

One of the earliest and best-known conceptualizations of expertise, which was developed from studies of chess masters (De Groot 1965; Chase and Simon 1973) is that experts remember patterns that lead to particular outcomes, such as configurations of chess pieces that lead to winning or losing.

A development of this idea is Recognition-Primed Decision Making (RPDM) that has primarily been applied to understanding decision-making in emergency or other high-pressure situations (Klein 1998). For example, expert fire-fighters recognize combinations of flame, smoke and building types indicating a particular type of fire, and this informs their fire-fighting strategy. Although RPDM describes categorical rather than quantitative judgements there is no reason why a particular configuration could not be associated with a set of quantities, for instance, amount and thickness of smoke, with a high ratio of combustible to non-combustible materials, means a fire of temperature x.

Cue-Based Judgements

Rather than learning relationship between patterns and outcomes experts can learn the covariation between several discrete or continuous cues and a criterion (e.g. Brunswik 1955). For example, a doctor learns that symptoms such as a temperature of x1, blood pressure of x2, and pulse-rate of x3 are associated with disease y. If the criterion is continuous, for instance, recovery time, then this can be a means of quantitative judgment.

Support Accumulation and Argumentation

Toulmin (1958) proposed a model of argumentation whereby claims are supported by data, the relevance of which are established by ‘warrants’ and ‘backing’. An example would be as follows: “Company X can take over Company Y (claim) because Company X has acquired a majority share of Company Y (datum) permitting the takeover since owning more than 50% of a company gives a controlling stake (warrant) as set out in Company Law (the backing)” The same set-up could

be used to justify a quantitative claim, for instance, levels of sales of a product in a year's time. A similar idea in this category is that people seek evidence in support of a favoured (or 'focal') hypothesis relative to an alternative (e.g. Brenner 2003; Griffin and Tversky 1992; Koehler et al. 2002; Tversky and Koehler 1994). These approaches—referred to as 'support' theories—are all probabilistic (i.e. the probability of the focal or alternative hypothesis being true is increased as support is received) and as such are descriptive counterparts of normative Bayesian updating. While support theories typically describe revision of judged probabilities of events (e.g. death from a particular cause) there is no reason why the focal hypothesis could not be about the likelihood of particular value of a continuous variable being achieved (although it raises questions about which alternative hypotheses should be considered).

Some or all of these five methods might be equivalent in the sense that they could be substitutable for each other. For example, Kleindorfer et al. (1993, pp. 85–86) have pointed to similarities between argumentation and cue-based judgments, while the patterns in the relationship between outcomes could equally be modelled as the covariation of cues, and the covariation of cues could be expressed as rules. However, each method has somewhat different requirements in terms of data and processing that can have distinct influences on the quality of judgments at a descriptive level. For instance, as already noted, cue-based representations are better suited to judgements of continuous quantities, and pattern matching and argumentation/support systems to discrete events, while problem-solving and rule-based approaches are lacking in the ability to generalize to new situations relative to the other approaches. Meanwhile, pattern-matching and cue-based systems seem to model what people actually do quite well and thus score highest in terms of psychological plausibility.

The way in which uncertainty is assessed is also likely to be linked to the way in which the initial judgment of the quantity is made, although the strength of this link may vary. For instance, in the cue-based approach criterion values, and the validities of cues used to predict them, are intimately related in Brunswik's 'Lens' model (to be described further in the next section) so value estimation and uncertainty assessment are two sides of the same coin for a Brunswikian judge. In contrast, a judge who judges quantities using problem-solving may have to use a combination of methods to assess uncertainty including top-down computations of errors in problem steps and bottom-up assessments of the effectiveness of heuristics on the basis of experience (which could be similar to cue-based assessments). Of course, it is an empirical question as to both how experts make both sorts of judgment and whether the same process is used for both, even for a Brunswikian judge.

16.2.4.2 Assessing Uncertainty

The cue-based model is perhaps the most well-developed of the above models regarding the process by which uncertainty surrounding judgments of uncertain

quantities might be assessed.[13] For this reason—and because my point is just to give an example of the kind of theoretical context that I believe is needed for expertise assessment—I will discuss the cue-based model only in this section.

Brunswik (1955) proposed the Lens model whereby a person's proximal judgment of a distal variable or state of the world is formed from an analysis of the relative strength of a set of cues acting like a lens between objects in the world and the perceiver. More specifically, cues are stochastically related to a criterion: the correlation between a cue and the criterion is the 'ecological validity' of that cue and the correlation between a cue and the judge's perceived value of the criterion is a judge's 'learned validity' for that cue. The subjective probability of a value of the criterion given a cue can be obtained from the learned validity of the cue while the subjective R^2 value for regression of the perceived value of the criterion on all cues provides the subjective probability of the criterion given those cues (which can be calibrated against the R^2 value for regression of the corresponding true value of the criterion, if available).

Gigerenzer et al. (1991) proposed a theory of how people make probability judgments when asked to make probabilistic forecasts of binary events (e.g. rain, no-rain) or judge the likelihood of correctly answering 2-alternative forced-choice (2AFC) questions (e.g. which is further north, New York or Rome?). Their theory of Probabilistic Mental Models (PMM) is a Brunswikian model in that it is proposed that people learn from experience the predictive validity of cues in their environment and then use these cues both to choose between alternatives and judge probability. Rather than using multiple cues, as in the Lens Model, in PMM judges use just one cue: the best available one. So for example, to answer the question of whether New York or Rome is further north a judge (who is unable to retrieve the correct answer) might use climate as a cue and select New York as being on cooler on average. The cue validity (i.e. the long-run, experienced success of the climate cue in distinguishing northerliness between two alternative locations) is given as the probability of being correct.

Findings of overconfidence in judgments of likelihood of correctly answering 2AFC general-knowledge questions (one of the two main ways in which the realism of confidence has been assessed, see e.g. Lichtenstein et al. 1982) have been attributed to test items being selected so as to be counter-intuitive (i.e. good tests of general knowledge). Such selection means that the cues are successful less often than experience tells people that they should be. In the case of the climate cue, this may have quite high validity across all instances of its application so when used people will state a high probability of being correct. Rome is, however, further north than New York so if the climate cue is used to answer this question people will get it wrong: several incorrect answers associated with high stated probability correct will, of course, manifest overconfidence. When items are not misleading in this way

[13] Support Theory also is well-developed with regard to uncertainty assessment but its primary focus is the assessment of the likelihood of categorical assessments (e.g. of event occurrences, correctness, or truth) rather than values of continuous variables.

(e.g. representatively sampled from an appropriate reference class for a cue) then overconfidence reduces or disappears (Gigerenzer et al. 1991; Juslin 1993, 1994).[14]

In the following discussion I will adopt the cue-based account of how experts might make judgments of uncertain quantities.

16.2.4.3 Limits of Expertise

When selecting experts for an EKE it is necessary to consider what levels of performance we can reasonably expect from our experts, so that we do not waste time trying to find experts performing at levels that are impossible to achieve, or nearly so. Task domains vary greatly in how much 'hard' data is available, how easy it is to generalize from one situation to another, how much variability there is in both phenomena of interest and measures of them, and so on. For instance, sometimes there are very distinct indicators of a target event and sometimes there are not (e.g. measles vs. meningitis), and sometimes there are clear precedents or analogues for events and sometimes there is not (e.g. the launch of an additive vs. a transformative technology). Both these will affect the levels of judgement accuracy attainable by even the most expert of experts.

One way of characterizing judgement tasks that may be useful, is in terms of predictability and learnability: that is, some domains of knowledge are more predictable than others, while some offer more opportunity to improve performance through experience than others. Often the unpredictable and unlearnable domains are one and the same, but not always. For example, predicting the eruptions of volcanoes is difficult and the task is not very learnable either, due to the (thankfully) infrequent occurrence of eruptions, and the rather unique circumstances surrounding them. Likewise, predicting the likelihood of rain one week ahead (in somewhere like the UK with rapidly changing weather) is also tricky but, in contrast to the volcanology situation, there is plenty of data from similar weather situations, with fairly rapid feedback on success of predictions, thus learnability is fairly high. Although learnability and predictability impact on the quality of expert judgements of quantity per se, they also have profound effects on the quality of uncertainty assessment: I will therefore now devote some attention to this matter.

Effects of Learnability and Predictability on Calibration

Bolger and Wright (1994) analysed 20 studies of expert judgement—all of which required the qualification of the judgements with probabilities—and concluded that in half of the studies good and/or well-calibrated judgement (i.e. a good correspondence between objective and subjective probabilities) would not be anticipated *a priori* because the task assessed could not be learned and/or predictability was low.

[14] Although as already indicated, there is still an ongoing debate as to the source and degree (and even existence) of overconfidence in such tasks—see Olsson 2014, for a recent review.

We argued that to learn a task well enough to be able to produce realistic assessments of uncertainty it is necessary to receive regular, rapid and reliable outcome feedback. Returning to the example of weather forecasters making short-term forecasts of precipitation, forecasters quickly find out if these forecasts are correct enough and so can potentially learn how the weather cues they use are related to outcomes. Further, the forecasts cannot affect the outcomes.

Some studies have shown such precipitation forecasts to be well calibrated (e.g. Murphy and Brown 1985). In contrast, life underwriters predicting whether or not claims will be made on applications do not receive good feedback on their risk assessments as claims are usually made a number of years after the application has been assessed. Further, even where feedback is received, it is not as diagnostic as it could be because the underwriter does not know what happened to those applicants who were not given cover—underwriters' risk assessments have been found to be similarly biased to those of students with no underwriting or actuarial experience (see Wright et al. 2002).

Receiving good feedback is necessary, but may not be sufficient, for experts to perform above non-experts—there has to be some predictability in a task to be learned in the first place. For example, one can bet on the outcomes of a roulette wheel, and get feedback very rapidly about outcomes that you cannot influence, but unless the wheel is biased it is not predictable, so your forecasts will be as good as anyone else's—the same may be true of some other domains where expertise is presumed and sought, such as the short-term movements of stocks (which are, in essence unpredictable; see, for example, Malkiel 2011).

Clearly it is important to ascertain whether experts being considered for a knowledge elicitation exercise are likely to produce useful estimates; thus, I have included questions about the nature, availability and speed of feedback in the E-SQ that can be used to screen potential experts for EKE (see Part II); additional questions request assessment of the difficulty of making judgements in the domain. If the results of most experts' questionnaires indicate that learnability and predictability is very low, then the nature of the elicitation exercise may need to be reassessed. However, if high learnability and predictability are generally indicated, it may be considered whether some other approach, such as statistical modelling, might be more appropriate.

Data Available to Experts Regarding the Judgement to be Made

Both experts and statistical models perform better if they have access to good data. If data are sparse we cannot expect experts to be able to make accurate judgements; for instance, predicting the success of a new technology will be difficult because experts will, by definition, have no previous data relevant to the task. Instead, he or she must rely on analogy with similar technological developments in the past—analogies that are likely to be only approximate because the technology will have different features and the world is continually changing. Further, it has been questioned whether probabilities attached to such essentially unique or one-off judgements can be assessed against calibration or coherence criteria (see, for example, Keren 1991).

To establish the quality and quantity of data available to experts, and thereby establish whether expert judgements provided in the planned elicitation exercise are likely to be useful, I pose several questions of potential experts in the E-SQ regarding the nature, calibre and amount of data perceived in the task domain. However, it should be noted that if a large amount of high-quality data are available to experts, then it might be possible to form a statistical or mathematical model to assess the target quantities and related uncertainties rather than use expert judgement. Such models are to be preferred to expert judgements in that they are more consistent (see, for example, Hardman 2009, pp. 10–13), permit experimentation with parameters through simulation, and are readily available if future forecasts are required. Expert judgement may be the only choice, though, if few relevant data are available, or a risk assessment needs to be made quickly, or there are significant new factors not represented in available data.

Ecological Validity of the Elicitation Task

In addition to learnability, Bolger and Wright (1994) also proposed that the quality of expert judgment can be affected by the ecological validity of the elicitation tasks, where 'ecological validity' refers to the match between the tasks used to elicit judgments, and those for which the experts concerned make judgments during their everyday professional activity (see Section "Ecological and Face Validity" above). Clearly with regard to substantive areas of expertise, you would not expect, say, a chess Grand Master to be able to answer questions about food safety, or food scientists to solve chess problems: however, much smaller differences in substantive focus—between food scientists specializing in microbiological versus chemical risks, for example—can also significantly impact on the quality of judgment. For this reason I propose that details of the relevant workday judgments of experts in an elicitation are collected in Part B of the E-SQ (see Table 16.2 below).

With regard to the quality of expert probability judgment, there is a similar need for ecological validity, thus if experts are not used to expressing their uncertainty as probabilities, but rather are quite familiar with odds, then the quality of the elicited judgments should be higher if odds are elicited rather than probabilities. Bolger and Wright (1994) found that in studies with high ecological validity of probability judgment task experts tended to be found to be better calibrated than those studies with low ecological validity of tasks, particularly if learnability was also high. Part C of the E-SQ attempts to determine experts' degree of experience with expressing uncertainty in various forms so that either the elicitation can be shaped around such experience or appropriate training provided to them.[15]

[15] Bazerman and Moore (2008) suggest that experts also need to have coherent models (mental or formal) in order to make good quality (probability) judgements. So models may be added to learnability and ecological validity of tasks as a criterion for well-calibrated experts.

The Nature of the Judgment Process

In Sect. 16.2.4.1 above I proposed several different cognitive processes whereby uncertain quantities might be judged. It can be noted that these processes can vary along a number of dimensions, such as how many steps there are, and the degree of accessibility to consciousness (e.g. problem solving is a relatively conscious analytic approach, potentially requiring many steps, whereas pattern matching is more intuitive and holistic with just one or few steps). It seems plausible that these differences impact on the quality of judgment so that, for instance, more steps may lead to more error but, as is often assumed, use of intuitive, heuristic reasoning leads to more bias: systematic investigation of the relationship between judgment processes and judgment quality in the context of judging uncertain quantities in EKE is clearly warranted.

Similarly, the role of affect (mood and emotion) in judgment and its implications for resulting quality is, as yet, scarcely researched. For example, it is known that mood can influence reasoning and performance (e.g., Ellis and Ashbrook 1989; Isen and Erez 2002)—including mood that is incidental to the task at hand, such as that due to the weather (e.g. Kliger and Levy 2003), and thus could impinge on an EKE session—but clear effects on the quality of probability judgment have yet to be demonstrated. Assuming such effects do exist it is plausible that they are stronger for some judgment processes than others: perhaps those of a more holistic, intuitive nature than those that are analytic and rational.

16.2.4.4 Determinants of the Quality of (Probability) Judgements

From the discussion above I suggest that there are several factors that determine the quality of expert quantitative judgments in general, and probability assessments in particular. These include:

- The number and comparability of instances/data
- The amount and quality of feedback
- Experience with expressing judgments in the required metric (e.g. uncertainty probabilistically)
- The number of steps in the judgment process
- The nature of the judgment process (e.g. whether it is analytic or holistic, rational or intuitive, conscious or unconscious, cognitive or affective etc.)

Some conclusions for probabilistic EKE that can be drawn from this analysis are that the quality of probability judgment depends on the nature of the judgment task in interaction with the cognitive, emotional and dispositional characteristics of the experts, as well as, of course, their experience. This leads me to question the status of expert judgments as probabilities that can be meaningfully calibrated. In ideal situations (dispassionate, analytic judges, with lots of experience of both making probability judgments and receiving rapid and useable feedback, for a judgment task where there are several cues related well to the criterion for which

information is readily available) then there is at least the potential for well-calibrated judgments. Bolger and Wright's (1994) review supports this view that, under the right circumstances, expert probability judgment can be reliable. However, in most cases where we actually need to perform an EKE many of these ideal features will not be present (indeed where they are, there is a case for arguing that there is no need for expert judgment in the first place as statistical or machine learning procedures might produce better results, not least by reducing error in application of models[16]). In the least favourable conditions the likelihood judgments of experts may well be based on task-irrelevant factors such as the weather (e.g. Saunders 1993), or how hungry the expert is (Hoefling and Strack 2010) and, as such it might be counter-productive to elicit them in the first place—for instance, giving a spurious scientific veneer to the EKE—and other ways of expressing uncertainty and criteria might be better used (e.g. qualitative statements, coherence rather than calibration).

16.2.5 How Many Experts and How Many Judgements from Each?

If EKE is considered a data collection methodology then one way of increasing its reliability is to elicit estimates from as many experts as possible: aggregation over these multiple experts then helps reduce error and bias. Other ways to improve reliability is to ensure the quality of expert judgment by selection and training—or by giving the more expert greater weight in the aggregation than the less expert—and to measure expert performance reliably for these purposes generally requires each expert to provide several judgements. Is it possible to determine how many of each are required?

16.2.5.1 How Many Experts?

In a 'standard' probabilistic EKE, such as is the focus of this volume a few 'top' experts are sampled[17]: these may be the only ones available (period, or within resource limitations), or good enough to pass screening. The Sheffield Method (Chap. 3) and the Classical Method (Chap. 2) are both optimally carried out with between 5 and ten experts (e.g. Bolger et al. 2014) as there are diminishing returns

[16]For example, 'bootstrap' linear models—models derived from regressing expert judgments onto the cues that they are presumed to use—make better predictions of a criterion than do the original unaided experts (e.g. Goldberg 1970) because they apply the judgment model more consistently.

[17]In contrast to the 'crowdsourcing' approach mentioned above whereby a large number of people with little or no domain expertise are polled (see, e.g. Budescu and Chen 2015).

to accuracy improvement as more experts are added. For example, Aspinall (2010) comments regarding expert risk elicitation using the Classical Method:

> *"My experience with more than 20 panels suggests that 8–15 experts is a viable number—getting more together will not change the findings significantly, but will incur extra expense and time."* (p. 295)

However, as Meyer and Booker (1991) note:

> *"Having less than five experts reduces the chances of providing adequate diversity or information to make inferences."* (p. 88).

Recommendations regarding the optimal number of experts have largely based on practitioners' opinion rather than on the theory or evidence, however, Budescu and Chen (2015) offer an empirical analysis of the benefits of adding additional experts in relation to their weighting system. They conclude that the best performance is derived with between 3 and 16 experts, with six being optimal (however, note that this is assuming that all experts make positive contributions to the group—groups may need to be somewhat larger than six if there is some redundancy in expertise). Budescu and Chen's result support the claim of Mannes et al. (2014) that averaging over a few top experts selected on the basis of Mean Absolute Error of their judgments (i.e. an absolute measure, in contrast to Budescu and Chen's relative measure) can equal or outperform averaging over all a larger group of experts, or just using a single top-performing judge (Mannes et al. 2014).

In the case of the Sheffield Method (see Chap. 2), the experts interact face-to-face in a workshop, and this process gets more difficult to manage as the number of experts increase: with each added expert the length of discussion is potentially increased without a corresponding increase in information, while it becomes difficult to ensure that everyone's opinions are heard and discussed thoroughly. Also with behavioural aggregation the desired endpoint is usually to achieve consensus (although not *necessarily*, as some accounts suggest): clearly this will be more difficult to achieve with larger groups.

In principle, if there were sufficient resources, it would be possible to have more than one workshop: this might be a desirable strategy if there were distinct groups with expertise relevant to a particular problem but with different knowledge bases and, perhaps, technical language, who might be difficult to put together into a single workshop. An example of this is an EFSA probabilistic-EKE exercise to determine risks for European consumers from Rift Valley Fever, a disease endemic in North African cattle, sheep and goats. Relevant experts included distinct groups such as veterinary scientists, microbiologists, food transportation and preservation specialists, and those with knowledge of the illegal trade in animals across the Sahara (see Bolger et al. 2014). Another way in which a larger number of experts could be accommodated within the Sheffield Method is by designating some experts as advisors who provide the other—'judging experts'—with information and interpretations from their specialist knowledge, but who do not make judgements themselves.

In the Classical Method, experts are usually interviewed face-to-face and separately, this is to ensure that they properly understand what they are being asked to do, and to ensure that they do not attempt to 'cheat' by looking up the answers to seed questions. This resource-intensive process is the principle practical limitation to the number of experts that can be sampled, beyond identification of potential suitable experts in the first place, which in itself can be a particular problem for the Classical Method due to the need to also find seed questions to pose: this might, for instance, be particularly challenging for experts with more practical than theoretical kinds of expertise since the practical corpus of knowledge is less likely to be written down then the theoretical (Bolger and Rowe 2015a). Probably the requirement for face-to-face interviewing in the Classical Method could be relaxed with an 'honour code', or checks for cheating, so that experts could be evaluated and elicited remotely, thereby reducing the costs of using larger number of experts.

Similarly, the Delphi method can be applied remotely: in fact, in most respects it does not matter where the experts are located since there is no need for direct interaction between them. It could be useful for the facilitator to have face-to-face communication with individual experts, though, in order to assist in explaining the task, answering queries, and encouraging them to respond: any such interaction increases demands on resources but, in my experience, not severely, and the effects can be mitigated with appropriate software (e.g. for aggregation and provision of feedback, as well as management of the interactions) so hundreds of experts can potentially take part in a Delphi elicitation with only a couple of facilitators (see e.g. Rowe and Bolger 2016).

If there are many more experts than can be managed or are needed then the various instruments discussed above can be used to cut down the number to the 'best' for the purposes of the EKE. If there are many potential experts (i.e. who pass the basic suitability criteria for the elicitation) and also the possibility of usefully employing a large number of them, for instance, in a remote Delphi application, then we might consider some kind of sampling. For example, in our large-scale Delphi on future food-risk-assessment priorities we were required to sample experts broadly across the member states of the EU, and across different domains of expertise (microbiological, chemical, environmental and nutrition), and different roles (e.g. risk management vs. risk assessment). To achieve these goals, while also minimizing sampling bias, we combined probability sampling and quota sampling (e.g. randomly sampled within categories such as broad geographical regions with similar characteristics until a set of criteria were satisfied—see Rowe and Bolger 2016 for details and Lohr 2009 for a discussion of sampling methods more generally). We also over-sampled because the remote administration and requirement for repeated responses in Delphi leads to much reduced final response rates relative to methods requiring experts to physically attend meetings. Whatever the method, it is likely to be useful to have some experts in reserve in case of drop-out.

Another consideration when determining how many experts to recruit, which is related to the quota-sampling example above, is whether there is sufficient heterogeneity in the expert panel. The main reasons for including several experts in an EKE are first to broaden the information base by gathering together a number of

different perspectives on an issue, and second to reduce bias and error by averaging over differences in quality of judgment (assuming you cannot simply pick the most accurate expert). With regard to the first of these reasons, it should be clear that just adding more experts with exactly the same knowledge, and interpretation thereof, will not improve the result of the EKE one bit, although it should help with regard to the second reason. Further, many methods of mathematical aggregation assume independence between expert judgments (as is discussed by Wilson, Chap. 9), so excessive homogeneity can have an impact on the accuracy of judgments, unless the dependencies are accounted for.

As I mentioned above, some selection methods have a tendency to introduce homogeneity into groups. For example, selecting experts using social criteria such as a senior position within an organization, or on the basis of number of publications, can introduce age and gender biases (and also a bias towards academics in the second instance). Snowballing can also encourage homogeneity as the invitation to participate will tend to be passed around within a particular social network, rather than between (although contacts can be asked to try and circulate to out-groups, including those they know to hold different opinions to themselves). However, since true random sampling of experts will rarely be a realistic prospect it may be necessary to artificially introduce heterogeneity of opinion into panels by use of methods such as 'devil's advocacy' and 'dialectical inquiry' (e.g., Bolger and Wright 2011). On the other hand, if the heterogeneity is too great, it can be difficult to reach a consensus, and aggregation may not make sense.

16.2.5.2 How Many Judgements?

The answer to this question depends on several factors and what you are trying to do. For example, if you are trying to use psychometric tests to measure an aspect of expertise that has been shown in the past to be well measured by a handful of items then you will only need an expert to answer these few items. Commonly reliability of test items is indicated by values of Cronbach's alpha of at least 0.7, indicating that there is a high average correlation between items that purportedly measure the same construct. Meanwhile predictive validity of test items is indicated by high R-squared values (typically around 35% of variance accounted for in tests of job performance, comparing to 16% for personality questionnaires and 14% for semi-structured interviews), indicating that the items collectively predict well the target variable, e.g. accuracy of the expert judgements. However, the application of psychometrics to the selection of experts is in its infancy and thus the reliability and predictive validity of instruments such as GEM or the E-SQ are not known. Until such time that they are, reliability and predictive validity might be improved by using multiple test instruments (i.e. increasing the number of questions each expert has to answer) but the cost of this strategy is losing potential experts who may see themselves as too busy (or important) to complete lots of tests.

When attempting to assess expertise by examining the accuracy or calibration of judgements then there are costs to asking for additional judgments beyond simply deterring potential contributors to the EKE. As already discussed, finding seed questions that are closely related to the target, have known answers, yet are not too easy or too hard for the experts is challenging. For this reason studies using the Classical Method tend to use rather few seed questions: modally only ten per study in the 45 applications reported by Cooke and Goossens (2008). There is some debate as to whether this is a sufficient number. We (Bolger and Rowe 2015a) calculated that ten seed variables have insufficient power to detect all but the most poorly calibrated experts with at least 60 judgments being a more reasonable number to aim for. However, Quigley et al. (Chap. 2) show in a Monte Carlo simulation that if we simply wish to discriminate between experts in terms of calibration—rather than test the hypothesis that they are different from a well-calibrated expert—then ten seeds is sufficient to achieve good discrimination.

I argue that further research is still needed to establish whether ten seeds is a sufficient number for reliable assessment of expert performance across a range of judgment tasks (e.g. varying in difficulty and heterogeneity of items). Also even ten questions—with three or more probability judgments per question (to elicit a distribution)—can prove burdensome to experts, so, as with psychometric testing, incentives may be needed (if not financial, then anonymized feedback of performance relative to others might suffice). Finally, it is important that judges are encouraged to provide true expressions of belief so if any incentives are given they should be in accord with a 'proper scoring rule' that is designed to do just this, and that is also understandable to the experts (see Bolger and Rowe 2015a for a discussion of proper scoring rules in relation to the Classical Method).

The number of experts and judgments has implications for the choice of elicitation method. If there are numerous experts but relatively few judgments, this might favour the use of methods that can easily be run remotely such as IDEA—which has been carried-out using video or phone conferencing—or Delphi, especially if the experts are widely geographically distributed and/or resources for conveying them to an elicitation venue are limited. In contrast, if there are only a few experts, but many judgments are to be made, a behavioural aggregation method such as Sheffield, could be indicated. The Classical Method could potentially be used well in either case—the main constraint here being the availability of seed questions (however, if there are several judgments to be made regarding rather disparate types of target variables then this could create additional problems for finding seed variables in sufficient numbers as well as increasing the load on the experts who might have to answer scores of seeds as well as making numerous target judgments).

16.3 Part II: A Structured Approach to the Selection of Experts

I propose a two-stage expert recruitment process that is designed to maximize the quality of experts in an EKE. The first recruitment stage occurs prior to determination of elicitation method and thus is focussed on identifying those experts who may have knowledge relevant to assessing the required target quantities. I propose that this is achieved by means of an instrument referred to as a *profile matrix* that defines both the type of knowledge required and the roles of experts who might have such knowledge. Once this profile matrix has been used—in conjunction with relevant databases and snowballing—to construct a *long list* of as many potential experts as is practical within budget and time constraints, then the elicitation method can be chosen. Both the profile matrix and the long list are needed to select the method because different numbers, types and mixes of experts have implications for the best method to be used.

In the second recruitment stage, the number, type and mix of experts might be reappraised—possibly requiring a return to the first stage—and, if there are more experts on the long list than is required by the method, *screening* might be applied to create a *short list* of those experts thought to best fit the profile matrix. Screening might also be used to achieve representation of various opinions, roles and so on, or to create heterogeneous groups. I suggest a template for a *questionnaire* (the E-SQ) that can be used for screening purposes, and can also be used to assess any training needs: in particular, experts may need training in expressing uncertainty in the form of probabilities. Other measures of expertise discussed in Part I can also be used either independently or in conjunction with the E-SQ at this stage, both for screening and *weighting* of experts, if the EKE method requires weighting. Once the short-list is created experts can be invited, trained and elicited: steps need also to be taken during this recruitment stage to ensure the retention of experts throughout the elicitation process (which can be lengthy, depending on the method) and the quality of their contributions.

16.3.1 The EKE Process

It is important to locate the two-stage process of the selection and recruitment of experts described above within the broader context of the EKE process as a whole, and the way such a process is managed. In a report for the European Food Safety Authority (EFSA: Bolger et al. 2014) propose that once a problem has been identified a Working Group (WG) is set up by the stake holder. The WG consists of

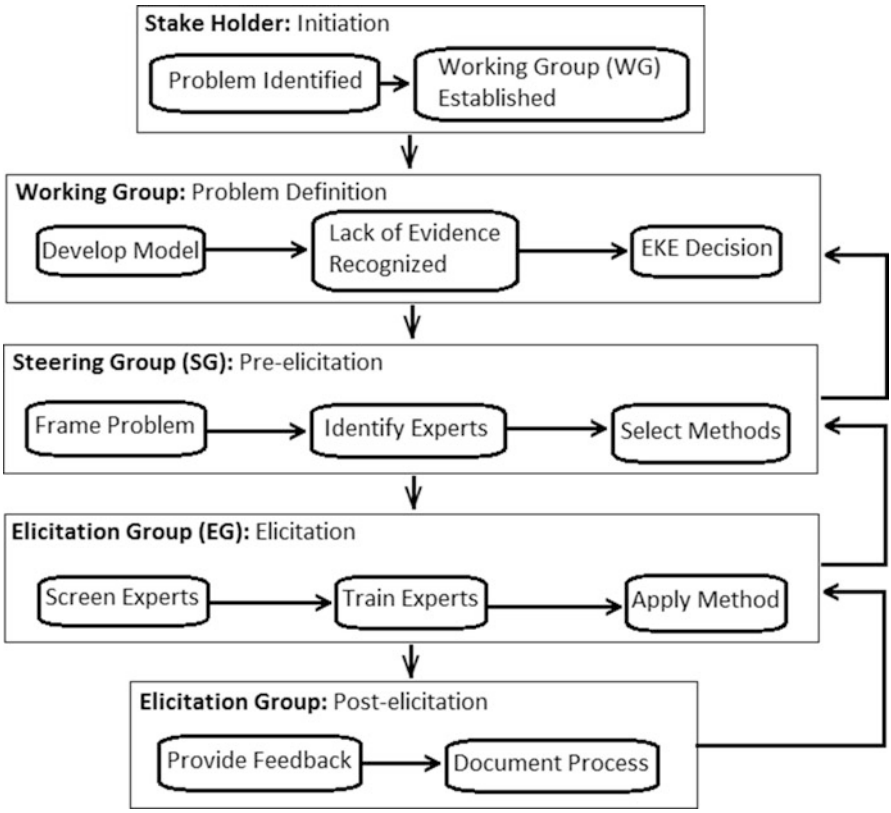


Fig. 16.1 A possible EKE process (adapted from Bolger et al. 2014)

a set of experts [18] in the recognized problem area and is responsible for determining the model and instigating the initial search for data. If a lack of 'hard' data is identified then the WG makes a decision to conduct EKE and select a new group, the Steering Group (SG), to manage the EKE process: The SG may be a subset of the WG and/or comprise new members, perhaps with some knowledge of EKE, and is responsible for analysing the problem of concern to a sufficient degree to allow them to identify appropriate domain experts and select an appropriate EKE method. The SG is also responsible for selecting a third group—the Elicitation Group (EG)—to conduct the elicitation. The EG will consist of experts in the chosen elicitation method and probably one or more members of the SG/WG with expertise in the target problem. The EG will be responsible for managing the elicitation and post-elicitation phases which involve not only administering the EKE method but also training experts and evaluating and documenting the process. This entire

[18]These experts will be referred to later as 'super-experts' since they have an overview of the problem as a whole and are responsible for recruitment, selection and management of any other experts used in the process.

process is illustrated in Fig. 16.1: the arrows pointing to the right and downwards represent progression through the process and the arrows pointing up and to the left feedback and communication: the latter can lead to changes, for example, in composition of the groups, model specification, criteria for expert selection, and choice of method.

A scheme similar to that proposed in Fig. 16.1 is being applied to risk assessment projects in EFSA but at the time of writing an evaluation of its effectiveness is not yet available.

16.3.2 From Problem Identification to Long-Listing (Stage 1)

In order to identify required expert knowledge, Bolger et al. (2014) recommend the construction of an expert profile matrix. This is a table of essential and desirable expert characteristics based on definitions and types of expertise described in Part I and relevant 'roles' (e.g. the sort of job roles where required expertise might be expected to reside—these roles help in the search for suitable experts but should not be overly constraining i.e. attempts to look beyond these roles should ideally be made). To illustrate what an expert profile matrix might look like we present the following real example of a risk assessment exercise where a probabilistic EKE was performed (EFSA 2013).

> "To assess the risk of introduction of Rift Valley fever virus (RVFV) through the movement of RVFV-infected (previraemic and viraemic) animals into designated countries of North Africa and the Near East, also referred to . . . as the Region Concerned (RC)."

Table 16.1 shows an example profile matrix that could be applied to this problem (it is not the one actually used as the concept of the expert profile matrix was still under development at that time). The table should be read from left to right. Thus there are three kinds of substantive expertise that it is essential should be represented on the expert panel (epidemiological and/or virological, disease surveillance, and risk assessment) and two other kinds of expertise that we would ideally like represented on the panel (concerned with livestock transport and food safety). Normative expertise, in this case knowledge of probability, was only deemed essential for the risk assessment experts and was considered not applicable to livestock transport experts. Specific expertise with regard to RVFV is only ever considered desirable largely because of the few experts who have this specialism, however, where possible experts from the RC were to be selected, partly for political reasons, and partly for their specialist on-the-ground knowledge. An international perspective was also considered desirable in most cases and essential in the case of risk assessment and food safety issues. Finally, experts were sought from different countries within the RC and in different roles: again partly for political reasons and partly to introduce heterogeneity of perspectives on the problem.

The matrix suggests the kind of experts to search for, and the skills that they should have, but it usually will not be necessary to have a representative for each cell in the matrix (which is just as well because finding sufficient numbers of experts

Table 16.1 Example of a profile matrix

Knowledge requirements				Expert roles			
Substantive expertise (Importance)	Normative expertise (Importance)	Substantive specificity (Importance)	Regional specificity (Importance)	Country	Industry/NGO	Government	Academia
Epidemiology and/or virology (Essential)	Desirable	RVFV (Desirable)	RC (Essential)	AA			
				BB			
			International (Desirable)				
		General (Essential)	RC (Essential)	AA			
				BB			
			International (Desirable)				
Disease surveillance (Essential)	Desirable	RVFV (Desirable)	RC (Essential)	AA			
				BB			
			International (Desirable)				
		General (Essential)	RC (Essential)	AA			
				BB			
			International (Desirable)	BB			
Risk assessment (Essential)	Essential	RVFV (Desirable)	RC (Desirable)	AA			
				BB			
			International (Essential)				
		General (Essential)	RC (Desirable)	AA			
				BB			
			International (Essential)				
Livestock transport (Desirable)	N/A	N/A	RC (Essential)	AA			
				BB			
			International (Desirable)				
Food safety (Desirable)	Desirable	N/A	RC (Desirable)	AA			
				BB			
			International (Essential)				

is often difficult). For instance, in the example in Table 16.1 there may not need to be an expert of every role from every country, but merely a selection across both dimensions (but see Sect. 16.2.5.1 above regarding quota sampling). Similarly a balance of specialists and generalists might be sufficient rather than one of each for each kind of substantive expertise, for which expert skills that are merely desirable might be dispensed with (or perhaps provided through training). In the Rift Valley case 18 experts attended Sheffield elicitations and this was deemed sufficient—most experts could cover several cells of the matrix—however, it was noted that no livestock transport experts were invited and that that was a potential limitation (EFSA 2013).

Once a matrix, or some other list of criteria, has been created a long-list of potential experts can start to be compiled on the basis of, for instance, SG and WG suggestions, the sponsor's and others' databases, and social/professional networks. The long-list can be further augmented via snowballing, in other words, asking experts who have already been identified to nominate other potential experts etcetera. It may be noted that snowballing can lead to biasing the sample of experts towards particular interest groups and networks relative to others so, as a recruitment technique, it should be used with open eyes (but the same can be said of any expert recruitment method since it is not usually possible to sample randomly). The GEM questionnaire and/or other peer assessment instruments could be sent out during the snowballing process.

Some further issues about approaching experts for potential recruitment include: who should approach prospective experts (e.g. someone known and trusted might be more persuasive than a stranger, and someone with status might be more persuasive than someone without); how approaches should best be made (e.g. a phone call might be more effective than an e-mail); and how experts should be compensated (e.g. while some financial reward may be a useful incentive to participation many experts will have intrinsic motivation to take part and this can be 'crowded-out' by large financial rewards).

As much information about experts should be collected as is necessary to determine their fit to the Expert Profile Matrix. Some information may be available in advance of approaching the experts, for instance, if an expertise database is held. Other information, such as a CV can be requested when experts are first contacted or, perhaps better, after they have indicated an interest in participation. As already discussed, CV's contain useful information for profiling experts, but may not hold all the information required. It may therefore be necessary to construct a special questionnaire regarding the particular abilities of the prospective experts relative to the foresight task in question.

Bolger et al. (2014) present an example questionnaire for use in identifying particular expert skills, which I already introduced above as the E-SQ: some example questions are shown in Table 16.2.[19] This questionnaire is designed find

[19]This questionnaire was an adaptation of the one first developed by Wright et al. (2004), based on earlier work by Bolger and Wright (1994) and Rowe and Wright (2001): the example questions presented in Table 16.2 have been further edited by the current author since the 2014 version.

Table 16.2 Example items from the E-SQ

Part A. The nature of the expert's job
• What is the title of your job?
• How would you describe your area of expertise?
• How many years of experience would you say you had in your area of expertise?
• Would you describe that experience to be practical and/or field-based vs. theoretical and/or lab-based?
Part B. The type of judgments the expert makes while performing his/her job and what help, if any, s/he receives in making these judgements
• Describe the most important judgements that you make on a regular basis in your job.
• When you have to make work judgements, to what extent do you rely on your judgement alone, and to what extent do you rely on other information sources (such as manuals of statistics, computer databases or programs, etc.)?
• Describe any other information sources you use.
Part C. Whether the job requires an expert to make probabilistic judgments and how such judgments are made
• Considering the uncertainty you assess at work, do you ever make any of the following types of judgments (I estimate the likelihood/probability of . . . , I estimate the chances of . . . , I estimate confidence in . . .)?
• How often, on average, are you called upon to make risk judgments of these types?
• When you make uncertainty judgments, what forms do they take? For example:
- Numerical estimates (e.g. 0.5, 50%, 1 in 2)
- Verbal estimates (e.g. likely, infrequent)
- Comparative (e.g. 'this likelihood is similar to another likelihood')?
• If you make numerical estimates of uncertainty, what form do they take? For example:
- Percentages (e.g. 50% chance) ; Point probabilities (e.g. 0.5 chance)
- Confidence intervals (e.g. range within which you are 95% confident the true value falls)
- Probability distributions (as previous but more than one range assessed for each quantity)
- Frequencies (e.g. 3 out of 10 chances of occurring); Odds (e.g. odds of 2 to 1 against it occurring)
- Ratings on scales (e.g. point 2 on a 7-point scale of likelihood)
- Other type of numerical judgement: please provide details
Part D. The nature of data and models used to make judgements, whether any feedback is received about the quality of judgements, and whether any training has been received
• In making your work judgements, do you receive any feedback about their accuracy?
• If you receive some feedback, what form does this take?
• How soon after a judgement, on average, do you receive feedback?
• How would you rate the ease of making good judgements in your work?
• Do you make use of a formal model for making your work judgements?
• How would you rate the availability of data that you use for your work judgements?
• How would you rate the quality of data that you use for your work judgements?
• Did you receive any training to make judgements? If so please describe.

out about the nature of an expert's job (Part A), and the type of judgments that he or she makes while performing it and what help, if any, the expert receives in making these judgements (Part B). In particular, we are interested in whether or not the job requires an expert to make probabilistic judgments and how such judgments are made (Part C). Finally, we are interested to find out what sort of data and models

are used in making judgements whether any feedback is received about the quality of judgements, and whether any training in making judgements has been received (Part D). The questionnaire is inspired by Bolger and Wright's (1994) analysis in terms of learnability and ecological validity of tasks that we discussed above in Sect. 16.2.4.2. For example, match between the judgments normally made by the expert and those s/he will have to answer in the EKE can be determined from parts A to C and possibilities for learning to make (probability) judgements from part D. Information about training received given in Part D can be useful in determining the training requirements for the experts prior to the elicitation exercise.

The E-SQ can be sent out either as expert names come available, in which case the responses can be used to assist long-listing (by comparison with the Profile Matrix), or can be distributed after a long list has been constructed. Once responses are received they can not only be used for screening to produce a short list, and assessing training needs, as I will describe in the next section, but also can be helpful for determining the elicitation protocol to be used. For example, the responses to the questionnaire could help to determine whether the nature of the potential panellists' expert practice is more qualitative, thus perhaps favouring behavioural aggregation methods where there is 'discussion' between the experts (e.g. Sheffield,), or more quantitative, thus favouring mathematical aggregation approaches (e.g. the Classical Method), or a mixture of quantitative and qualitative, thus favouring mixed aggregation methods (e.g. IDEA, or Delphi with exchange of rationales). E-SQ responses could also indicate the potential availability of seed questions for the Classical Method.

16.3.3 From Short-Listing to Wrap-Up (Stage 2)

16.3.3.1 Screening, Short-Listing, and Weighting

Depending on the choice of elicitation method it may be necessary to short-list experts from the long-list created at Stage 1. For example, as I have already noted above, behavioural aggregation methods usually require bringing experts together physically to a workshop where experts discuss the topic at hand with each other, and make their judgments, under the supervision of the elicitor. In the Classical Method experts are usually asked to answer the seed questions individually and again under the direct supervision of the elicitor. For such methods it may therefore be too difficult or expensive to have more than a few experts. As I also noted in Sect. 16.2.5 of Part I, there may be diminishing returns to having more than a few experts. In the unusual situation where there are more experts on your long-list who are willing to take part in the elicitation exercise than you need, those experts with the most, and most relevant, expertise as indicated by responses to the questionnaire could be put onto a short-list.

The E-SQ described above can be used for screening here, for instance, by comparing responses to the essential and desirable features of the Expert Profile Matrix: this is equivalent to common procedures for short-listing job applicants.[20] Some other EKE methods, such as the Classical Method, require weighting the experts in terms of their performance on similar judgment problems ('seed questions'). Although in the Classical Method less capable experts are included in the elicitation but given lower (or possibly, no) weight, such specially designed tests could also be used to determine who is, or is not, short-listed.

Information from CV's and other sources can similarly be used for screening but if the intention is to weight experts on the basis of their perceived ability (other than the crude weighting of accept or reject for the exercise) it is probably necessary to ask additional questions to those in the E-SQ[21] and to supplement the data normally found on CV's. Potential questions include specific ones relating to required aspects of substantive or normative expertise: these are akin to the seed questions of the Classical Method. Any questions designed to test substantive aspects of expertise will presumably have to be posed by super-experts (probably from the WG). Peer or self-assessments, or simple metrics such as years of experience in role or number of publications and their impact (e.g. number of citations) could also be used, subject to the caveats raised in Part I. Weighting of experts who have passed screening, and are therefore participants in the elicitation exercise, is relevant if there are several experts who may differ in judgment accuracy (defined variously from simple hit-rate or judgment error to the realism of uncertainty assessments).

16.3.3.2 Training, Retention and Documentation

In the course of expert screening training needs may be identified thus this is a good place to briefly discuss training (but only at a general level because details of training requirements, delivery and content are specific to each particular protocol).

Training will most usually be related to expression of uncertainty (i.e. normative rather than substantive aspect of expert judgment): the E-SQ has questions for this purpose. Another approach is to give *all* experts normative training as part of induction, perhaps using distance-learning materials such as a video presentation or an e-learning package. Although not usual, training could also be given in substantive expertise. For example, if forecasting would benefit from expertise from several different specialisms then some training, for instance in terminology and basic concepts, could be given across specialisms so as to assist communication (i.e. knowledge exchange between experts). Such training in substantive issues might

[20] However, in some other important respects, expert selection is not like job selection i.e. you want to find people with the right skills rather than reject people with the wrong skills . . . so it is more akin to head hunting.

[21] Potentially, scores for weighting could be derived from this questionnaire but this has not been attempted yet, let alone any validation of weights thus derived, therefore this is something for future research.

most easily be accomplished face-to-face in a facilitated workshop, but could also conceivably be achieved online, for instance within a Delphi process.

Training in normative aspects is most usually given by the elicitors whereas training in substantive aspects would most appropriately be given by domain experts, perhaps the super-experts who have a broad view of the problem: as noted earlier, super-experts are often those in the Working Group who identified the need for EKE in the first place and are responsible for initial problem formulation, and developing the Expert Profile Matrix and Expert-Selection Questionnaire. If training is done remotely then there is more flexibility regarding who it is performed by (and when it is done; i.e. if it is face-to-face then it is likely that you will want to do it at the same time as the elicitation exercise itself).

As should by now be clear, experts are a valuable resource so you want to make sure that they complete the EKE (and complete it to the best of their ability; i.e. have not only appropriate training, as just discussed, but also sufficient motivation to try hard to perform well at what are often difficult tasks): depending on the protocol, some EKE's may take many months to complete. It may also be the case that you wish to re-use experts for follow up probes or future complete EKE exercises. How to motivate and retain experts are therefore important considerations when conducting an EKE.

I have already touched upon some aspects of motivating experts, such as appropriate payment and proper-scoring rules that reward truthful answers (which could be the basis for performance-related payment), however, in my experience, the key to both motivation and retention is the maintenance of good—accurate, speedy, friendly yet respectful—channels of communication between those conducting the EKE and the experts. Intrinsic motivation can be kept at high levels if experts feel involved and valued, and this can best be achieved by 'keeping them in the loop'. Thus any queries from experts should be answered rapidly (and without making them appear that they are being a nuisance or ignorant, even if they are) and feedback should be given about their contributions in particular, and the outcomes of the EKE more generally, as soon as they become available. This may require some management of expectations if the results may be limited in some way (e.g. due to confidentiality issues) or, as is often the case, a long time in gestation.

This brings me to what is usually the final step in an EKE: documentation. Documentation may be produced for many audiences (and purposes) including the consumers (e.g. for decision and policy making), the elicitors (e.g. to inform future elicitations), the commissioning organization (who may not necessarily be either the consumers or the elicitors—also to inform future elicitations), and the general public (e.g. to justify spending tax-payer money). The experts are another audience and should receive appropriate documentation in a timely measure with acknowledgment for their contributions if desired (some experts may prefer anonymity so agreement over the level of disclosure must be agreed in advance of publication).

16.4 Conclusions

We have seen that expert judgments of uncertain quantities are frequently needed to complete risk and forecast models used to inform decision and policy making. While an increasing amount of research is being devoted to developing protocols for eliciting such judgments in an unbiased manner as possible, relatively little attention has been paid to the question of finding and managing the experts to be used in an elicitation. In this chapter I have reviewed the 'state-of-the-art' with regard to identifying, measuring and cultivating expertise and attempted to locate this work within a framework that conceptualizes expert elicitation as a social-science methodology that has at its heart the goal of producing data of the highest possible quality in terms of its reliability and validity. To this end, I first explored the nature of expertise which, through a classification of types of expertise—along social, epistemic, psychological, and normative dimensions—produced a number of different indicators that can be ranked in terms of their capacity to differentiate those with real from those with apparent ability. Beyond indicators, the classifications also give rise to extant and potential measures of expertise, which I argued are likely to vary in terms of their reliability and validity.

Further research is needed to establish exactly how these indicators and measures stand against each other, and what their relative strengths and weaknesses are in terms of their utility for optimizing expertise in EKE. More generally there is a need for theoretically grounded and empirically based research to answer a number of questions relating to the use of experts in EKE, including: should we use expertise measures just for selection or also to weight the experts? how many experts should we use in an elicitation? and how many judgments? and if we use seed variables, how many of them? and what are the upper limits to expert performance?

As a starting point for answering such questions I posed another question: how might an intelligent system go about making judgments of uncertain quantities? Once you start to explore *this* question one discovers that there are potentially complex interactions between characteristics of the intelligent system on the one hand and the features of judgment tasks (and elicitation protocols) on the other: these interactions mean that answers to the superficially straightforward questions in the previous paragraph become rather nuanced, depending on the personality and past and current experience of the experts (including their ongoing affective states, and training/instructions they have received), the judgment task (e.g. its familiarity and similarity to those faced professionally), and the characteristics of the protocol (e.g. do experts make their judgments independently or in direct discussion with others?).

So, to conclude, there is still much work to be done but through careful analysis of characteristics of experts and the judgment tasks they perform (in the context of EKE) I believe we can move forward so as to develop both new instruments for identifying and measuring expertise, and improved systems for selecting and managing experts in EKE: 'meta-protocols' if you like. In Part II of this chapter I took a few tentative steps in this direction, outlining an example template for

conducting EKE's in an organization, and two new instruments—the Expert Profile Matrix and the Expert-Skills Questionnaire—which, although subject to evaluation, may assist in the long-term goal of improving the quality of expert judgment input to forecasting and risk models.

Acknowledgements I wish to acknowledge the support of the European Food Safety Authority (EFSA) who convened the "Working Group on Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment": the research conducted for which laid the foundations for this chapter. I also wish to thank Working Group members, Anca Hanea, Anthony O'Hagan, Jeremy Oakley, Gene Rowe and Meike Wentholt; and EFSA staff, Elisa Aiassa, Fulvio Barizzone, Eugen Christoph, Andrea Gervelmeyer, Olaf Mosbach-Schulz, and Sara Tramontini for their contributions to the original research. Finally, I am grateful to David Budescu, John Quigley, and Gene Rowe for their comments on earlier drafts of this manuscript.

References

Alvarado-Valencia J, Barrero LH, Onkal D, Dennerlein JT (2017) Expertise, credibility of systems forecasts and integration of methods in judgmental demand forecasting. Int J Forecast 33(1):298–313

Anderson JR (1982) Acquisition of cognitive skill. Psychol Rev 89:369–406

Aspinall W (2010) A route to more tractable expert advice. Nature 463:294–295

Bargh JA (1994) The four horsemen of automaticity: intention, awareness, efficiency, and control as separate issues. In: Wyer R, Srull T (eds) Handbook of social cognition. Lawrence Erlbaum, Hillsdale, NJ, pp 1–40

Barron G, Erev I (2003) Small feedback-based decisions and their limited correspondence to description-based decisions. J Behav Decis Mak 16(3):215–233

Bayindir M, Bolger F, Say B (2017) An investigation of the role of some person and situation variables in multiple cue probability learning. Q J Exp Psychol 70(1):36–52

Bazerman MH, Moore DA (2008) Judgment in managerial decision making, 7th edn. Wiley, New York

Bolger F, Rowe G (2015a) The aggregation of expert judgment: do good things come to those who weight? Risk Anal 35(1):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 Anal 35(1):21–26

Bolger F, Wright G (1993) Coherence and calibration in expert probability judgement. OMEGA Int J Manag Sci 21:629–644

Bolger F, Wright G (1994) Assessing the quality of expert judgment: issues and analysis. Decis Support Syst 11:1–24

Bolger F, Wright G (2011) Improving the Delphi process: lessons from social psychological research. Technol Forecast Soc Chang 78:1500–1513

Bolger F, Wright G (2017) Use of expert knowledge to anticipate the future: issues, analysis and directions. Int J Forecast 33(1):230–243

Bolger F, Wright G, Rowe G, Gammack J, Wood R (1989) LUST for life: developing expert systems for life assurance underwriting. In: Shadbolt N (ed) Research and development in expert systems VI. Cambridge University Press, Cambridge, pp 128–139

Bolger F, Hanea A, Mosbach-Schulz O, Oakley J, O'Hagan A, Rowe G, Wenholt M (2014) Guidance on expert knowledge elicitation in food and feed safety risk assessment. European Food Safety Authority (EFSA), Parma

Brenner LA (2003) A random support model of the calibration of subjective probabilities. Organ Behav Hum Decis Process 90(1):87–110

Brunswik E (1955) Representative design and probabilistic theory in a functional psychology. Psychol Rev 62(3):193
Budescu DV, Chen E (2015) Identifying expertise to extract the wisdom of crowds. Manag Sci 61(2):267–280
Burgman MA, McBride M, Ashton R, Speirs-Bridge A, Flander L et al (2011) Expert status and performance. PLoS One 6(7):E22998
Chase WG, Simon HA (1973) Perception in chess. Cogn Psychol 4(1):55–81
Clemen RT (2008) Comment on Cooke's classical method. Reliab Eng Syst Saf 93:760–765
Cooke RM (2014) Validating expert judgments with the classical model. In: Martini C, Boumans M (eds) Experts and consensus in social science—critical perspectives from economics, sociology, politics, and philosophy. Ethical economy—studies in economic ethics and philosophy. Springer, Heidelberg, pp 191–121
Cooke RM, Goossens LLHJ (2008) TU Delft expert judgment database. Reliab Eng Syst Saf 93:657–674
Cooke RM, ElSaadany S, Huang X (2008) On the performance of social network and likelihood-based expert weighting schemes. Reliab Eng Syst Saf 93(5):745–756
Dalgleish LI (1988) Decision making in child abuse cases: Applications of social judgment theory and signal detection theory. In: Brehmer B, Joyce CB (eds) Human judgment: the SJT view. Elsevier, Amsterdam, pp 317–360
Dalkey NC (1975) Toward a theory of group estimation. In: Linstone HA, Turoff M (eds) The Delphi method: techniques and applications. Addison-Wesley, Reading, MA, pp 236–261
De Bondt WF, Thaler RH (1989) Anomalies: a mean-reverting walk down Wall Street. J Econ Perspect 3(1):189–202
De Groot AD (1965) Thought and choice in chess. The Hague, Mouton
Dhami MK, Wallsten TS (2005) Interpersonal comparison of subjective probabilities: toward translating linguistic probabilities. Mem Cogn 33(6):1057–1068
Dror IE (2011) The paradox of human expertise: why experts get it wrong. In: Kapur N (ed) The paradoxical brain. Cambridge University Press, Cambridge, pp 177–188
Eden C (1988) Cognitive mapping. Eur J Oper Res 36(1):1–13
Edwards W (1968) Conservatism in human information processing. In: Kleinmuntz B (ed) Formal representation of human judgment, CMU cognition series, vol 3. Wiley, New York, pp 17–51
Eggstaff JW, Mazzuchi TA, Sarkani S (2014) The effect of the number of seed variables on the performance of Cooke's classical model. Reliab Eng Syst Saf 121:72–82
Ellis HC, Ashbrook PW (1989) The state of mood and memory research: a selective review. J Soc Behav Pers 4(2):1–21
European Food Safety Authority (2013) Technical meeting of the EFSA Scientific Network on EFSA Scientific. Network for risk assessment in Animal Health and Welfare – Risk of introduction of Rift Valley fever into the Southern Mediterranean area through undocumented movement of infected animals. EFSA Supporting Documents, 10 (4)
Fama EF (1965) The behaviour of stock market prices. J Bus 38:34–105
Ferrell WR, McGoey PJ (1980) A model of calibration for subjective probabilities. Organ Behav Hum Perform 26(1):32–53
Frederick S (2005) Cognitive reflection and decision making. J Econ Perspect 19(4):25–42
Genre V, Kenny G, Meyler A, Timmermann A (2013) Combining expert forecasts: can anything beat the simple average? Int J Forecast 29:108–121
Germain ML (2006) Development and preliminary validation of a psychometric measure of expertise: the generalized expertise measure (GEM). Unpublished doctoral dissertation. Barry University, Florida, USA
Germain M-L, Tejeda MJ (2012) A preliminary exploration on the measurement of expertise: an initial development of a psychometric scale. Hum Resour Dev Q 23(2):203–232
Gibbons AM, Sniezek JA, Dalal RS (2003) Antecedents and consequences of unsolicited versus explicitly solicited advice. In: Budescu D (Chair), Symposium in Honor of Janet Sniezek. Symposium presented at the annual meeting of the Society for Judgment and Decision Making, Vancouver, BC (2003 November)

Gigerenzer G, Hoffrage U (1995) How to improve Bayesian reasoning without instruction: frequency formats. Psychol Rev 102(4):684
Gigerenzer G, Hoffrage U, Kleinbölting H (1991) Probabilistic mental models: a Brunswikian theory of confidence. Psychol Rev 98(4):506–528
Goldberg LR (1970) Man versus model of man: a rationale, plus some evidence, for a method of improving on clinical inferences. Psychol Bull 73(6):422
Griffin D, Brenner L (2004) Perspectives on probability judgment calibration. In: Koehler DJ, Harvey N (eds) Blackwell handbook of judgment and decision making. Blackwell, Malden (MA), pp 177–199
Griffin D, Tversky A (1992) The weighing of evidence and the determinants of confidence. Cogn Psychol 24(3):411–435
Hardman DK (2009) Judgment and decision making: psychological perspectives. Wiley, Chichester
Hayes-Roth F, Waterman D, Lenat D (1983) Building expert systems. Addison-Wesley, Boston
Hertwig R, Barron G, Weber EU, Erev I (2004) Decisions from experience and the effect of rare events in risky choice. Psychol Sci 15(8):534–539
Hodge R, Evans M, Marshall J, Quigley J, Walls L (2001) Eliciting engineering knowledge about reliability during design-lessons learnt from implementation. Qual Reliab Eng Int 17(3):169–179
Hoefling A, Strack F (2010) Hunger induced changes in food choice. When beggars cannot be choosers even if they are allowed to choose. Appetite 54(3):603–606
Howard RA, Matheson JE (2005) Influence diagrams. Decis Anal 3:127–143
Isen AM, Erez A (2002) The influence of positive affect on the components of expectancy motivation. J Appl Psychol 87(6):1055–1067
Ivlev I, Kneppo P, Bartak M (2015) Method for selecting expert groups and determining the importance of experts' judgments for the purpose of managerial decision-making tasks in health systems. Ekon Manag 2(18):57–72
Janis IL (1982) Groupthink: psychological studies of policy decisions and fiascoes. Houghton Mifflin, Boston
Juslin P (1993) An explanation of the hard-easy effect in studies of realism of confidence in one's general knowledge. Eur J Cogn Psychol 5(1):55–71
Juslin P (1994) The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. Organ Behav Hum Decis Process 57(2):226–246
Keren G (1991) Calibration and probability judgements: conceptual and methodological issues. Acta Psychol 77(3):217–273
Klahr D, Langley P, Neches R (1987) Production system models of learning and development. MIT Press, Cambridge, MA
Klein G (1998) Sources of power: how people make decisions. MIT Press, Cambridge, MA
Kleindorfer PR, Kunreuther H, Schoemaker PJ (1993) Decision sciences: an integrative perspective. Cambridge University Press, Cambridge, p 484
Kliger D, Levy O (2003) Mood and judgment of subjective probabilities: evidence from the US index option market. Eur Finan Rev 7(2):235–248
Koehler D, Brenner L, Griffin D (2002) The calibration of expert judgment: heuristics and biases beyond the laboratory. In: Gilovich T, Griffin D, Kahneman D (eds) Heuristics and biases: the psychology of intuitive judgment. Cambridge University Press, Cambridge, pp 686–715
Koriat A, Lichtenstein S, Fischhoff B (1980) Reasons for confidence. J Exp Psychol Hum Learn Mem 6(2):107–118
Lave J, Wenger E (1991) Situated learning: legitimate peripheral participation. Cambridge University Press, Cambridge
Lichtenstein S, Fischhoff B, Phillips L (1982) Calibration of probabilities: the state of the art to 1980. In: Kahneman D, Slovic P, Tversky A (eds) Judgment under uncertainty: heuristics and biases. Cambridge University Press, Cambridge, pp 306–334
Lin SW, Bier VM (2008) A study of expert overconfidence. Reliab Eng Syst Saf 93:711–721

Lin SW, Cheng CH (2009) The reliability of aggregated probability judgments obtained through Cooke's classical method. J Model Manag 4:149–161
Lohr S (2009) Sampling: design and analysis. Nelson Education, Toronto
MacGregor DG, Lichtenstein S (1991) Problem structuring aids for quantitative estimation. J Behav Decis Mak 4:101–116
MacGregor DG, Lichtenstein S, Slovic P (1988) Structuring knowledge retrieval: an analysis of decomposed quantitative judgments. Organ Behav Hum Decis Process 42:303–323
MacIntyre A (2007) After virtue: a study in moral theory, 3rd edn. Duckworth, London
Malkiel BG (2011, October) The efficient-market hypothesis and the financial crisis. In: Rethinking finance: perspectives on the crisis (Proceedings of a conference). Russel Sage Foundation
Mannes AE, Soll JB, Larrick RP (2014) The wisdom of select crowds. J Pers Soc Psychol 107(2):76–299
Meyer MA, Booker JM (1991) Eliciting and analyzing expert judgment: a practical guide. Academic, London
Moore G, Beadle R (2006) In search of organizational virtue in business: agents, goods, practices, institutions and environments. Organ Stud 27:369–389
Moreira C, Wichert A (2013) Finding academic experts on a multisensor approach using Shannon's entropy. Expert Syst Appl 40:5740–5754
Murphy AH, Brown BG (1985) A comparative evaluation of objective and subjective weather forecasts in the United States. In: Wright G (ed) Behavioral decision making. Plenum, New York, pp 329–359
Newell A, Simon HA (1972) Human problem solving. Prentice-Hall, Englewood Cliffs, NJ
Newell A, Shaw JC, Simon HA (1959) The processes of creative thinking. Rand Corporation, Santa Monica, CA
Nisbett RE, Wilson TD (1977) Telling more than we can know: verbal reports on mental processes. Psychol Rev 84(3):231–259
Oliver RM, Smith JQ (1990) Influence diagrams, belief nets and decision analysis. Wiley, New York
Olsson H (2014) Measuring overconfidence: methodological problems and statistical artifacts. J Bus Res 67:1766–1770
Oskamp S (1965) Overconfidence in case study judgments. J Consult Psychol 29:261–265
Parenté FJ, Anderson-Parenté JK (1987) Delphi inquiry systems. In: Wright G, Ayton P (eds) Judgmental forecasting. Wiley, Chichester, pp 129–156
Phillips LD (1987) On the adequacy of judgmental forecasts. In: Wright G, Ayton P (eds) Judgmental forecasting. Wiley, Oxford, pp 11–30
Phillips JM (1999) Antecedents of leader utilization of staff input in decision-making teams. Organ Behav Hum Decis Process 77:215–242
Price PC, Stone ER (2004) Intuitive evaluation of likelihood judgment producers: evidence for a confidence heuristic. J Behav Decis Mak 17(1):39–57
Rakow T, Newell BR (2010) Degrees of uncertainty: an overview and framework for future research on experience-based choice. J Behav Decis Mak 23(1):1–14
Rowe G, Bolger F (2016) The identification of food safety priorities using the Delphi technique. EFSA Supporting Publications, 13 (3). EN-1007, 141 pp
Rowe G, Wright G (1996) The impact of task characteristics on the performance of structured group forecasting techniques. Int J Forecast 12:73–90
Rowe G, Wright G (2001) Differences in expert and lay judgments of risk: myth or reality? Risk Anal 21(2):341–356
Rowe G, Wright G, McColl A (2005) Judgment changes during Delphi-like procedures: the role of majority influence, expertise, and confidence. Technol Forecast Soc Chang 72:217–238
Saunders EMJ (1993) Stock prices and Wall Street weather. Am Econ Rev 83:1337–1345
Solomon I, Ariyo A, Tomasini LA (1985) Contextual effects on the calibration of probabilistic judgments. J Appl Psychol 70:528–532
Tetlock P, Gardner D (2016) Superforecasting: the art and science of prediction. Random House, New York

Toulmin S (1958) The uses of argument. Cambridge University Press, Cambridge, p 272
Tversky A, Koehler DJ (1994) Support theory: a nonextensional representation of subjective probability. Psychol Rev 101:547–567
Walls L, Quigley J, Marshall J (2006) Modeling to support reliability enhancement during product development with applications in the UK aerospace industry. IEEE Trans Eng Manag 53(2):263–274
Wallsten TS, Budescu DV (1995) A review of human linguistic probability processing: general principles and empirical evidence. Knowl Eng Rev 10(1):43–62
Waterman DA, Hayes-Roth F (1978) Pattern-directed inference systems. Academic, New York
Wenger E (1998) Communities of practice: learning, meaning and identity. Cambridge University Press, Cambridge
Wright G, Rowe G, Bolger F, Gammack J (1994) Coherence, calibration and expertise in judgmental probability forecasting. Organ Behav Hum Decis Process 57:1–25
Wright G, Bolger F, Rowe G (2002) An empirical test of the relative validity of expert and lay judgments of risk. Risk Anal 22:1107–1122
Wright G, Rowe G, McColl A (2004) A framework for future study of expert and lay differences in the judgment of risk. Risk Decis Policy 9(2):91–106

Chapter 17
Eliciting Probabilistic Judgements for Integrating Decision Support Systems

Martine J. Barons, Sophia K. Wright, and Jim Q. Smith

Abstract When facing extremely large and interconnected systems, decision-makers must often combine evidence obtained from multiple expert domains, each informed by a distinct panel of experts. To guide this combination so that it takes place in a coherent manner, we need an integrating decision support system (IDSS). This enables the user to calculate the subjective expected utility scores of candidate policies as well as providing a framework for incorporating measures of uncertainty into the system. Throughout this chapter we justify and describe the use of IDSS models and how this procedure is being implemented to inform decision-making for policies impacting food poverty within the UK. In particular, we provide specific details of this elicitation process when the overarching framework of the IDSS is a dynamic Bayesian network (DBN).

17.1 Introduction

In our increasingly interconnected world, large systems are becoming more common and progressively more complex. This means that statistical modelling protocols must also evolve to accommodate these changes. Typically, in this new situation, decision-makers need to gather evidence from a variety of different expert domains. Each such domain has a limited number of people who are deemed experts in particular aspects of the interdependent system. So in such systems, we must develop ways to combine together evidence from these different domains into a coherent whole. The evidence we want to accommodate will typically be framed probabilistically and will often be supported by domain-specific probabilistic

M.J. Barons (✉) • S.K. Wright
University of Warwick, Coventry CV4 7AL, UK
e-mail: Martine.Barons@warwick.ac.uk; S.K.Wright@warwick.ac.uk
Grant EP/K007580/1, Grant EP/L016710/1

J.Q. Smith
University of Warwick, Coventry CV4 7AL, UK

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

L.C. Dias et al. (eds.), *Elicitation*, International Series in Operations Research & Management Science 261, https://doi.org/10.1007/978-3-319-65052-4_17

predictive models. Fortunately, there is now a technique which makes possible coherent inference over a network of these multi-faceted probabilistic systems that can provide such decision support for policymakers. We call this composition an integrating decision support system (IDSS), see Smith et al. (2015a,b).

17.1.1 A Probabilistic IDSS: Its Genesis and Functionality

One complication facing the coherent combination of judgements within an IDSS is that in the twenty-first century users are now typically teams, here called *decision centres*, rather than individuals. The implicit (albeit virtual) owner of beliefs expressed by this team will henceforth be referred to as the *supraBayesian* (SB). This SB embodies the beliefs of the decision centre. Through this construction we are able to address issues such as statistical coherence or rationality as it applies to the system as a whole.

Once a coherent system has been built, being probabilistic in nature, these huge composite models enjoy many of the advantages seen in probabilistic models of smaller systems. In particular, the algorithms to determine the efficacy scores are based on widely accepted formulae. Furthermore, these algorithms permit the smooth combination of expert judgements with any information obtained from experimental or survey data that might be available to the centre.

The need for such integrated systems became clear to one author of this chapter, whilst working with Simon French and others as part of a team designing a decision support system for operations crisis control after an accidental release of radiation from a nuclear plant. This research was organised as part of a large EU programme called RODOS (Real-time On-line DecisiOn Support) (Caminada et al. 1999; French and Smith 2016). Part of the decision support led to fast evaluations of the effectiveness of various candidate countermeasures, integrating information from a variety of sources in what was then called an "evaluation subsystem". When addressing this massive problem, we were forced to separate the description of the unfolding processes and their threats into components where, just as we described above, each component was informed by a separate panel of experts. Each panel was charged with providing its own domain information which was then delivered to the centre and combined with the results from other panels to score the efficacy of the different candidate countermeasures. To process information in this way, once the prescribed inputs from other components were delivered, each component would need to be able to take and then produce its outputs *autonomously*.

Although the RODOS development was seminal for its time and the long collaboration produced a valuable decision support engine, urgencies in its development meant that the architects and methodologies on which RODOS depended were often necessarily naïve. Furthermore, due to the constraints on computational availability, such a composite system was constrained to fairly coarse scales.

The project raised some important questions about the construction of decision support systems for multi-faceted problems like this. In particular there were two main concerns:

1. Firstly, this early system was unable to suitably allow uncertainties to be incorporated into the combined assessment in a sound or comprehensive manner. This meant that for example, if the expected consequences of one policy were marginally better than another, but the uncertainties arising from the first policy were much greater than the second, then suggesting to the decision centre that the first option was better than the second was often not formally correct and it may also have had disastrous consequences.
2. Secondly, the actual dynamics of the situation were often only partially incorporated into the evaluation of the system, meaning that it could only provide snapshots of how scenarios were unfolding.

Now, at last, general probabilistic methodologies are in place that can be implemented to properly process the dynamics of the system (Leonelli and Smith 2015, 2013a). We describe in this chapter how such an implementation is now being enacted. The theoretical development for IDSS's developed in Smith et al. (2015a) needed to address two fundamental questions:

1. "Is it even theoretically and logically justifiable to compose inferential methodologies using an overarching system like the one used for RODOS?".
2. "Can the scores associated with these formally correct and justifiable support systems be structured in such a way that the necessary calculations can be made quickly enough for the IDSS to be feasible?".

The first of these two questions, regarding the logical justification of creating such overarching systems, can be broken down into more specific queries: "What conditions guarantee all uncertainties expressed by experts are appropriately represented and processed in such an integrating system? Can the dynamic nature of such processes be captured and the progressively fine-grained information be processed appropriately? Can these, at least in principle, be constructed to guide the evaluation of policies in a way that takes proper account of all the component uncertainties and dynamics within such a composite system?". In order to answer these questions fully it was necessary to show that it was possible to break up a composite DSS into different, autonomously updated components and subsequently aggregate them together in a formally justifiable way.

Recently, we have been able to show that under commonly satisfied conditions, and with a careful construction of the components of the system and their interfaces, expected utility scores of candidate policies can be either perfectly calculated in this distributed way or approximately so. The proofs of these recent discoveries are necessarily rather mathematical and so beyond the scope of this book. However, these are presented in the public domain formally appearing in Smith et al. (2015a) and less formally in Smith et al. (2015b) and Leonelli and Smith (2015, 2013a,b). Furthermore, we can show that within such an integrated system the rationale behind the different component forecasts can be delegated to panels overseeing the corresponding components delivering these outputs. Here, we simply present the relevant summary results of this technical work.

In addition to these methodological advances, the second question raised the issue of computational feasibility in relation to speed of such systems for the IDSS to be feasible. Recent results show the answer to this question also to be "Yes". The theory for this assertion, as it applies to various types of such systems is published in Leonelli and Smith (2015). Therefore, the methodological obstacles that had faced both the justifiability of the RODOS IDSS and its ability to quickly and effectively calculate estimates and their associated uncertainties, have now been surmounted.

We note that both the theory and the algorithms referred to above apply to probabilistic models where both the composite model and its components have associated probability distributions. Here the changing, probabilistically expressed beliefs of the expert panels need to be elicited and then processed. Precisely because these systems are probabilistic, the relevant uncertainties associated with different candidate policies can be expressed directly via distributions and the stochastic dynamics represented in terms of various stochastic processes, using very well understood methods of uncertainty handling.

Of course it is one thing to demonstrate that this type of IDSS is *formally* justifiable and feasible to implement *in principle*, and quite another to apply the proposed methodology to construct an *actual working system* that can be used successfully to help inform a particular domain. Over the last couple of years we have begun to build such a system for UK food security in collaboration with policy-makers, using the theoretical development described above. This chapter describes how we are currently implementing the methodologies and the practical challenges we face when doing this.

17.1.2 The Running Example of Food Security

High on the UK government agenda is the issue of food poverty. The world population reached 7.3 billion as of mid-2015, which is the result of an expansion of approximately one billion people in the last 12 years and projections expect the total to pass nine billion by 2050 (DESA 2015). It is therefore vital that optimal use is made of the world's finite resources for food production (Collier 2009). With such a growth in population, the demand for food and its affordability has changed everywhere in the world. Stresses have been exacerbated by an increasingly extreme division of wealth between the rich and the poor, within and across nations, and the emergence of food riots in 2008 and 2011 (Lagi et al. 2011). This has resulted in making food poverty endemic worldwide and an increasingly serious threat to even wealthy nations such as the UK. As a consequence, it is very timely to develop decision support tools designed to help government policy-makers to assess the effects of policies they might enact in order to address the various threats of food poverty within their local domain. Strains felt by local government in the UK are currently severe because of progressively decreasing budgets which requires them

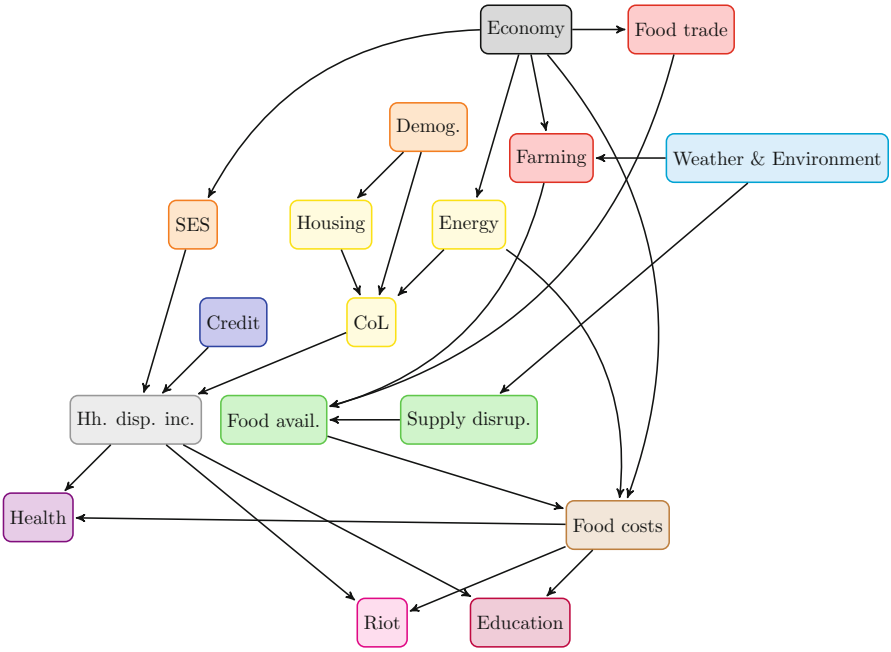


Fig. 17.1 A plausible schematic of information flows for the modules of a UK food security IDSS, with colours depicting different expert panels. KEY: CoL—cost of living; Credit—access to credit; Demog—Demography; Economy—UK economic forecasts; Farming—food production; Food Avail—UK Food availability; Hh. disp. inc.—household disposable income; SES—Socio-economic status; Supply disrupt—food supply disruption

to implement different forms of cutbacks to financial benefits and spending on social support, within the populations for whom they are responsible.

In this context, local governments need an IDSS. The overarching food security IDSS model integrates together all social, political and socioeconomic factors which may affect food poverty within the UK, shown in the schematic below, see Fig. 17.1.

It is possible to use a variety of different overarching frameworks to embody this integration. Here, we shall focus our attention on how we perform this integration when the overarching framework is a dynamic Bayesian network (DBN) which we define in Sect. 17.3.4.1. DBNs have already proven to be particularly useful when we have multi-faceted, interdependent systems which need to incorporate multiple models, natural processes and contextual information, for an example see Johnson and Fielding (2010). As well as modelling the expected course of a process, these networks can be used as a framework to intelligently integrate uncertainties about the impact of different events which would result from selecting different policy choices, driving mechanisms or external shocks. Furthermore, this framework can be embellished into a full probabilistic model to enable the decision centre to

combine expert judgement with data that tracks the unfolding process, as well as utilising new experimental evidence as and when this arrives. These methods use general probabilistic machinery, such as Bayes Rule, in ways discussed in Chap. 6 of this book (see Hartley and French 2018).

Whilst developing the IDSS we have found that, perhaps rather predictably, the expert panels largely mirror panels already created by organisations and governmental agencies. In the schematic, each colour depicts a different expert panel. The model is informed through a number of sources, for example:

- Demography and socio-economic status statistics (SES) distributions are available from the Office for National Statistics (ONS),
- Costs of housing, energy and general cost of living (CoL) are available via the consumer prices index (CPI),
- Food trade (imports and exports) and farming yields can be obtained from the Department for Environment, Food and Rural Affairs (DEFRA),
- Supply chain disruption and overall food availability can be obtained through DEFRA, the foods standards agency and the food and drink federation,
- Access to credit is available from the Bank of England,
- Household disposable income distributions come from ONS,
- Food costs are via the cost of a typical basket of food, which is systematically calculated as part of the UK CPI as calculated by the ONS.

Within each of these expert panels lies a complex sub-network. For example, the consumer price index (CPI) describes a typical basket of goods and services which are purchased by an average UK household, see Gooding (2016), ONS (2013). The average price of such a basket is reviewed both monthly and yearly. As we are interested in variables such as food costs, we can use the food element of the basket of goods as a measurable proxy. The model of the basket of food is disaggregated into various pre-defined food subgroups such as meat, fruit, vegetables etc. By probabilistically modelling each of these diverse subgroups individually and then reflecting on natural dependencies between them we can derive sub-networks which can be combined together to provide forecasts for the cost of the entire food basket over time.

Another sub-network is concerned with modelling pollination of food crops within the UK and this will form our detailed example in Sect. 17.3.2.5. It is estimated that 70% of important food crops are pollinated by bees (Datta et al. 2013) and the contribution of other insect pollinators is also significant (Rader et al. 2016). So the status of pollinators, and of bees in particular, is a key concern in global food security (Blaauw and Isaacs 2014; Lonsdorf et al. 2009).

In this chapter we will describe in some detail how we constructed the probabilistic component relating to pollination within this system, itself a DBN. We will discuss the implications of the sort of elicitation required within this context and various sorts of diagnostics that can be used to check its plausibility.

17.2 Framing a Complex Dynamic System

Although IDSSs may seem to be a suitable structure within which to model complex dynamical processes, it is not always appropriate to formulate a problem in such a manner. There are a number of criteria which must be fulfilled for an IDSS to be fully justifiable. Building an IDSS for a multi-faceted, interdependent system requires the following properties to hold:

1. Since the decision maker is typically *a centre rather than an individual* it is necessary that the centre needs to be populated by individuals who want to act constructively and collaboratively. In other words, the 'centre' must be motivated to strive to act as a single coherent unit for a common goal.
2. In particular, there must be consensus about the appropriate *utility structure* on which the efficacy of the candidate policies could be scrutinized, were certain unfolding of events to be certain. For example, such a consensus might be that the centre's utility functions should have preferentially independent attributes, see Keeney and Raiffa (1993).
3. Consensus also needs to be attained about an overarching description of the dynamics driving the process. This can be allowed to take a variety of forms depending on the context. In this chapter, we have assumed that this overarching structure can be represented by a DBN (defined in Sect. 17.3.4.1) although other frameworks can be used, for example see Leonelli and Smith (2015).
4. The necessary coherence of the group requires there to be a consensus about *who is expert about what* in order to identify appropriate expert panels. In a formal sense, this consensus in turn implies that individuals outside a domain should be prepared to adopt the beliefs of the expert panel of that domain as their own, see Smith et al. (2015a). It can then be proved that they should then delegate their reasoning about every domain to the appropriate panel.

Within our two contexts, the overarching food poverty system and the pollinator abundance sub-network, the first condition of a collaborative decision making group was broadly met. This means that all local government officers agreed that food poverty is an issue requiring action, although meeting this need impinged on several budget-holding departments, so requires negotiation and co-operation. Also, the recent National Pollinator Strategy was developed collaboratively. However, this co-operative approach is not always taken. Many decision centres work more like a court of law, especially those in public health policy making. In these cases, advocates fight both the case for and then against a policy and win the argument in a competitive way. In such a context, the support system we define here can then at best only support one side of the argument. Note that the sorts of methods we describe here have been used for exactly this purpose in a court of law (Puch and Smith 2002).

The second condition, utility consensus, is often delivered through decision conferencing. Very briefly, a facilitated discussion encourages the centre to first determine sets of broad objectives and preferences the centre should bear in mind when assessing the efficacy of different candidate policies. Then agreement is negotiated about a vector of attributes of the utility function that best capture the essence of these objectives, see Fig. 17.2 and Sect. 17.3 for more details. These vectors of attributes of the utility function must be specific enough to be measured in an unambiguous way, i.e. passing the clarity test, see Howard (1988). Thus, for example, an objective described as "minimising the number whose quality of diet was low over the coming year" could not be treated as an attribute, while on the other hand "reducing the number of individuals treated in hospitals A for conditions which included explicit mention of malnutrition within a geographic region B from Jan 1st 2020–2021" would be a candidate attribute.

The vector of attributes of the utility function then provides not only a transparent and unambiguous picture of what might happen in the future, but also one that balances different aspects implied by a given objective. In the context of the types of decision support we give here, it is very common for the utility function to depend on attributes that estimate the impacts of different candidate policies well into the future.

The quantitative form of a utility function will eventually need to be elicited in a manner which is appropriate to the centre. This utility is a function of the possible set of future values of the whole composite vector of attributes. There are many ways to do this that are described in detail in Edwards et al. (2005), French et al. (2009) or Chaps. 6 and Chap. 10 in this book (see Hartley and French 2018; Gonzalez-Ortega et al. 2018).

The next step in the elicitation process needs only to have identified the attributes of the utility functions, but not the quantitative form of the utility function itself. So although discussion needs to draw out those clearly specified measures on which preferences might depend, at this stage it is unnecessary, and indeed often unwise, to assign values of these attributes. This is often most efficiently performed once the overall structure of the problem and its relationships to other processes has been at least provisionally mapped out in a way we will describe below.

Remark 1 The elicitation of an IDSS can only be done robustly if performed in an *iterative* manner, meaning we review the qualitative structure of the IDSS repeatedly. At periodic intervals, previous steps of the process are reviewed and checked in light of the more profound understanding of the process acquired through further elicitation. This enables modification and improvement to the various contributing elements defined in the construction of the qualitative form of the model. The process continues until the decision centre is content that the structure is fit for purpose, called "requisite" (Phillips 1984).

Remark 2 Since the process of elicitation is an iterative one, it is often wise to begin with some simple measures, proceed with an initial structural elicitation, and then to revisit the initial list of attributes of the utility to consider whether these need to be adapted or supplemented so as to better measure the efficacy of the possible

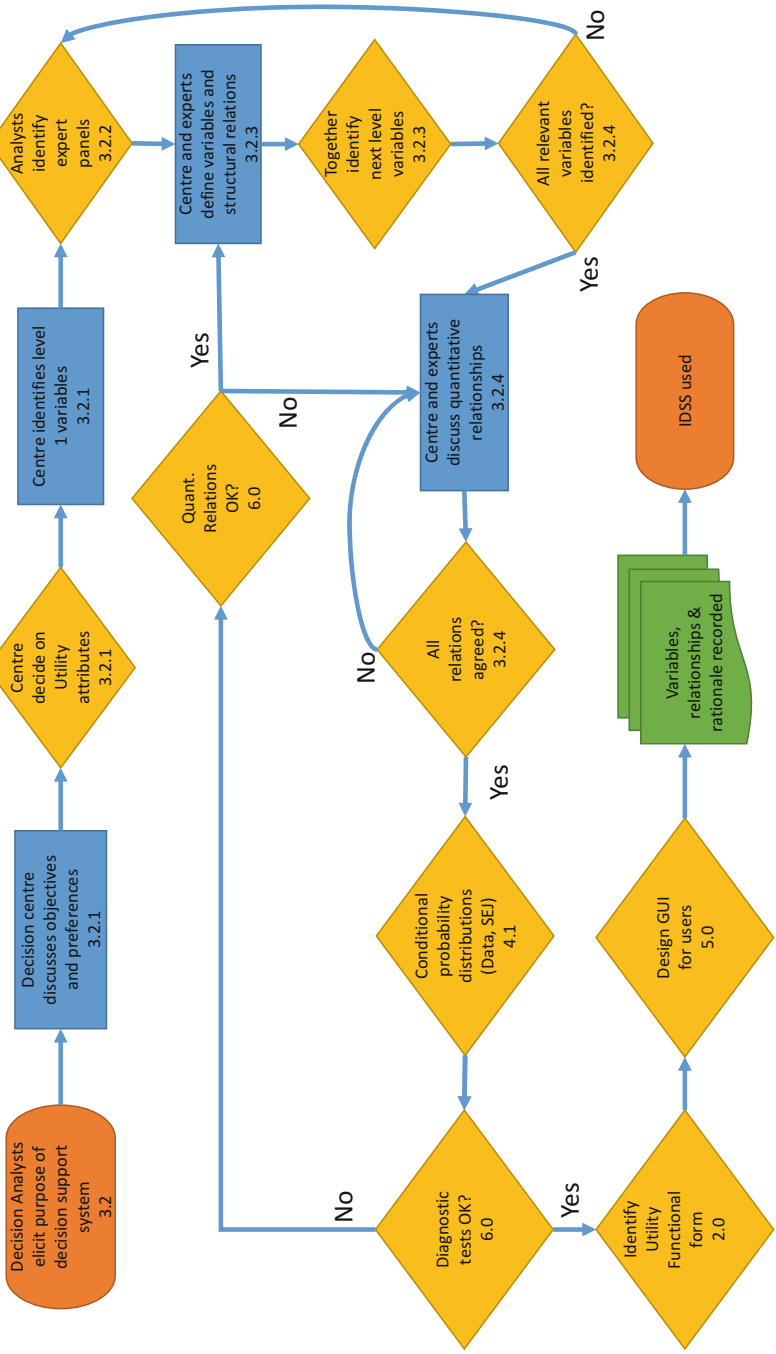


Fig. 17.2 A flowchart mapping the creation of an IDSS with section numbers to refer readers to specific sections of this chapter for more detail

candidate policies. In our experience it is often only after the science, economics or sociology driving the process has been more fully discussed that the decision centre can become fully aware of the suitability of certain types of utility attribute measures.

Since household food security (food poverty) is not directly measured in the UK, it was necessary for the local government decision-makers to devise a suitable proxy for this measure which is both relevant to their areas of responsibility and measurable. Through a sequence of decision conferences, we elicited from council officers three areas which they expect would be impacted by increasing household food insecurity within their jurisdiction: educational attainment, health (as the effects of malnutrition, or threats of malnutrition, on health in the short medium and long term), and social cohesion. Finally, of course, the cost and resource implications of applying any ameliorating strategy to address these negative consequences must be taken into account. Discussions also determined suitable measurements of these four attributes or proxies which would make suitable surrogates.

In the case of educational attainment, it is well-established that on average, pupils who are entitled to free school meals have a lower educational achievement in public examinations than pupils not entitled to free school meals. Eligibility for free school meals is a proxy measurement for deprivation since it is based upon household income. Educational attainment is assessed by a vector of different examination results: students are expected to gain SATs level 4 or above at age 11, and five or more GCSEs, including English and Maths, at grade C or higher at age 16. The UK Department for Education defines disadvantaged pupils as any child in care of the local authority, or any pupil who has been eligible for free school meals at any time over the last 6 years. Pupils classified as disadvantaged have a lower average educational attainment record than other pupils and there is a direct correlation between level of qualification and future employment and earnings.

Health concerns are captured by admission to hospital with a primary or secondary diagnosis of malnutrition, and death with malnutrition listed as a contributory cause. Figures are available from the UK Hospital Episode statistics for records of diagnoses of malnutrition and death records indicate whether malnutrition was a contributing factor. It is well documented that malnutrition has long term effects on the health of an individual, such as a greater risk of high blood pressure. As well as directly impacting the quality of life for the individual concerned, long term effects of malnutrition place a greater burden on the healthcare system. These longer term health effects routinely measured using mortality and morbidity indices.

Social discontent might be expressed in terms of, for example, food riots provoked by the inaccessibility of food stuffs, Lagi et al. (2011). The term ‘riot’ is tightly defined in UK law and although they are rare, they are a costly occurrence which the decision-maker is keen to avoid. Crime statistics data collected by the ONS record civil unrest including riots, criminal damage and looting.

Cost is measured directly by the amount of money that local councils must spend to implement and uphold any change in policy. For example, estimated costs associated with different intensities of disturbance can be estimated from the resource implications and damage to infrastructure in past riots.

Remark 3 Although the initial problem may look overwhelmingly complex, by focusing the centre and its expert panels on those issues that really impact on final outcomes we can vastly reduce the scope of deliberation. In particular, it soon becomes apparent to all that it is not necessary to capture *all* available expert judgements for decision support, but only those features that might be critical in helping to discriminate between the potential effectiveness of one candidate policy against another.

Once attributes of the utility have been decided, the next task is to construct a description of the processes leading to the different values of the utility attribute vector. This enables the centre to capture qualitatively how the system might respond to key unfoldings of events, see Sect. 17.3.2. Although this issue has not yet been addressed within this book, in our experience, this part of the elicitation process is a critical part of the elicitation process. Fortunately, many useful techniques for capturing these explanations have been widely documented and tested against thousands of applications (Edwards et al. 2005). In the following sections we review the basis of this work. We then describe how we are currently applying these methodologies to construct an IDSS for policies related to household food poverty and the sub-module within that IDSS pertaining to insect pollination of food crops.

17.3 An Agreed Picture of the Whole Probability Process

We shall now discuss in detail the process of creating and populating the IDSS. The steps below are based upon published theory and experience of applying these models to specific scenarios. We conclude this section with a look at our food security application. Throughout this section it may be useful to refer to the following flowchart (Fig. 17.2) which synthesises the process of creating an IDSS and refers the reader to specific section numbers as required.

17.3.1 An Overarching Structure and Common Language

In Remark 1, we required the centre to agree an overarching qualitative structure to provide a plausible description about how different features of the development relate to one another and how the future might potentially unfold. Obviously this description needs to be transparent enough to be understood by all experts in the system. In this application it can be expressed as a graph of vertices and edges which together express how variables in the system are believed to relate to one another. The logical foundation and probabilistic compatibility of our description ensures the graph can be expanded later into a full description of the stochastic process driving the utility attributes in any given unfolding of events. This in turn will enable the centre to evaluate the expected utility scores associated with those policy choices open to it.

This structure needs to be elicited and this elicitation should ideally include representatives of all domain experts across the system as a whole, and is best conducted using common language (as far as is possible). It follows that the type of elicitation we use for this stage is most often behavioural: the experts discuss various options and try to arrive at a consensus about the possible dependences, see Korb and Nicholson (2011), Smith (2010).

Remark 4 If there is strong disagreement about whether or not a dependency exists in the system then the group of domain experts should assume initially that a dependency does exist and only later explore whether in the face of evidence such dependence is supported. On the other hand, if there is a broad consensus that if a dependence might exist it will be weak then it is usually wise to omit this dependence in the description and only revisit and re-examine this assumption later when the understanding of the underlying process is more mature. In this way we are often able to contain the structure to a manageable size.

The question remains about how exactly we can capture such dependence hypotheses formally without first eliciting probabilities. Fortunately, a mature industry of graphical modelling has now provided us with a new set of inferential axiomatic systems. One language is based on the semigraphoid axioms (Dawid 2001; Pearl 1988, 2000; Smith 2010). These rules simply specify two simple properties we might expect to hold whenever someone asserts that knowing one collection of measurements $\mathbf{X}$ does not help the prediction of a vector of measurements $\mathbf{Y}$ once the vector of measurements $\mathbf{Z}$ is known. More formally and precisely:

Definition 1 Suppose that the client believes that the measurement $\mathbf{X}$ is *irrelevant* for predicting $\mathbf{Y}$ given the measurement $\mathbf{Z}$ (written $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{Z}$) so that once she learns the value of $\mathbf{Z}$ then the measurement $\mathbf{X}$ will provide her with no *extra* useful information with which to predict the value of $\mathbf{Y}$.

We assume that all members of the decision centre accept that when they make an irrelevance statement like the one above then two properties hold (Smith 2010). The first, called the *symmetry* property, asks that for any three disjoint vectors of measurements $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ **:**

$$\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z} \Leftrightarrow \mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{Z}.$$

In words this states that if $\mathbf{Y}$ is irrelevant for predicting $\mathbf{X}$ once $\mathbf{Z}$ is known then also $\mathbf{X}$ is irrelevant for predicting $\mathbf{Y}$ once $\mathbf{Z}$ is known. It is simple to check that this property holds for most probabilistic and many non-probabilistic methods of measuring irrelevance.

Even more compelling is a second property, called *perfect composition*, see for example Pearl (1988), for an explanation of this. This property asks that for any four disjoint vectors of measurements $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W}$;

$$\mathbf{X} \perp\!\!\!\perp (\mathbf{Y}, \mathbf{Z})|\mathbf{W} \Leftrightarrow \mathbf{X} \perp\!\!\!\perp \mathbf{Y}|(\mathbf{W}, \mathbf{Z}) \& \mathbf{X} \perp\!\!\!\perp \mathbf{Z}|\mathbf{W}.$$

More informally, this means that when forecasting **X** already knowing the value of **W**, the statement **(Y, Z)** are irrelevant to **X** is the same as saying **Z** is irrelevant to **X** and that after learning the value of **Z**, **Y** provides no relevant information about **X** either. Bayesian inference automatically satisfies this reasoning rule (as do many other alternative inferential systems).

These two reasoning rules allow us to elicit a collection of irrelevance statements from experts and deduce many others as logical consequences. More importantly, we note that this can all be done without mentioning probability and therefore can be done using common language. Furthermore, the graphs of the BN can be drawn *directly* from these types of irrelevance statements, requiring only one assertion for each vertex in the graph. The BN so constructed has a logical integrity and can be used as an overarching framework based on verbal statements from experts that can then be later embellished into full probability models. Each irrelevance statement of the form above then just transforms into a conditional independence statement in that probability model.

17.3.2 Defining the Features and Variables in a Problem

In any decision analysis we need to determine which parts of the process are both intrinsic to the description of the process, and are as yet uncertain. Uncertainty in a system can arise either due to lack of hard data in a certain element of the process or because the knock on effects from the implementation of a specific policy are not fully predictable.

17.3.2.1 What Are the Centre's Attributes and Time Frames?

The decision analysts along with the decision centre first import into the discussion the preliminary vector of utility attributes of the problem that might inform a one step ahead prediction of the value of each utility attribute vector, (Smith 2010). These are termed the "Level 1" vectors, see Fig. 17.2. We call the vectors which have direct relationships with the attributes of the utility, the level one vectors. Those variables which impact on the level one vectors, the level two vectors, and so on.

Within this early phase, the decision centre also need to decide what time step is the most natural one to use for the purposes of the support of the IDSS. The appropriate choice of these steps depends on a number of factors: for example the speed of the process, how relevant data is routinely collected on some of the components, and some technical acyclicity assumptions that are typically known only to the decision analysts. In our pollination application, the agricultural cycle is typically annual whilst pollination is required only for specific months of the year and pollinator life cycles are measured in weeks. However, the abundance of honey bees is measured in an annual survey. The most crucial factor in determining the

most appropriate step length is the timing in which the success of the system will be appraised. This early stage of the elicitation is perhaps the most delicate because there is often conflict between the granularity of for example, informing economic models of the process and sample survey regularity, and the needs of the system. This is an important issue. The granularity needed is driven by the granularity of the attributes of the utility. Decision analysts take a great deal of care to match precisely the outputs of a donating panel with the requirements of a receiving panel, see Remark 5. When these do not naturally align, then some translation may be needed between them, the nature of which will be advised by the relevant expert. It can also be shown that the further from the utility vectors a vector is, the less precision is required, as these have a much smaller impact on the utility scores for the candidate policies. Typically, the decision analysts would advise that the needs of the centre and the time horizons at which they work should be prioritised when making this trade-off.

17.3.2.2 Who Can Inform These Attributes and How?

We then need to *identify experts* who might understand how to *forecast* the future value of each utility attribute vector, albeit in the light of some necessary input information on these features, see the next point. Note that this elicitation will need to be framed within the granularity of the support system as discussed above. These experts will provide a provisional list of further candidates to populate the panels, as described in bullet point 4 in Sect. 17.2. Note that sometimes it can be appropriate for a single panel to be responsible for more than one utility attribute vector.

It is important to note that the entire process is highly iterative. General experts on the system (for example academics) use literature and any prior knowledge to sketch the potential process. Then experts are identified who might lead panels of experts and they confirm or adjust the understanding of the systems component variables and the dependencies between them. If new variables or subsystems are so identified, then experts from these domains are consulted and their contributions incorporated. This process continues until all the relevant variables are identified and the relationships agreed.

17.3.2.3 Firming Up Meaningful Inputs and Outputs

Representatives of the expert panels are then asked what past or contemporaneous values of the attribute vector or other "Level 2" features (see Smith 2010) they would need to know before being able to calculate the forecast distribution of the utility attribute they take responsibility for—again over the required granularity. The newly identified Level 2 features must then be firmed up and clarified so that they are able to pass the clarity test (Howard 1988, 1990). If the panel is supported by its own domain specific probabilistic software then this question is often easy

to answer. Since such software is usually more fine grained than is needed by the encompassing IDSS, the required attribute forecasts are often simple aggregates—across time and variables—of statistics already calculated by the software. However, in this case our task is often simply to ask the panel to use their software and so in this way to contribute various means and variances conditional on the values of their required inputs. This, for example, is true of many of the economic components used in our Food IDSS. On the other hand it is also often the case that no such probabilistic model is currently available. In this situation the IDSS designers will need to construct a probabilistic module that is able to do this job. Later in this chapter we will describe in some detail how we have performed this task for the pollination services sub-module within the food IDSS. We will see that this usually involves the construction of another qualitative sub-system—often itself either a BN or dynamic BN (DBN). First, new provisional panels need to be identified that can discuss and describe these processes. We will then need to elicit from these discussions a vector of measurement variables that can act as a surrogate for the expert's elicited understanding of the underlying processes. Out of this construction the experts will need to specify inputs it needs to be able to forecast its output. In the pollination example below, one of these concerns different types of weather conditions provided through meteorological forecasting systems, determining how insect pollinator abundance might fluctuate within a given year.

17.3.2.4 Iterations to Provide Causal Chains

We now take the Level 2 vectors of input measurements needed for the forecasts of the utility attributes and repeat the process substituting these clearly defined vectors for the attribute vectors above. The collections of any vectors needed as input to these models that have not already appeared as either Level 1 or Level 2 vectors we label as "Level 3" vectors. So in the weather example above these new components might involve various measurements of climate change like average earth temperature that will affect weather changes.

We iterate this process, deepening the levels until all input variables have currently known values or are sufficiently remote from the attributes of the process to not have a major impact on the forecasts needed by the system to determine high expected utility scoring policies.

17.3.2.5 Example

Within our Food IDSS one attribute of the utility function of the centre is health of the given population as a function of possible malnutrition.

The predictions of this vector of health indicators appear as functions of other "input" variables as defined by the relevant expert. One of the components of this input variable is the cost of feeding a household, based on the cost of an

appropriately defined basket of food for a household (a Level 2 variable), which passes the clarity test, (Howard 1988). To provide an appropriate joint distribution over components like these will be the primary task of our team. In this sense it will constitute one of the attributes of this expert panel's utility.

One of the inputs needed to forecast this food cost is a specific clarity tested measure of the abundance of particular fruits which will be processed into products in the basket of food above (a Level 3 variable). These in turn will be influenced by pollinator abundance measures (a Level 4 variable). These measures will need models as functions of other Level 5 variables, like weather and environmental factors.

Once this process has been completed the centre will have elicited a collection of random vectors partitioned into depth levels where the level of this depth reflects its distance from the attributes of the decision centre's utility function.

Remark 5 Within this process it is absolutely critical to ensure that the outputs delivered by expert panels at deeper levels of the process *precisely* match the input requirements of the receiving expert panel. Since the donating panels and the receiving panels work in different domains there is a clear danger of confusion at the interface. Unlike in many expert systems, note that in these composite systems the output variables delivered by the expert panels are determined by the needs of the *receiving* panel and are not self-determined. This helps simplify the system, but the needs of the receiving panel can be unfamiliar to the donating panels. Sometimes quite deep discussions are needed between representatives of the adjacent panels before this delivery can be successfully understood and addressed. In extreme cases this might require the panel to develop some interface software, using their domain knowledge in conjunction with a statistical model to transform their standard output to the needs of the composite system.

Remark 6 It can be shown both in a formal sense and also empirically that the distribution of variables furthest away from any attribute tends to have the least effect on the scores of competing candidate policies. So often ignoring uncertainties and using a naïve plug in estimate for these variables, based on for example official statistics and predictions, sample surveys or estimates from an expert who know the field, will be sufficient. An experienced analyst will be able to guide the decision centre when sufficient levels are in place. Typically we would go one level deeper than needed so that deliberations of how things are affecting each other and any associated significant dependences are informed by this extra level.

Notice that the random vectors informing levels of the system described in Fig. 17.4 are often in practice vectors that will need to be indexed by time. So, for example, pollination by bees is affected by weather conditions the previous winter—affecting colony overwintering survival—and also contemporaneous weather conditions affecting bee activity levels. We often need to study and depict these different categories of relationships in the next stage of the elicitation process.

17.3.3 Listing Measurements in a Causal Order

Following the process described above we obtain a hierarchy of well defined random vectors $\{\mathbf{X}_{1,t}, \mathbf{X}_{2,t}, \ldots, \mathbf{X}_{n,t}\}$, where $t = 1, 2, \ldots$ for the prediction of the process in the coming time period. The vectors associated with a given time period are called a *time slice*. Each of these vectors can be associated with a delivered output from one of the panels. In both our running example of the overarching food system and pollinator sub-network, the natural forecast time periods turned out to be a year. Note that the overarching model need not have the same time period as component models. Here, we index these vectors so that the last vectors in the list corresponded to the attributes of the decision centre's utility function, whilst the earliest random vectors correspond to the random vectors in the deepest levels of the process, $\mathbf{X}_{i,t}$, conventionally called its *parents*, as $Pa(\mathbf{X}_{i,t})$, $i = 1, 2, \ldots, n,$. Notice that from the construction and labelling above $Pa(\mathbf{X}_1)$ is the null vector—because by definition $\mathbf{X}_1$ will need no inputs from other components in the system. We call those vectors $\mathbf{X}_i$ for which $Pa(\mathbf{X}_i)$ is the null vector, $i = 1, 2, \ldots, n$, *founder vectors*.

We have consciously chosen the elicitation above to be consistent with the principle that it is easiest to think about a vector by seeing this output as an effect whilst inputs can be loosely seen as causes. We are now able to construct pictures, as in Figs. 17.3 and 17.4, that depict processes consistent with this perceived causal structure. For example, obviously it will not be possible to use variables associated

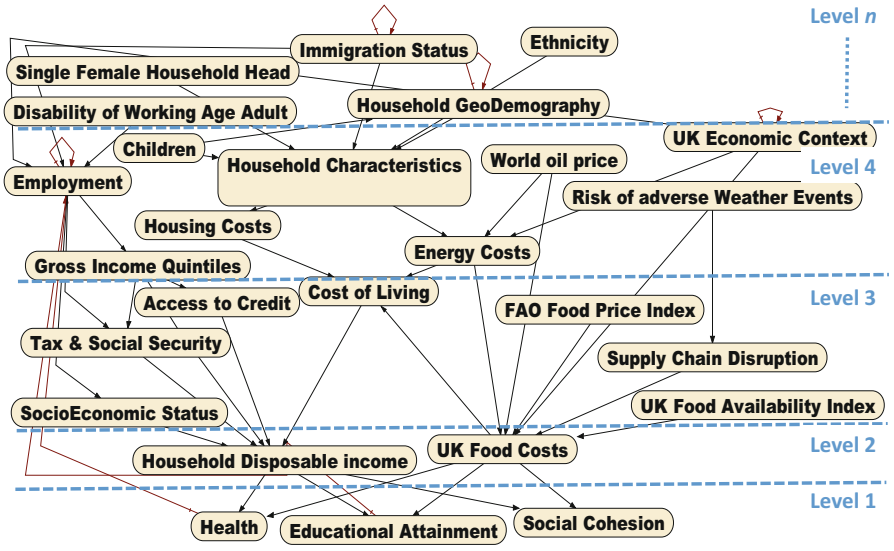


Fig. 17.3 A more detailed view of the IDSS, highlighting the target variables in Level 1 and higher level nodes. This is a DBN framework, note that the *red lines* indicate variables influencing another variable in the *next* time step, see Sect. 17.3.4.2 or Korb and Nicholson (2011), Smith (2010)

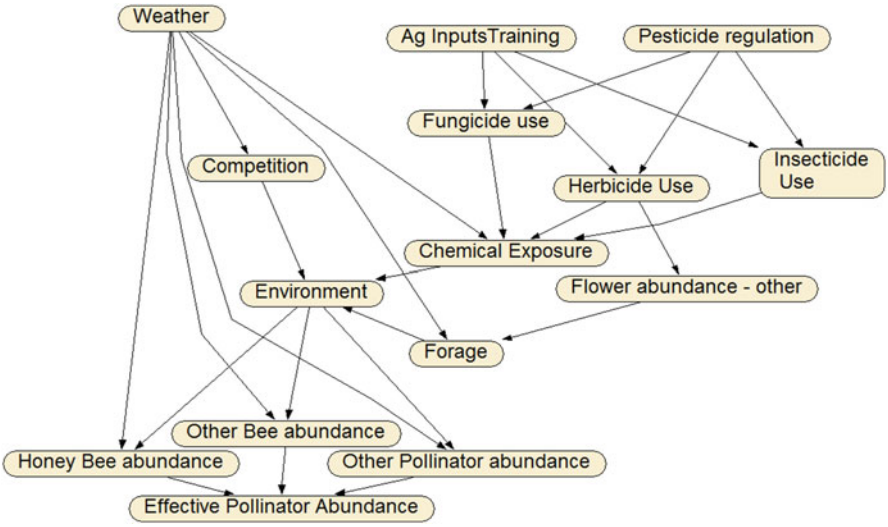


Fig. 17.4 In this fragment of the network, agricultural inputs training and pesticide regulation operate through several layers of nodes which have more direct effects on pollinator abundance, so neglecting the uncertainty in the estimates for these will have a less deleterious effect on the utility scores that neglecting the uncertainty on the estimates of, for example forage or weather. Produced in Netica, Norsys (1994–2016)

with future vectors along with predictive elements from the past and now at the current time. Thus suppose at each time t there are K_l vectors on level l, where $l = 1, 2, \ldots, L$ and let

$$\overline{\mathbf{X}}_{lt} = (\mathbf{X}_{1lt}, \mathbf{X}_{2lt}, \ldots, \mathbf{X}_{klt}, \ldots, \mathbf{X}_{Klt})$$

Definition 2 We call a listing *causally compatible* if under the notation above inputs of the component having $\mathbf{X}_{klt}$ as an output must be indexed before $\mathbf{X}_{klt}$.

For simplicity we will henceforth assume here that for our particular IDSS the elicitation admit at least one casually compatible ordering. This will enable us to represent the overarching process as a DBN.

Remark 7 If this condition cannot be met initially then we can often induce it by choosing a finer temporal step. Alternatively, if this transformation does not work then it is sometimes possible to omit some of the entries in the parent set corresponding to elicited contested or weak dependences. If neither of these reconstructions is possible then we will need to represent the problem by a reciprocal graph (Koster 1996). The formal methods we describe below then still apply. However, the outputs of the system are then less transparent and the calculations we can make directly from the system are far more costly and time consuming to make.

Through the construction above we can allow that contemporaneous vectors $\mathbf{X}_{klt}$ on the same level to depend on one another. However, the causal compatibility

condition above ensures that inputs of the component having this as an output must be indexed before $\mathbf{X}_{klt}$. With this constraint we can now choose a listing of variables time slice by time slice as follows:

Time	1	1	...	1	2	2	...	2	⋯	T	T	...	T
Depth	L	$L-1$	...	1	L	$L-1$	...	1	...	L	$L-1$		1
Vector	$\overline{\mathbf{X}}_{L1}$	$\overline{\mathbf{X}}_{(L-1)1}$		$\overline{\mathbf{X}}_{11}$	$\overline{\mathbf{X}}_{L2}$	$\overline{\mathbf{X}}_{(L-1)2}$		$\overline{\mathbf{X}}_{12}$		$\overline{\mathbf{X}}_{LT}$	$\overline{\mathbf{X}}_{(L-1)T}$		$\overline{\mathbf{X}}_{1T}$

Since through our elicitation process we preclude the possibility that higher time indexed outputs can serve as inputs to lower time indexed ones, by concatenating the vectors $\overline{\mathbf{X}}_{lt}$ in the order above we obtain an ordering of the vectors that will describe a causally compatible listing of vectors.

17.3.4 Bayesian Networks and Dynamic Bayesian Networks

The IDSS model works for many different frameworks depending on the underlying purpose of the DSS. For our food security application we have chosen to use DBNs and we shall therefore define this specific type of graphical model in this section.

17.3.4.1 Defining a Graph

Once this process is completed it is straightforward to express these elicited dependence statements graphically.

Definition 3 A Bayesian Network of a causally compatible listing of random vectors has as its vertices the set of these component vectors. A directed edge exists from $\mathbf{X}_{k'lt}$ into $\mathbf{X}_{klt}$ if and only if $\mathbf{X}_{k'lt}$ is a component of $Pa(\mathbf{X}_{klt})$, with $Pa(\mathbf{X}_{klt})$ as defined above.

In Smith (2010) we proved that this is indeed a Bayesian Network. In fact it is also a Causal Bayesian Network in the sense of Pearl (2000) and can be used as the basis of a DBN. So in particular, not only does this represent the genuine beliefs of the expert panels about how one component of the system affects another, it is not merely a representation but has a formal interpretation. This means that the directed graph itself—representing the composite of dependences expressed in terms of *individual local judgements*—can be interrogated for its plausibility of the logically implied global picture of interdependences as a whole! So once the first parse of elicitations takes place we can examine, through for example the use of the d-separation theorem (Pearl 2000; Smith 2010), whether the composite structure representation really can stand up to scrutiny. The way that such an interrogation might proceed is discussed and illustrated in Smith (2010).

Finally, because of the formal compatibility of the system, after a sequence of interrogation steps pick out a structure which to all responsible parties seems

plausible we can immediately use this graph as a framework for constructing a completely specified stochastic process. This is done by eliciting, for each vector $\mathbf{X}_{kLt}$ in the system, a probability distribution of this vector conditional on each configuration of its parents. Explicitly the joint density of the whole process can be defined by

$$p(\mathbf{x}) = \prod_{k,l,t} p_{k,l,t}(\mathbf{x}_{k,l,t}|Pa(\mathbf{x}_{k,l,t}))$$

where $p_{k,l,t}(\mathbf{x}_{k,l,t}|Pa(\mathbf{x}_{k,l,t}))$ are simply the conditional density or—in the discrete case—conditional probability table (CPT) of $\mathbf{X}_{k,l,t}|Pa(\mathbf{X}_{k,l,t})$. Each of these components in the formula may already be available from existing probability models specific to a given domain. In reality once the system has been specified we will inevitably find gaps where no such formal analysis has taken place. So to fill the gaps, to obtain quantitative measures of the required expected utilities, we need to elicit such distributions directly: see pollination example below. Often these distributions will actually be margins of other BNs.

Remark 8 When experts design their own systems, sometimes the *internal structure* of one component can share variables with the internal structure of another. So, for example, flooding could disrupt both the production of food and its distribution and yet these might be forecast using different components. In such cases, the coherence of the system will be lost and the most efficient way to ensure ongoing coherence is to separate out the shared variables and ask the panels concerned to take as inputs, probability distributions from the expert panel in, for in this example, flood risk.

17.3.4.2 Feasible Graphical Models and Simplifying Structures

The construction described above, whilst formally powerful would also lead us to build directed graphs which, in the contexts we are describing above might have hundreds or even tens of thousands of vertices. Obviously, as pictures, such graphs are very difficult to read or synthesise. Furthermore, the elicitation of the various conditional probability tables or distributions needed for the system would be prohibitive. However, there are now various techniques available to reduce the number of specifications and to make the depiction of the underlying processes more accessible.

There are two established collections of assumptions which restrict the underlying BNs to take a particular form and so vastly reduce the necessary inputs to the system which make even very complex models amenable to a parsimonious analysis. The two families of model are; the multiregression dynamic model (MDM) (Queen and Smith 1993) whose use in conjunction with these complex dynamic systems is discussed in, for example (Leonelli and Smith 2015; Smith et al. 2015a,b) and the two time slice dynamic Bayesian network (2TSDBN), see Korb and Nicholson (2011). Accommodating various additional assumptions about the

process reduces the picture to a graph over one or two time slices. Since the 2TSDBN is less technical and easier to describe and is used within our example we will detail only this particular construction in this chapter.

The 2TSDBN simply elicits the DBN up to the variables in the second time slice. It then makes an additional Markov assumption: it assumes that all the parent components can be written as

$$Pa(\mathbf{x}_{k,l,t}) = \left\{\mathbf{x}_{k',l',t'} : k' \in K', l' \in L', t' = t or t-1\right\}.$$

The key point here is that we assume once dependencies between contemporaneous values of measurements and the most recent past values are known, then the more distant past provides no further useful information. Furthermore, we assume that the process is time homogeneous after the first time point. Suppose the time step is a year. Then under these assumptions determining the conditional distributions associated with processes with what will happen next year and how this is affected by what is happening this year, the process can be fully specified see Figs. 17.5 and 17.6. In particular, to depict the process we simply need to specify a graph on

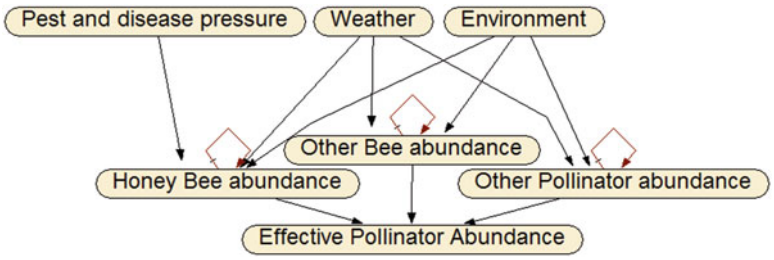


Fig. 17.5 We model the fact that pollinator abundance in the current time is influenced by pollinator abundance in the previous season, through the numbers entering overwintering. This is shown as a *red* self-loop here, but this violates the DAG requirements. Produced in Netica, Norsys (1994–2016)

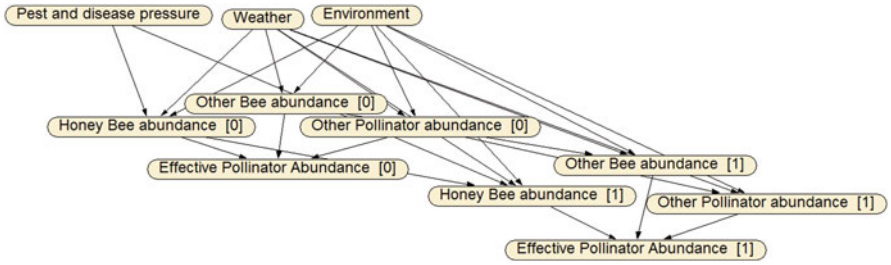


Fig. 17.6 Here the 2 time-slice DBN has been expanded in time. Note that, as well as weather and environment effects, at time [1] honey bee abundance is affected by honey bee abundance at time [0] and the same is true for the other insect pollinator types. Produced in Netica, Norsys (1994–2016)

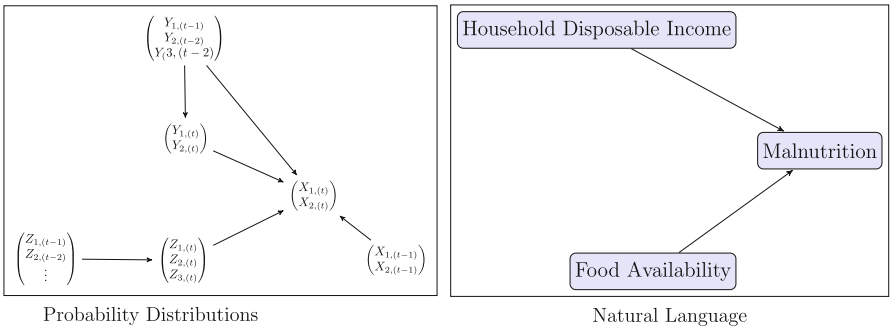


Fig. 17.7 Process of moving from natural language, to probability distributions, or vice versa

two time slices: containing vertices associated with measurements for this year and measurements for next and the dependencies between variables in the two time-slices. We illustrate this in the example below by unwrapping a 2TSDBN, Fig. 17.5, into a non-dynamic Bayesian network, Fig. 17.6:

Each node in our schematic, shown in full detail in Fig. 17.1, represents a list of random variables. An arrow is drawn between nodes when *at least one* variable is causally connected to at least one other variable in another node, or if there is a temporal relationship present. For specific classes of BNs such as Object Orientated Bayesian Networks (OOBNs) we usually begin with the probability distributions and then group similar objects together to create the overarching class nodes, moving from left to right in Fig. 17.7. Note that the schematic for Fig. 17.7 can be formally interpreted as a BN provided that we understand the schematic as representing a BN *conditional* on the relevant past variables from the time $t-1$ slice. Such conditional BNs have recently proved a useful modelling tool in other contexts, see for example Oates et al. (2016). However, in our application we are trying to construct our BN using literature and experts, so we first use common language to derive a general schematic and then more formally break these class nodes into more specific and detailed probability distributions. Working from a coarse to fine level in this way is a much more natural process for applications such as ours and is equivalent to moving from right to left in Fig. 17.7.

Of course the formally elicited graph is still there and can be used for formal deductions and construction. The simplified summary graph is nevertheless useful for depicting some important features of the elicited dependence structure and feeding this back to the panels for discussion and verification.

Remark 9 When studying and verifying these pictures of the elicited process of the process it is often useful to compare a structure with various mind maps and other schematic depictions informed by deep reflection by experts in the field. Although such pictures of dependences rarely have a formal interpretation, they are often the

result of deep reflection and provide supporting narratives which might produce compelling reasons for adopting modified dependencies. Furthermore, they are useful for helping populate provisional initial lists of vectors on which the process needs to be built.

Remark 10 Before the elicitation starts it is always necessary to do some preparatory work. With the help of various friendly domain experts, the analyst will need to trawl any relevant literature and check which hypotheses found there might still be current.

In this way, the elicitation of the qualitative structure of the determinants of household food security in the UK began with the decision-makers, local government, determining the attributes of their utility function.

Beginning with the schematic structure of the UK food system presented in Collier (2009), a relevant literature search was used to construct a plausible model of the qualitative structure, starting with the Level 2 elements which influence directly the attributes of the utility defined by the decision-makers. This qualitative structure was then iterated in consultation with food poverty domain experts from nutrition, politics, sociology, crop science and the local authority decision-makers in person, as well as using the reports and other resources they recommended. A final version was produced which represents the consensus of these experts.

We stated that it is often necessary, especially in the context of food poverty modelling, that the designers of the system will have to enable the panel to build a bespoke probabilistic model as one of the components. Perhaps the most straightforward way of doing this is to use DBN technologies with their supporting softwares to define this, especially when the probabilities expressed in the processes represent expert judgements. From a coding perspective we then have a DBN "object" that represents a node and its dependency structures.

Over the last 2 years this element of our research programme has been so important we will spend the next section describing one such elicitation in some detail. The process for the elicitation of the sub-component is more focussed and fine-grained, but otherwise identical to the elicitation methodology for the overarching system. So in this way, since both structures we choose to define here are dynamic forms of the BN, we can achieve both of these objects simultaneously.

So we next describe the process of eliciting the structure of the process leading to pollination, by a panel who were expert in the process of pollination as it applied to crop yield. This output would then inform any inputs associated with both the availability and price of certain items in a basket of food available to a person with stretched means and hence their health, educational attainment and social discontent. The analysis gave some of the probabilistic judgements we needed with a suite of BNs that were needed in relationships between the broad category labels such as Weather & Environment and Farming, displayed in Fig. 17.1.

17.4 Bayesian Networks for a Component Model: A Case Study

Embedded in the crop production element of the UK food security model is a need for pollination services. Since a large proportion of important food crops are insect-pollinated and the current concern about falling numbers of pollinators impacts on food production means this is an important element to model. There is also a need for decision support for those charged with ensuring the implementation and ongoing development of the UK's National Pollinator Strategy. However, there is considerable uncertainty and a dearth of evidence for some key parts of the DBN representing the pollination system, so for this we conducted a structured expert elicitation using the IDEA protocol: described in Chap. 5 of this book (see Hanea et al. 2018).

17.4.1 Development of the Bayesian Network Structure

The UK pollinator strategy recently published by DEFRA (DEFRA (2014)), provided the backbone for the development of the implicit utility function (pollinator abundance) and the relevant expert panels to provide evidence on factors influencing pollinator abundance. The abundance of pollinators can be subdivided into three categories; abundance of managed bees, abundance of other wild bees and abundance of other pollinators. Having established that pollinator abundance with regards to pollination of UK crops is the target of our model, the variables affecting this directly and indirectly were then identified in an iterative process as follows.

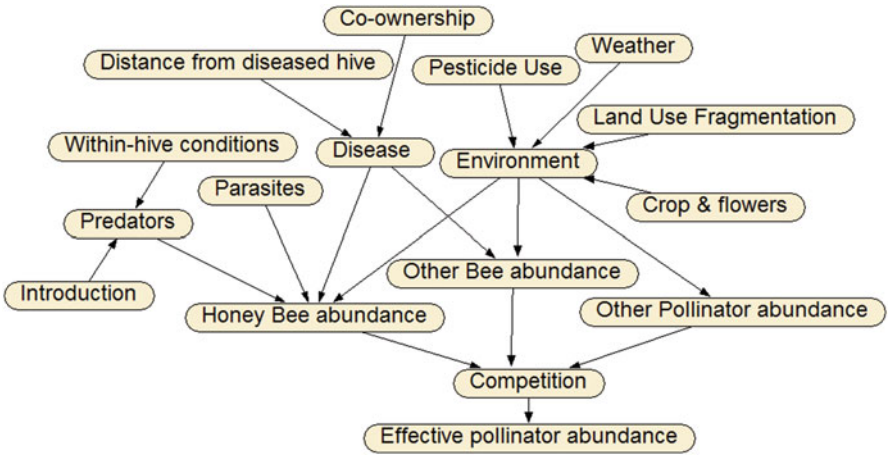


Fig. 17.8 2014: first draft of DBN with academics. Produced in Netica, Norsys (1994–2016)

The first draft of the BN was produced by academics, see Fig. 17.8, including one of the authors of this chapter, contributing respectively, expertise on BNs, honey bee disease dynamics and species distribution models. Using their background knowledge the first sketch was drawn, challenged and re-drawn until the underlying probability statements implied by the structure seemed plausible. The variables identified at this stage were the availability of forage, suitable nesting sites and the prevalence of disease. These in turn were influenced by the weather, pesticide use, competition and land use including crop type distribution. A greater quantity of documented research is available on the diseases, parasites and predators affecting honey bees than wild bees and other pollinators, and this is reflected in the first draft of the BN, Fig. 17.8.

The second step was to search the academic literature on pollinating insects and incorporate the new insights gained into the next iteration of the BN. It became clear that weather conditions affect the dynamics of the system on many more levels than previously thought, for example affecting the prevalence of parasites of honey bees and the availability of forage. Through its effects on weeds, fungi and insects that attack food crops, weather also influences the specific pesticide product employed, which in turn may affect pollinating insects. This was incorporated into the second draft, Fig. 17.9.

The third stage was to conduct a series of interviews with pollinator experts, undertaken by one of the authors of this chapter, ensuring we included as many different types as possible: beekeepers, government agency experts, Queen breeders, academics, honey producers, wild bee experts and government researchers. The next iteration of the qualitative structure included the direct effect of weather on the insects themselves—bees are prompted to forage by daylight and temperature and colonies can fail to survive if a winter is too long or too cold and therefore

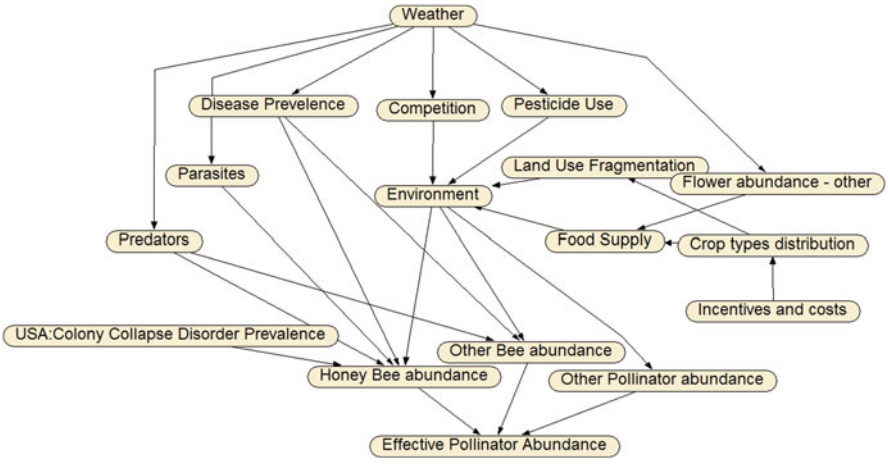


Fig. 17.9 The second draft incorporated insights from a detailed search of the academic literature. Produced in Netica, Norsys (1994–2016)

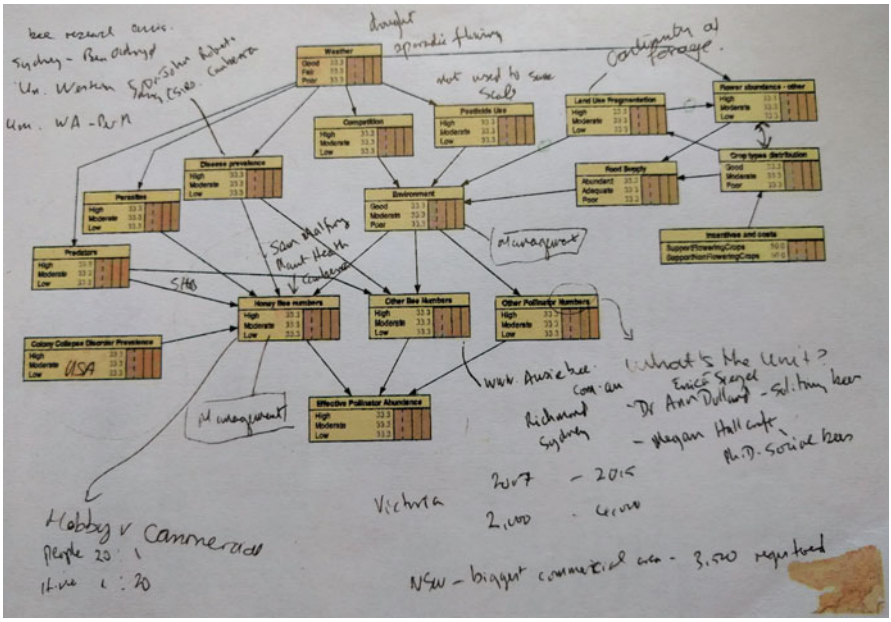


Fig. 17.10 Some experts annotated the network themselves

weather also affects numbers directly. These experts also informed us that the BN should include human factors; in the case of honey bees, the competence of the bee-keeper and for the environment, some measure of its management. As many of these interviews as possible were carried out face-to-face and the experts, having had the network explained to them, indicated changes that needed to be made or annotated the network themselves as in Fig. 17.10. For those who were available only by telephone, a list of questions was constructed to guide the conversation, but they were also allowed to comment freely from their perspective. Further resources recommended by these experts, such as government reports, policies, research articles and other experts, were followed up and the structure adjusted accordingly. Happily, there were no mutually exclusive opinions expressed, so the network structure expresses the consensus of the domain experts. These interviewees also helped to populate the decision space with the candidate policies which could be enacted with an assessment of the likely efficacy under given circumstances, although some of the interviewees were from overseas and so some of the strategies were not appropriate to the UK due to the varying bee species and ecological systems. These interviews were also very useful in helping to refine the definition of each of the variables which informed us how they could be measured. Finally, a DBN expert with extensive experience in using DBNs for decision support in an ecological domain suggested that the disease/treatment pairs be explicitly included, to evaluate fine-grained specific interventions for example the inclusion of Antibiotics, Miticides and Pesticides nodes, Fig. 17.11.

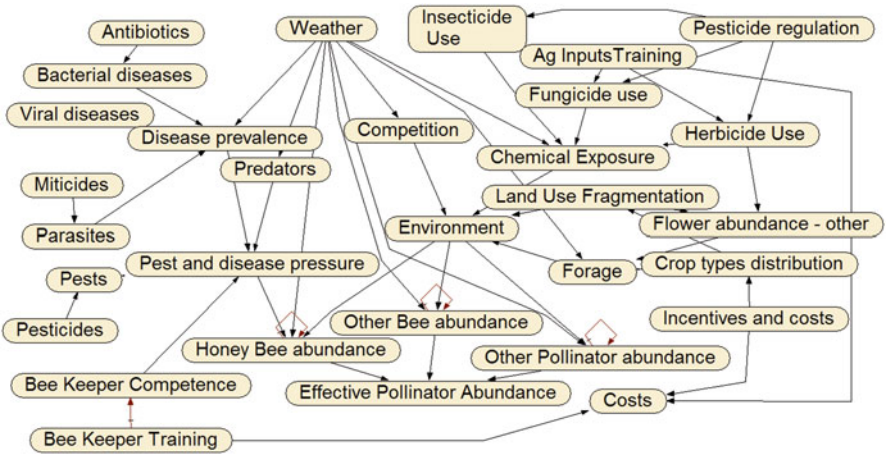


Fig. 17.11 Fully detailed DBN, including the time dependencies. Produced in Netica, Norsys (1994–2016)

The final addition to our network was the time dependencies (seen in Fig. 17.11 as red loops or lines)—all variables influence each other within the same season, but the abundance of pollinators entering the over-winter period directly affects their abundance in the next season and similarly bee-keeper training affects management in each subsequent season as well as the current one, so these are represented by time-delay links in the DBN.

This completes the structural elicitation phase of this component.

17.4.2 Eliciting Conditional Probability Tables

Having established the qualitative framework and checked the implied conditional probabilities make sense to domain experts, the next task was to populate the conditional probability distributions of the variables at each node. The academic literature is able to provide some estimates of these, particularity the marginal probabilities, but the conditional probabilities were not so readily available. The IDSS theory developed in Smith et al. (2015a) shows that we can legitimately and coherently admit expert judgement as evidence in an IDSS alongside experimental and observational studies. We therefore turned to expert judgement to populate the conditional probability distributions for the parts of the system with least evidence and most uncertainty.

A new set of UK experts in pollinators were gathered for a structured elicitation exercise, using the IDEA protocol, to provide probability distributions for the effects of weather, environment and disease on abundance of honey bees, other bees and

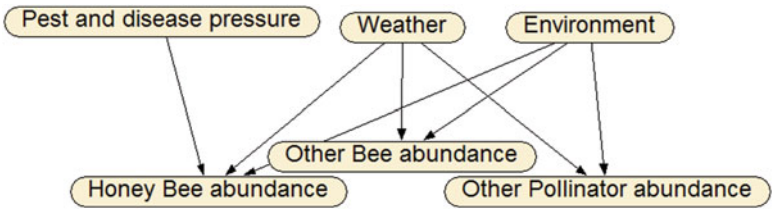


Fig. 17.12 Expert elicitation was used to derive conditional probability distributions for the parts of the system with least evidence and most uncertainty. Disease pressure, weather and environment each had two levels, giving eight combinations to elicit for honey bees and four each for other bees and other pollinators, 16 questions in all. Produced in Netica, Norsys (1994–2016)

other insect pollinators, such as hover flies. The fragment of the Bayesian network for which quantities were elicited is shown in Fig. 17.12.

This new panel of experts were identified with the assistance of the government agency responsible for pollination in the UK. Eleven experts participated in the workshop, their expertise covering honey bees, plant-pollinator relationships, bees and farming, hoverflies, pollinator viruses, wild pollinators and mathematical epidemiology, including epidemiology of honey bee diseases and pests, in line with the IDEA protocol. A list of resources was circulated as background reading, to make the evidence available equally to all experts, and experts also forwarded suggestions for further resources which were circulated.

One expert also attended for the previous day to lend domain knowledge to the definition of the variables and refinement of the questions of interest to ensure that they made sense to the domain experts, and were precisely and unambiguously worded. The workshop started with a presentation of the problem and presentation of the IDEA protocol. The question of biases was explained and the importance of answering the questions in the order they are posed, in order to avoid anchoring, was also discussed. The definitions of the variables were scrutinised and there were some changes made at the request of the assembled expert panel. The experts then gave their individual first-round subjective estimates of the lowest plausible, highest plausible and best estimate of probabilities of interest, 16 combinations in all.

The subjective estimates were collected and data entered in anonymised form to produce graphs for each question showing each expert's estimates and upper and lower bounds together with the group mean for each question side-by-side on a graph. A facilitated discussion, led by Dr. Anca Hanea, of results followed and each expert had the opportunity to contribute and give reasons for their estimates. Discussion was encouraged especially when different opinions on values were depicted in the graphs, but also when agreement was observed. During the discussions experts were encouraged to note down anything that they considered to be significant new information or compelling argument, especially if it had altered their thinking about a question.

After the completion of the discussions, the experts gave a second set of individual, subjective estimates for the same questions, having their original estimates

returned to them for comparison. Finally, an on-line protocol was discussed for running calibration questions, where the experts would be given several days to fill in the first round estimates, have an on-line facilitated discussion and then give a second round of estimates for the calibration questions. The experts indicated that they understood the value of this element of the elicitation to the final results and all agreed to participate.

Second-round estimates were received from nine experts who had provided first-round estimates. The measures of performance we considered are:

- The Brier score (per question, per expert)—scores close to 0 are good
- The average Brier score (per expert)—scores close to 0 are good (a big score corresponds to poor performance; a 0.5 score can be achieved by setting all answers to 0.5)
- The length of the uncertainty interval (per question, per expert)—small scores are better
- The calibration term of the Brier score (one number per expert calculated from all questions)—smaller scores are better
- Relative informativeness (one score per expert calculated from all answers)—departure from the [0.5 0.5] distribution—larger scores are better

The differences in scores are not significant. This means that the original questions can be combined with equal weighting. Following this, the experts' judgements were aggregated mathematically.

Now we have arrived at the first versions of the DBN for pollinator abundance. Of course, this may need further refinement once the functionality of the composite system is calibrated—see the next section. The probabilistic output of this pollination sub-module provides the CPTs needed for part of the components of the overarching food system, the "Farming" module, represented in Fig. 17.1. Once all the components of the "Farming" model have been populated with probabilities in analogous ways, the "Farming" module will have a fully specified set of CPTs conditional on its input variables, here listed under the broad heading of Economy and Weather & Environment.

Finally, this process needs to be repeated for each of the vertices in Fig. 17.1. We have then elicited a probability model of features in the process needed to score various policy options. The population of this model is obviously a massive task. However, our small team has made significant strides in producing this IDSS which we plan to deliver in prototype form in 2 years time.

17.5 Communicating the Results

Having successfully completed the construction, quantification and diagnostic checks on the IDSS, the final and very important step is to make its outputs available to the decision centre in a way that they find accessible, useful and compelling. There are no set rules for how to do this; each decision centre and

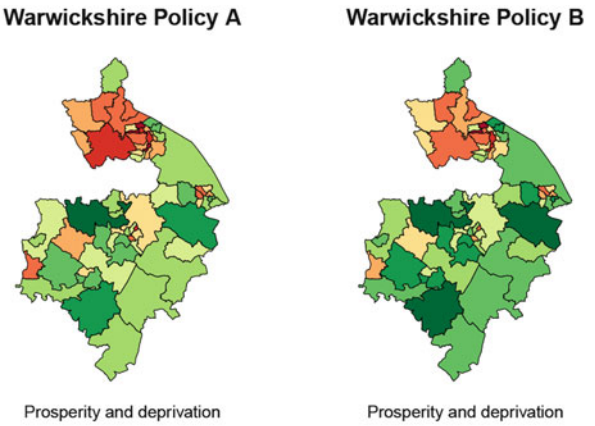


Fig. 17.13 UK food security: an illustration of the use of regional maps for decision support. Here an indicator of prosperity shows deprived areas in *red*. A clustering of deprivation (*left map*) is a risk factor for social unrest, such as food riots. Therefore, policies which specifically reduce and fragment large areas of poverty (*right map*) are to be preferred

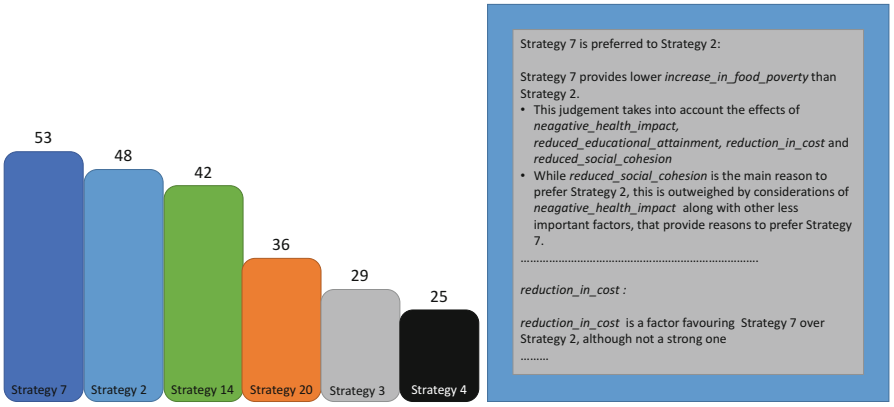


Fig. 17.14 Often, a straight comparison of the top-scoring candidate policies is sufficient. A natural-language report explaining in simple terms what has led to the ordering of one policy ahead of the next will aid the decision-centre's deliberations

domain will be different. In UK local government, historic data is often displayed on the regional map since decision-makers understand the social, political and administrative geographies within their region very well and the results they see are immediately contextualised in their thinking. One option, then, is to follow suit with the outputs of the IDSS, so that the effects of the various policies can be compared, including as they vary over time, within the domain context in a familiar way. The comparison of Policies A and B in Fig. 17.13 shows a stronger effect on the northernmost region, known to be an area of deprivation, so knowing where the effect is strongest can form part of the decision support. In other contexts, a simple side-by-side comparison of utility scores is required (Fig. 17.14). In all cases a natural language output explaining why one policy is scored higher than another is valuable to support human decision-making.

17.6 Quality Control of Integrating systems, Diagnostics and Robustness

In any complex model it is absolutely critical that once built, the integrating system produces credible composite forecasts of the processes of interest. This is true of each subcomponent, component and of the whole IDSS itself. Even when the overarching system is requisite to everybody and individually each component appears plausible, at least from a theoretical point of view it is quite possible that vital components or dependencies in the structure have been missed.

Our IDSS systems are usually intended for use by policymakers within governmental agencies or industrial companies who may not include expert statisticians. For example, our food IDSS is being co-created with local councils within the UK. It is therefore imperative that we build into the software on-line diagnostics and robustness alerts, so that red flags instantly appear if there is a potential error in the information entered or if an outcome is predicted which is infeasible. This will enable the user of the system to note any discrepancies and account for any additional uncertainty they may obtain due to these causes.

However, because the integrating system has been modelled so that it is probabilistically coherent, standard methods for checking the robustness and diagnostically checking any probabilistic system can be devised to address this domain. Here we simply have a massive system, but the underlying suites of methodologies are the same.

There are various methods we can use here to check out whether the integrating system is fit for purpose and if not to modify it so that it is. These are often called *diagnostic checks*. Here we illustrate quality checks that we will perform on the Food IDSS, once fully populated, which we inherited from other modelling activities.

Perhaps the most important consistency checks are ones that ensure that the IDSS works *predictively* well. This is because it is these distributions which will determine the way the centre scores various options open to it. The easiest element of this is to check one-step ahead forecasts with observations that are actually seen. This can be performed by retrospectively using the methods on a suitable number of past time points, using informative forecasts available at the time. These can be constructed retrospectively and the system run against this systematically, for example over or under estimation of attribute vectors can then be identified.

We can also compare future predictions from our model to predictions that the experts forecast. This would enable us to use any new information in the formulation of the model and would also highlight any discrepancies which occur, for example our model may predict certain outcomes which the expert knows will never occur. This can then be fed back into the system, helping to improve future forecasts.

Diagnostic checks for BNs have been studied over the years, and we shall briefly discuss three in particular which were introduced by Dawid et al. (1999) and further discussed in Cowell et al. (2007) as a way of quantifying the difference between the

specified prior and the data. We shall briefly discuss three of their diagnostics here: a parent-child monitor, a node monitor and a global monitor

1. The *Parent-Child Monitor* examines the forecasting capability of the conditional probability between a parent and child node. More specifically, if we can observe the parent node, we can use the parent-child monitor to quantify the performance of the BN when predicting the outcome of the child nodes.
2. The *Node Monitor* quantifies the performance of predictions on a specific node, given all available evidence and can be categorised into two subsections: unconditional node monitors and conditional node monitors.
3. The *Global Monitor* determines the overall performance of the graphical model.

These diagnostic checks all aim to quantify how well the model is performing after the data is observed. This measurement of performance is done using formulae which form proper scoring rules.

Any real probabilistic system has a degree of misspecification in it, either within its assumed structure or the specification of the probabilities within it. Of course, one way of addressing this is to perform a one at a time sensitivity analysis. However, when examining the robustness of the types of massive system we consider here, limited information about robustness can be gleaned in this way. Rather it is better to consider the robustness of the outputs to perturbations of the whole system. One spin-off of a formal analysis of this type is that we can discover *before* the BN is fully elicited that the precision of probability specifications in parts of the system can have little impact on the process *however* we fill in the CPTs. In this case little effort needs to be expended on these elements of the process.

We note that if we are deciding between two models, Bayes Factor methods can be utilised very simply to appreciate how much gain can be obtained in explanatory power from using a more complicated model rather than a simpler one. In our context here, all things being equal, we would advise the choice of a simpler model over a complex one if the gain in using the complex one is marginal. These techniques are widespread for standard graphical models, see Dawid et al. (1999), Korb and Nicholson (2011), Smith (2010), and are currently being examined in this specific context by two of these authors. Early results suggest that relative scores are most affected by misspecification of input distributions and structural features close to the attribute vector of the graph in the BN specification of the process.

17.7 Conclusions

Throughout this chapter we have introduced the idea of Integrated Decision Support Systems (IDSSs) to provide coherent and transparent decision support for complex systems. Although IDSSs can be applied within a wide range of overarching frameworks, we have focussed solely on probabilistic graphical models known as Bayesian networks (BNs). BNs have the advantage that they are a visual representation of the complex problem and therefore can be easier for non-statisticians, such as the problem-owner, to understand.

We have presented our ongoing case study which provides decision support for decision makers interested in enacting local policies regarding food poverty within the UK. We have discussed two separate features of this IDSS for comparison and clarity: the overarching food poverty model which aims to reproduce the cost of an average household weekly shopping basket, and the pollinator subsystem which forecasts pollinator abundance with regards to UK crop pollination. We have demonstrated the iterative process of identifying the variables of interest and the structure of the relationship between them. Iteration of the process also ensures that the experts and decision centre are fully involved at every stage of the creation process giving them ownership of the finished DSS and are therefore more likely to use and update it. Due to the nature of complex systems it is common to find that the model has hierarchical, interrelated components, however we have shown here that many of the techniques for creating and populating the model can be used on the overarching framework as well as the subsystems. In addition to describing how to design and structure the model, we have provided an in-depth guide on how we executed the elicitation workshop which populated the conditional probabilities in part of the pollinator subsystem and results which entailed.

We have demonstrated that building a fully probabilistic IDSS is feasible and been able to communicate some of our experiences in engaging in this activity. We hope that others will follow our lead and start to produce similar tools to address other large scale decision analyses. We believe that surmounting the challenges of implementing such large scale tools will be increasingly important in the coming years.

References

Blaauw BR, Isaacs R (2014) Flower plantings increase wild bee abundance and the pollination services provided to a pollination-dependent crop. J Appl Ecol 51(4):890–898

Caminada G, French S, Politis K, Smith JQ (1999) Uncertainty in RODOS. Doc. RODOS(B) RP(94) 05

Collier RA (2009) Identify reasons why food security may be an issue requiring specific attention. DEFRA Research Project Final Report

Cowell RG, Verrall RJ, Yoon YK (2007) Modeling operational risk with Bayesian networks. J Risk Insur 74(4):795–827

Datta S, Bull JC, Budge GE, Keeling MJ (2013) Modelling the spread of American foulbrood in honeybees. J R Soc Interface 10(88). doi:10.1098/rsif.2013.0650

Dawid AP (2001) Separoids: a mathematical framework for conditional independence and irrelevance. Ann Math Artif Intell 32(1–4):335–372

Dawid AP, Cowell RG, Lauritzen SL, Spiegelhalter DJ (1999) Probabilistic networks and expert systems. Springer, New York

DEFRA (2014) The National Pollinator Strategy: for bees and other pollinators in England

DESA U (2015) World population prospects: the 2012 revision, key findings and advance tables. Working paper no. ESA/P/WP. 227. United Nations Department of Economic and Social Affairs, New York, Population Division

Edwards W, Miles RF, Von Winterfeldt D (2005) Advances in decision analysis. Cambridge University Press, Cambridge

French S, Smith J (2016) Decision analytic framework for a decision support system for nuclear emergency management. In: UK success stories in industrial mathematics. Springer International Publishing, Berlin, pp 163–169

French S, Maule J, Papamichail KN (2009) Decision behaviour, analysis and support. Cambridge University Press, Cambridge

Gonzalez-Ortega J, Radovic V, Rios Insua D (2018) Utility elicitation. In: Dias LC, Morton A, Quigley J, Elicitation: The science and art of structuring judgment. Springer, New York

Gooding P (2016) Consumer price inflation: the 2016 basket of goods and services. Office for National Statistics

Hanea A, Burgman M, Hemming V (2018) IDEA for uncertainty quantification. In: Dias LC, Morton A, Quigley J, Elicitation: the science and art of structuring judgment. Springer, New York

Hartley D, French S (2018) Elicitation and calibration: a Bayesian perspective. In: Dias LC, Morton A, Quigley J, Elicitation: The science and art of structuring judgment. Springer, New York

Howard RA (1988) Decision analysis: practice and promise. Manag Sci 34(6):679–695

Howard RA (1990) From influence to relevance to knowledge. In: Oliver RM, Smith JQ (eds) Influence diagrams, belief nets and decision analysis. Wiley, New York, pp 3–23

Johnson S, Fielding F, Hamilton G, Mengersen K (2010) An integrated Bayesian network approach to Lyngbya majuscula bloom initiation. Mar Environ Res 69(1):27–37

Keeney RL, Raiffa H (1993) Decision with multiple objectives: preferences and value trade-offs. Cambridge University Press, Cambridge

Koster JT (1996) Markov properties of non-recursive causal models. Ann Stat 24(5):2148–2177

Korb KB, Nicholson AE (2011) Bayesian artificial intelligence. CRC press, Boca Raton

Lagi M, Bertrand KZ, Bar-Yam Y (2011) The food crises and political instability in North Africa and the middle east. arXiv preprint:1108.2455

Leonelli M, Smith JQ (2015) Bayesian decision support for complex systems with many distributed experts. Ann Oper Res 235(1):517–542

Leonelli M, Smith JQ (2013a) Using graphical models and multi-attribute utility theory for probabilistic uncertainty handling in large systems, with application to the nuclear emergency management. In: 2013 IEEE 29th international conference data engineering workshops (ICDEW), April. IEEE, New York, pp 181–192

Leonelli M, Smith JQ (2013b) Dynamic uncertainty handling for coherent decision making in nuclear emergency response. In Proceedings of the winter meeting of the ANS

Lonsdorf E, Kremen C, Ricketts T, Winfree R, Williams N, Greenleaf S (2009) Modelling pollination services across agricultural landscapes. Ann Bot 103:1589–1600

Norsys (1994–2016). Netica. Norsys

Oates CJ, Smith JQ, Mukherjee S (2016) Estimation of causal structure using conditional DAG models. J Mach Learn Res 17(54):1–23

ONS (2013) Consumer price indices: a brief guide

Pearl J (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, San Francisco

Pearl J (2000) Causality: models, reasoning and inference. Cambridge University Press, Cambridge

Phillips LD (1984) A theory of requisite decision models. Acta Pschol 56:29–48

Puch RO, Smith JQ (2002) FINDS: a training package to assess forensic fibre evidence. In: Coella CAC, de Albornoz A, Sucar LE, Battistutti OS (eds) Advances in artificial intelligence. Springer, Berlin, pp 420–429

Queen CM, Smith JQ (1993) Multi-regression dynamic models. J R Stat Soc B 55(4):849–870

Rader R, Bartomeus I, Garibaldi LA, Garratt MPD, Howlett BG, Winfree R, Cunningham SA, Mayfield MM, Arthur AD, Andersson GK, Bommarco R et al (2016) Non-bee insects are important contributors to global crop pollination. Proc Natl Acad Sci 113(1):146–151

Smith JQ (2010) Bayesian decision analysis: principles and practice. Cambridge University Press, Cambridge

Smith JQ, Barons MJ, Leonelli M (2015a) Coherent inference for integrating decision support systems, arXiv preprint:1507.07394

Smith JQ, Barons MJ, Leonelli M (2015b) Decision focused inference on networked probabilistic systems: with applications to food security. In: Proceedings of the joint statistical meeting, pp 3220–3233

Chapter 18
Expert Elicitation to Inform Health Technology Assessment

Marta O. Soares and Laura Bojke

Abstract In the face of constrained budgets, unavoidable decisions about the use of health care interventions have to be made. Decision makers seeking to maximise health for their given budget should use the best available information on effectiveness and cost-effectiveness, and for this purpose they may use a process of gathering and combining existing evidence in this context called Health Technology Assessment (HTA). In informing decisions, utilising HTA, expert elicitation can provide valuable information, particularly where evidence is missing, where it may not be as well developed (e.g. diagnostics, medical devices, early access to medicines scheme or public health) or limited (insufficient, not very relevant, contradictory and/or flawed). Here, formal methods to elicit expert judgements are preferred to improve the accountability and transparency of the decision making process, in addition to the important role in reducing bias and the use of heuristics. There have been a limited number of applications of expert elicitation in health care decision making, and in part this may be due to a number of methodological uncertainties regarding the applicability and transferability of techniques from other disciples, such as Bayesian statistics and engineering, to health care. This chapter discusses the distinguishing features of health care decision making and the use of expert elicitation to inform this, drawing on applied examples in the area illustrating some of the complexities and uncertainties.

18.1 Introduction

There are many factors required to achieve health, or more specifically good health, one of them being healthcare. Healthcare is viewed as an intrinsic right of every individual and its consumption confers benefits to everyone in a population, either for altruistic reasons or through externalities such as immunization (Folland et al. 2013).

M.O. Soares (✉) • L. Bojke
Centre for Health Economics, University of York, York, UK
e-mail: marta.soares@york.ac.uk

L.C. Dias et al. (eds.), *Elicitation*, International Series in Operations Research & Management Science 261, https://doi.org/10.1007/978-3-319-65052-4_18

The conditions under which health care is supplied are very different from the perfectly competitive model, and it has been argued that healthcare cannot be provided efficiently through free market mechanisms. As such, in many countries, the government steps in to provide healthcare either funded via social insurance, general taxation or hypothecated tax. Systems to provide health care to a population differ between countries, and are largely influenced by the specific social and economic conditions. In some countries there is still little government involvement and the majority of health care is funded through out of pocket payments. This is more typical in less developed countries and a significant degree of health inequalities exists as a consequence, with only those on higher incomes being able to afford access to all required healthcare.

Even in systems where healthcare is centrally funded, such as the UK, Canada and France, budget constraints will mean difficult decisions have to be made regarding what can and cannot be funded. Without explicit consideration of the budget constraint, the cost of healthcare would increase as consumers demand new pharmaceuticals and interventions, and industry develops these products to meet these needs.

In making these unavoidable decisions about the use of healthcare interventions, the bodies with responsibility for such decisions must determine the objective function of the healthcare system it is informing. In the UK, the objective function, as interpreted by the National Institute for Health and Clinical Excellence (NICE), is to maximise health and therefore the expectation is that new interventions will lead to better health. In many countries, equity of access may be an equally important criterion (Romanow 2002). There may be circumstances where an intervention that is seen to improve the health of particular groups of individuals to a greater extent than another group of individuals, for example on the basis of income, will be adopted so as to achieve a 'fair' distribution of health (vertical equity) (Culyer and Wagstaff 1993). Regardless of the objective function, where resources are limited, these benefits must be put into context considering the resources required to generate them, as any additional costs incurred will impact on access to healthcare for other patients, and thus potentially health is foregone. The consideration of both the benefits and costs of a health intervention, or competing health interventions, is referred to as cost-effectiveness analysis (Bryan et al. 2007) and the process of assessment is called Health Technology Assessment (HTA).

18.2 Representing Uncertainty in Adoption Decisions

A key feature of HTA is that it is unlikely that a single piece of evidence is entirely informative, for example a trial is unlikely to capture all the costs and benefits of all competing interventions over a sufficient time horizon (Sculpher et al. 2006). Also, in supporting an accountable and transparent decision making process, it is

essential that decisions are grounded on comprehensive evidence. Clinical evidence is often considered to be of highest quality if drawn from all available randomised controlled trials. This may be supplemented by longer term observational studies, surveys and real world studies, particularly when evidence on resource use and quality of life is required. This evidence needs to be synthesised to allow total costs and health benefits associated with competing interventions to be estimated. For this purpose, cost-effectiveness analysis employs decision modelling methods that define mathematical relationships between a varied set of input parameters, in a way that describes aspects of the history of the disease of interest and the consequences of the interventions (Drummond et al. 2005).

However, it is often the case that the assessment of important input parameters in decision models is supported by only limited empirical data; for example, the evidence may not be available on 'final' outcomes (e.g. cancer products licensed on evidence of progression-free survival). Due to uncertainty in the assembled evidence and/or underpinning assumptions required for analysis, the expected cost-effectiveness of an intervention is often not known with certainty, introducing uncertainty in the decision (Griffin et al. 2011). Indeed, there may be circumstances in which the inputs required to a decision model are missing entirely.

An assessment of uncertainty is required as models are typically complicated, with non-linear relationships between inputs and outputs (Griffin et al. 2011), and estimated expected cost-effectiveness can be biased if uncertainty in model inputs is not reflected in the analyses. Also, additional evidence can reduce uncertainty and provide a more precise estimate of cost-effectiveness. By explicitly quantifying uncertainty, it is possible to assess the potential value of additional evidence, inform the types of evidence that might be needed, and consider restricted use until the additional evidence becomes available (Claxton 1999).

Given the incomplete nature of evidence often used to support decision making in healthcare, expert judgements are often needed for a decision to be reached. In an accountable decision making process, these judgements should be made explicit and incorporated transparently into the decision making process, following the Bayesian view of decision making (Briggs 1999).

This creates a *primia facie* case for considering the use of expert elicitation. To date, formal expert elicitation has only been used to a limited extent in healthcare decision making, perhaps due to the lack of clear guidance of what methodologies may be appropriate in this context (Sullivan and Payne 2011). However, in informing decisions, expert elicitation can provide valuable information, particularly where evidence is missing, where it may not be as well developed (e.g. diagnostics, medical devices, early access to medicines scheme or public health) or is limited (insufficient, not very relevant, contradictory and/or flawed). Where expert judgements are used, analysts should strive for structured and formal methods of elicitation to reduce bias and the use of heuristics, and improve the quality and confidence in this evidence.

18.3 Distinguishing Features of Health Care Decision Making and Requirements for Expert Elicitation

Whilst there is a lack of guidance on the appropriate methodology for expert elicitation in health care decision making, methods have received significantly more attention in other disciplines, including engineering and Bayesian statistics (Babuscia and Cheung 2014). These developments may be useful in suggesting a range of possible methods for elicitation in health care. However, there are a number of features of healthcare decision making that distinguish it from these other disciplines and, thus, currently available guidelines and protocols for expert elicitation need to be subject to further consideration. This is particularly true where such protocols describe multiple options for particular elements of the design process, for example the choice between consensus and mathematical approaches.

The first distinguishing feature of health care decision making is the need for consistent use of methods. This warrants uniformity and transparency between decisions made in potentially very different contexts (e.g. public health screening compared to surgical intervention). Jurisdictions using HTA already define a set of methodological principles for effectiveness and cost-effectiveness evaluations—many define a reference case, departures from which need to be carefully justified. The use of elicitation in this context should thus follow a standardised set of principles, ideally under the form of an elicitation protocol sufficiently flexible to ensure it is useful across evaluations. Given the decision making context, such a protocol should have normative input from decision makers. It should focus on all aspects of elicitation: those related to the design (what and how to elicit), conduct (the role of the facilitator) and analyses (how to pool judgements).

It is also important that elements of the protocol for elicitation are tailored to the specificities and requirements of health care. For example, likely substantive experts in this area may be health professionals, which may not possess normative skills. Normative experts have been defined as those whose skills lie in elicitation methods or have good numeracy skills (O'Hagan et al. 2006). The lack of normative skills may restrict the method of elicitation used. Such experts may, for example, find it difficult to grasp the concept of quantiles required for the bisection method (O'Hagan et al. 2006).

For HTA, it is also important that the elicited information represents how uncertain experts are about the current state of knowledge regarding a parameter of interest, and in this way reflect the imperfect knowledge they have (referred to as epistemic uncertainty). An important concern is that uncertainty is misrepresented in the judgements elicited. One of the reasons for this is that experts, when reflecting on their own experiences, may include some level of variability. Variability refers to the fact that individual responses to an intervention will differ between patients with the same observed characteristics within the population. Another reason for uncertainty to be misrepresented may relate to systematic biases well-known in the elicitation literature, such as overconfidence; over-extremity (where bias affects extreme values) or discrimination (when the expert cannot distinguish likely events from those less likely) (O'Hagan et al. 2006, Gigerenzer and Hoffrage 1995).

This may have important implications for the design and conduct of elicitation; for example, consensus approaches have been shown to produce overconfident statements (Grigore et al. 2016) and this may be a consideration when choosing the method of aggregation.

Also, for HTA, it is generally accepted that the elicited information needs to reflect the range of reasonable judgements that may be expressed across experts (between-expert variation). This is for two reasons. Firstly, because the quantities of interest are generally never known with certainty and thus trying to appropriately reflect not just the individual's own epistemic uncertainty but also any additional variation between experts helps decision makers understand the full extent of the uncertainty. Secondly, experts may be exposed to different settings or case-mixes. This heterogeneity in the subpopulations that experts observe will be reflected in the judgements elicited. Heterogeneity refers to individual differences in, for example, response to treatment, that can be associated with differences in characteristics of the patients or their disease. It is recognised that gathering opinions from multiple experts is essential to quantify the current level of knowledge for a parameter, including any uncertainty (O'Hagan et al. 2006), however there is little guidance on how to aggregate these in a way that takes into account, in an appropriate way, what is known or unknown about heterogeneity.

18.4 Methods for Expert Elicitation in Healthcare Decision Making

A recent overview of methods for eliciting expert opinion to inform HTA, conducted to inform a funding stream in the UK (Gosling 2014), identified no specific guidance in the context of HTA, but did identify three generic protocols for the elicitation of uncertain quantities: the Sheffield elicitation framework (SHELF), a modified Delphi scheme extended from SHELF (European Food Safety Authority 2014) and Cooke's classical method (Cooke 1991).

The Sheffield elicitation framework—see Chapter 4: "SHELF: The Sheffield Elicitation Framework" of this book (Gosling 2018)—is a package of materials aimed at standardising the information recorded as part of a consensus group elicitation exercise. It provides templates for: pre-session briefing notes, pre-session pro forma to be sent out with the briefing notes to experts, elicitation records regarding the context and purpose of the elicitation, and records for the elicitation of each probability distribution. It also provides an R package for fitting distributions using least squares. In terms of elicitation methods, SHELF considers only univariate techniques, specifically SHELF offers the option of eliciting probabilities (for two ranges of values), eliciting percentiles (specifically quartiles through the bisection method and terciles), and the roulette method. It is designed to be applied in a face-to-face workshop led by a facilitator, and to reach consensus amongst participants, what is known as a behavioural approach.

The process however starts with individual experts making their own quantitative judgements, which are then linearly pooled by the facilitator and presented to the group to encourage discussions. Recently, a web-based interface for SHELF has been developed, called MATCH (Morris et al. 2014).

The Delphi iteration process proposed by the EFSA (2014) modifies the group stage of the SHELF protocol to be undertaken remotely through a Delphi iteration process. This entails that the experts' judgements (accompanied by a description of the rationale) are relayed back to all the experts individually (anonymously) and they are invited to review their judgements and revise if they deem appropriate. After two or more iterations rounds, the experts' individual probability distributions are averaged to provide the final aggregate distribution.

Contrary to SHELF and the modified Delphi, Cooke's classical method does not attempt to reach a consensus but uses mathematical elicitation coupled with linear pooling. It does, however, take into account the performance of each of the experts by using unequal weights. The weights are defined based on how accurately and precisely each expert answers a set of seed questions (where answers are known to the researcher but not to the expert). A software package using Cooke's method is available: EXCALIBUR. In addition to these three protocols, there exists a number of 'off the shelf' software packages for elicitation. Others have summarised these elsewhere (Expert Judgement Network 2016).

Despite SHELF (first released in 2008) and EXCALIBUR (latest version compiled in 2004) being freely available for a few years now, a recent review of applications reporting the use of formal methods of elicitation in the context of HTA has not revealed any examples of the use of these protocols (Grigore et al. 2013). Also, the review identified only 14 studies, which, in the context of the vast HTA literature, represents only marginal use. In the existing applications, all except one study used mathematical aggregation; three of the studies using mathematical aggregation explored the potential use of calibration, but justified not using it based on difficulty in choosing appropriate criteria. The most common methods of elicitation used were: the histogram method (equivalent to the roulette proposed within SHELF but using a fixed number of chips); the bisection method and the elicitation of percentiles (commonly 95% confidence intervals). The preferred aggregation method was unweighted linear pooling. Across applied examples, the methods of elicitation appear to be heterogeneous and the reasons underlying choice of methods are unclear.

The limited use of expert elicitation in an HTA context, and the limited use of existing protocols, suggests that the appropriateness of available methods and protocols may need to be considered specifically in the context of HTA. There also appears to be a need for guidance to homogenise the methods used across evaluations. Previous authors (Iglesias et al. 2016) do note that the quality of reporting for elicitation in health care is particularly poor, and recommend the development of guidelines for reporting, for example, in the form of a checklist.

18.5 Examples of Applications in Health Care Decision Making

This section describes the methods and results of two examples of formal elicitation exercises implemented to inform HTA. These help to demonstrate some of the practical and methodological challenges faced in this context.

18.5.1 Negative Pressure Wound Therapy (Soares et al. 2011)

Negative pressure wound therapy (NPWT), also known as topical negative pressure, is a medical device used to treat full thickness wounds such as severe pressure ulcers. It has been claimed that NPWT speeds healing and reduces infection rates and costs as well as assists in the practicalities of wound management; however, there is very little actual evidence for its clinical or cost effectiveness (Soares et al. 2013). NPWT is also a relatively expensive treatment used widely in the developed world; thus, it likely incurs a significant burden on health care resources. Therefore, there is a need to evaluate the cost effectiveness of NPWT and alternative treatments for its various indications, including severe pressure ulcers. Additionally, given the expected uncertainty surrounding the choice of treatment, it is important to explore whether investing in further research regarding the use of NPWT is worthwhile and, if so, what type of future research is most likely to offer the most value for money.

The study here used as an example aimed at evaluating the cost-effectiveness of NPWT in speeding up the healing of grade 3 or 4 pressure ulcers and explored whether further research regarding the use of NPWT was worthwhile and, if so, what type of future research would be likely to offer the most value for money. The latter analysis is an extension of the cost-effectiveness framework to consider decision making under uncertainty, using quantification of the consequences of making the wrong decision to establish the expected value of collecting further evidence (Claxton 1999).

To establish cost-effectiveness, a mathematical decision model was used that evaluated the long term costs and health effects of NPWT and relevant alternatives (dressings such as spun hydrocolloid—HC; alginate—ALG; and foam dressings—F). The model described how patients were expected to transit between three health states (unhealed, healed and dead) and also distinguished the means to healing (if through closure surgery or secondary healing) and the occurrence of complications and discontinuation from treatment. A review of the literature was conducted to identify evidence to inform the model, but the evidence-base was found to be limited and sparse. However, NPWT and comparators were used extensively within the NHS in the UK and excluding such experience could misrepresent the current level of knowledge regarding these treatments. Thus, a formal exercise was designed to systematically capture experts' knowledge and uncertainty around the treatment and progression of severe pressure ulcers. Evidence was collected

in the form of probabilistic judgments around the speed of transitions between health states (transition probabilities) and related events (except those associated with mortality), including beliefs about the impact of the alternative treatments on transition probabilities (relative effectiveness).

The authors identified a number of challenges in designing the elicitation exercise. Nurses were identified as substantive experts but were highlighted to have limited normative skills. The authors thus planned for a face-to-face meeting where substantial training could be delivered. Also, only parameters of binomial variables were elicited in order to simplify the task and abridge the training session. The method of elicitation used was the histogram method, for its intuitiveness. In specifying the quantities to elicit, a series of considerations were taken. Firstly, the quantities of primary interest were expressed in terms of others whose distribution(s) were thought to be easier to elicit. All quantities were thus directly observable. On some of the quantities there was available (though sparse) data, and the quantities elicited were thus defined keeping in mind the need for synthesising the elicited evidence together with the existing evidence.

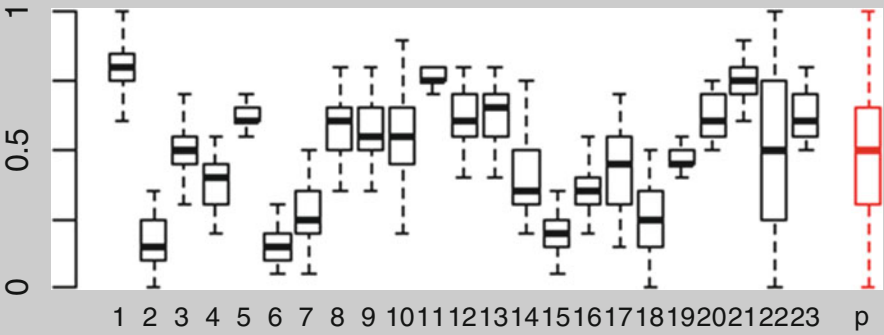
In total, twenty-three nurses attended and completed the elicitation exercise. Each expert answered more than 30 questions, 18 of which were uncertain quantities elicited through the histogram technique. Experts were asked to elicit their beliefs individually and were discouraged from interacting (a mathematical approach to elicitation was used); there was no attempt to achieve consensus.

For illustrative purposes, we will provide more detail on the methods and results of the elicitation of judgements over the relative effectiveness on healing of NPWT compared to a reference treatment, Hydrocolloid (HC). In this context, experts were asked to first elicit the probability of healing with the reference treatment (HC) at 6 months (question 1 in Box 1). A box plot summary of each expert's response is shown in Box 1. Experts used a variety of distributional shapes to characterise their strength of belief, showing that they felt comfortable in using the histogram method. Results show a high level of variability between experts, with the median ranging between 5% and 75%. This variation is far from unexpected: individual beliefs will inherently differ from each other, and experts were recruited from a variety of clinical contexts (e.g. primary care vs. hospital, specialist vs. community nurses). From a decision making perspective, it is desirable that all views are represented and the implications of between-expert variation explored.

Box 1: Elicitation

Six months after starting treatment with HC, what proportion of patients who are alive do you think would have a healed reference ulcer? This is regardless of whether patients are still receiving treatment with HC at this 6 months point.

(continued)



The judgements of individual experts were elicited using the histogram technique. Results for each individual expert (from expert 1 to expert 23, indicated in the x-axis) are summarised here using box-plots. The red box-plot identified on the x-axis with the letter P reflects the linearly pooled judgements from the 23 experts.

To elicit the effectiveness of the active treatment (NPWT), the authors elicited absolute effectiveness but used an approach based on conditional independence. The expert was asked to assume that the value they believe best represented their knowledge about the effectiveness of the comparator treatment, HC, was true. The same experts' mode of the distribution elicited for HC was used to represent this value. The absolute effectiveness of NPWT and its uncertainty was then elicited, with the expert bearing in mind the conditioning value for HC—see Box 2. The elicited information was *a posteriori* converted onto a relative effectiveness measure (a log hazard ratio) using the conditioning value, and only then pooled across experts. The relative effectiveness measure obtained in such a way assumes independence of the absolute effectiveness of the comparator.

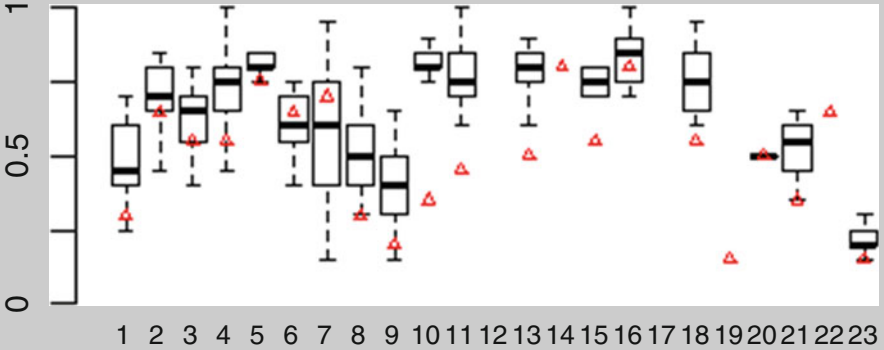
Experts expressed different views over the relative effectiveness of NPWT—Box 2. The majority believed that NPWT would achieve a higher proportion of healing at 6 months than HC (e.g. expert 1 and 11), with many experts retaining the possibility of it being equally or less effective (e.g. expert 1 but not expert 11); expert 7 was very uncertain, but indicated that the proportion healed with NPWT was more likely to be lower than with HC (with the elicited median being below the conditioning value) and expert 20 indicated a strong belief that NPWT and HC were equal in terms of healing at 6 months.

Box 2: Elicitation

Your strongest belief was that «mode of elicited distribution for HC (Box 1)»% of patients had a healed ulcer 6 months after starting HC. Assume that

(continued)

this value is true. Six months after starting treatment with topical negative pressure therapy, what proportion of patients who are alive do you think would have a healed reference ulcer?



The judgements of individual experts were elicited using the histogram technique. Results for each individual expert (from expert 1 to expert 23, indicated in the x-axis) are summarised here using box-plots. Red triangles indicate the reference value used for each expert (the mode of the response given to the question in Box 1). Also note that these histograms were preceded by filter questions: "Think of UK patients with at least one debrided grade 3 or 4 pressure ulcer (greater than 5 cm^2 in area). Assume that the deepest ulcer was treated with spun hydrocolloid/hydrofiber (HC) as the primary contact layer and a certain proportion healed. Do you think the proportion healed would be different if instead of a HC patients were treated with NPWT? ". Only experts suggesting treatments to be different proceeded to elicit relative effectiveness and its uncertainty. In this way we avoided eliciting through the grid when experts claimed to be fully certain that the treatment were similar or to be fully uncertain about this. In this case, we assumed the density to be fully allocated to the reference value or to be equally distributed by the range of values available, respectively.

In a separate stream of work, evidence from the literature was systematically sought to inform these same effectiveness parameters of the decision model (Soares et al. 2014). The search identified one randomised controlled trial (RCT) investigating NPWT, and 11 investigating dressings. The data from these trials were linked within an evidence network and synthesised. Because most links in the network were informed by a single study and the number of healing events in some trials was small or zero, it was not possible to obtain inferences on one of the comparators, foam (F). Therefore, in pooling the existing evidence an assumption was made that there was a common effect across dressing treatments (implemented using a random effect). Results of this analysis are presented as log hazards or

Table 18.1 Results

	Existing evidence Mean [95% CrI]	Elicited evidence Mean [95% CrI]	Existing and elicited evidence collated Mean [95% CrI]
Log hazard of healing for HC	−3.95 [−4.50 to −3.46]	−3.74 [−5.96 to −1.52]	−3.97 [−4.59 to −3.46]
Relative effectiveness in relation to HC			
Log hazard ratio of healing for F	0.03 [−1.97 to 1.86]	−0.96 [−6.32 to 4.40]	−0.91 [−2.14 to 0.21]
Log hazard ratio of healing for ALG	−0.19 [−1.76 to 1.13]	0.003 [−0.63 to 0.64]	−0.27 [−2.12 to 1.57]
Log hazard ratio of healing for NPWT	0.18 [−2.17 to 2.63]	0.45 [−0.66 to 1.56]	0.47 [−1.18 to 2.10]

log hazard ratios in Table 18.1, with the ratios expressed using HC as a reference treatment. Note that the assumption of common effect across dressings meant that, in this analysis, F is assumed to be as effective as other dressings.

The second column in Table 18.1 shows the elicited evidence pooled across experts on the log hazard scale. Experts, as a group, expressed different judgements over the expected relative effectiveness of the treatments, suggesting that foam dressings were expected to be less effective than the comparator (HC), alginate dressings to have the same effectiveness and NPWT to be slightly beneficial. This was a reflection of the individual replies of experts.

The two sources of evidence were collated using Bayesian updating, to generate a combined posterior distribution (Table 18.1, third column) incorporating both the prior distribution (elicited beliefs) and observations from the existing evidence. In this scenario, the use of elicited data allowed the assumption of a common effect across treatments to be dropped, and thus treatment effects for each dressing were able to be estimated without information being exchanged across dressings.

When these findings were included in the cost-effectiveness analyses (results not shown here but detailed in Soares et al. 2011; Soares et al. 2013), they had important implications as they allowed ruling out foam (F) as a relevant alternative to NPWT in clinical practice, and also determining that further research on foam was unlikely to be worthwhile. When considering all sources of evidence, NPWT was evaluated as cost-effective at expected values, but there was substantial uncertainty with a probability of error of approximately 55% (in which case one of the other treatments would have been better). The results of value of information analysis indicate that further research is worthwhile and established the most efficient design as a 3-arm trial comparing NPWT with ALG and HC, with a 2 year follow-up and recruiting with approximately 400 patients.

18.5.2 Photo Acoustic Mammography (PAM) (Haakma et al. 2014)

During the development of new imaging technology, Photo Acoustic Mammography (PAM) for the diagnosis of breast cancer, an expert elicitation was conducted to inform the cost-effectiveness of PAM compared to the currently available imaging technique, Magnetic Resonance Imaging (MRI). In addition, the expert elicitation aimed to establish its potential clinical value to guide further product development. Specifically the elicitation focused on generating priors, expressed as probability distributions, for the diagnostic performance of PAM (sensitivity and specificity). A mathematical approach to elicitation was used to allow full uncertainty in experts' beliefs and any between expert heterogeneity to be expressed.

Twenty radiologists were invited to participate in this study, although two were unable to attend. These were selected for the study using purposeful sampling, based on predefined characteristics such as expected level of knowledge and experience using MRI. The questionnaire was administered on a face-to-face basis, and a standardized script was used along with an Excel elicitation sheet designed specifically for the task. Eighteen radiologists specialising in examining MR-images of breasts first ranked tumor characteristics according to their importance in detecting malignancies. This was undertaken to allow the experts to refresh their knowledge on the features of a diagnostic imaging technique and start to explore how these concepts may relate to PAM. The intention was to reduce the possibility of bias in the elicited beliefs, including overconfidence and confirmation bias.

These tumour characteristics are identified from the BI-RADS (Breast Imaging–Reporting and Data System) classification system to grade breast lesions and include (1) mass margins; (2) mass shape; (3) mass size; (4) vascularization; (5) localization; (6) oxygen saturation; and (7) mechanical properties. First the experts were asked to express how important the tumor characteristics were in the examination of images. The importance of tumor characteristics was expressed by allocating 100 points between the characteristics. Following this, they were then asked how well MRI and PAM can visualize these characteristics by assigning each characteristic with a value between 0 and 100, where 0 indicates a low performance and 100 indicates a high performance. The expected performance of MRI and PAM was then determined by the following equations:

For MRI:

$$MRIp\left(tc_j\right) = \sum_{i=1}^{n} w_i{}^{*}\left(MRIp_i\left(tc_j\right)\right) \tag{18.1}$$

For PAM:

$$PAMp\left(tc_j\right) = \sum_{i=1}^{n} w_i\left(PAMp_i\left(tc_j\right)\right) \tag{18.2}$$

Table 18.2 Distribution of importance of tumor characteristics

	Mass margins	Mass shape	Mass size	Vascularisation	Oxygen saturation	Location mass	Mechanical properties
Mean	30.78	29.02	5.59	19.02	4.67	3.94	10.40
SD	8.48	12.73	6.09	10.95	5.68	4.71	7.62
95% CI	20, 43	15, 54	0, 18	5, 43	0, 14	0, 11	0, 26

Table 18.3 Calibration factors explored in the PAM application

Years of experience (overall weight assigned = 0.45)		Average number of MRI's examined per week (overall weight assigned = 0.45)		Examining MRI's in other areas (overall weight assigned = 0.1)	
Level	Score assigned	Level	Score assigned	Level	Score assigned
$X < 3$	1	$X < 5$	1	$X = 0$	1
$X \geq 3$	2	$5 \leq X < 10$	2	$X > 0$	2
		$10 \leq X$	3		

where (p) is the performance of each tumor characteristic (tc_j), and w_i accounted for the weight (w) of each individual expert (i).

The mean importance of each of the tumor characteristics with their standard deviations and 95% confidence intervals are shown in Table 18.2. Mass margins and mass shape are shown to be particularly important, with oxygen saturation and location mass seen as much less important in the examination of images.

In terms of the discriminatory ability of MRI and PAM in relation to these characteristics, MRI was thought to perform better at visualizing mass margins and mass shape, whereas PAM was thought to perform better at visualizing vascularization and mechanical properties.

Fourteen of these experts then expressed their beliefs about the true positive rate (TPR) and true negative rate (TNR) for PAM. Given that PAM is a new diagnostic device, and therefore experts have little or no practical experience using it, TPR and TPR were elicited relative to existing MRI data. Experts were asked to express the mode (the most likely value) as this was thought to be the most intuitive quantity. In addition experts expressed the lower, and the upper boundaries (95% credible interval). A probability density function (PDF) was then generated using the linear opinion pooling method in which weighting is applied to reflect the performance of individual experts. These weights were generated based on the clinical background of each expert. These characteristics were chosen to reflect substantiveness of experts (years of experience and average number of MRI images examined per week), and adaptive skills (examination of MR images in other areas). Characteristics and weights are applied are shown in Table 18.3 below along with the contribution each factor made towards the final weight (45%, 45% and 10% respectively).

There was considerable heterogeneity between experts (radiologists) in estimating the diagnostic performance of PAM. The overall probability density function indicated a sensitivity ranging from 56.1% to 86.9%, with a mode of 73.3%.

The specificity ranges from 48.1% to 78.2%, with a mode of 64.7%. Experts expressed difficulties estimating the performance based on limited practical experience, despite being asked to reveal their beliefs about the performance of MRI, where there is considerable practical experience, in the first instance. PAM is an early stage technology, for which only small-scale, experimental experience was available. However, it remains uncertain if this is a more general problem with providing estimations about technologies in the early stages of development or if this uncertainty is specific to the application in breast imaging. In the early stages of development, where experimental evidence is less developed, there is a strong case for the use of formal elicited judgements, however the lack of practical experience presents a problem for experts, and instead relies of their ability to use any adaptive skills (O'Hagan et al. 2006), i.e. the ability to transfer knowledge from one setting/technology to another. The method of weighting, based on clinical experience, is also contrary to other methods of weighting, such as Cookes classical method.

18.6 Conclusions and Requirements for Further Research

Around the world, decision making in health care is increasingly becoming explicit, accountable, evidence-based, and focused on an explicit normative framework defining an objective function as a metric of value, for example maximizing health for a given budget. Where initially many jurisdictions defined HTA processes focusing on medicinal products, there has been recently a move to expand these processes to evaluate interventions/programs in areas where experimental evidence may be difficult to collect or is not actively encouraged by the regulatory process, e.g. public health, diagnostics, genomics. Additionally, health care technologies are also being appraised earlier in the development pathway, especially as regulatory agencies such as the European Medicines Agency's (EMA) change the regulatory pathway in efforts to improve timely access for patients to new medicines, for example through the adaptive pathways approach.

In all these areas, the evidence supporting decision making processes may be less well developed. This implies that judgements are needed for a decision to be reached. In an accountable decision making process, these judgements should be made explicit and incorporated transparently into the decision making, in accordance with the Bayesian view on decision making. These judgements may be appropriately informed by one or more experts, in which case structured and formal methods of elicitation should be preferred to reduce bias and the use of heuristics, and improve the quality and confidence in this form of evidence.

Despite there being elicitation protocols proposed in the literature, there has not been, as yet, proper consideration of which protocols or elements of protocols are appropriate for healthcare decision making. Forming a view over methodology could form the basis for a reference case to be defined and used across evaluations and in this way ensue consistency. It is also important, alongside this, to consider

how formal expert elicitation would work alongside a HTA process, specifically those for which there are time constraints and possibly resource constraints. In some circumstances an optimal expert elicitation may not be achievable, in order for decisions to be made in a timely manner. It is therefore important to consider which elements of formal elicitation are necessary requirements for decision making in health care, and which elements, although methodologically optimum, offer a more marginal contribution. It may also be advantageous to consider which available software facilitates the process, or indeed if alternative software needs to be developed.

References

Babuscia A, Cheung KM (2014) An approach to perform expert elicitation for engineering design risk analysis: methodology and experimental results. J R Stat Soc 177(2):475–497

Briggs AH (1999) A Bayesian approach to stochastic cost-effectiveness analysis. Health Econ 8(3):257–261

Bryan S, Williams I, McIver S (2007) Seeing the NICE side of cost-effectiveness analysis: a qualitative investigation of the use of CEA in NICE technology appraisals. Health Econ 16(2):179–193

Claxton K (1999) The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. J Health Econ 18(3):341–364

Cooke RM (1991) Experts in Uncertainty. Oxford University Press, Oxford

Culyer AJ, Wagstaff A (1993) Equity and equality in health and health care. J Health Econ 12:431–458

Drummond M, Sculpher MJ, Torrance JW, O'Brien BJ, Stoddart GL (2005) Methods for the economic evaluation of health care programmes, 3rd edn. Oxford University Press, Oxford

European Food Safety Authority (2014) Guidance on expert knowledge elicitation in food and feed safety risk assessment. EFSA J 12(6):3734. http://onlinelibrary.wiley.com/doi/10.2903/j.efsa.2014.3734/epdf

Expert Judgement Network (2016) Selection of structured expert judgement software. http://www.expertsinuncertainty.net/Software/tabid/4149/Default.aspx. Accessed 05 June 2017

Folland S, Goodman AC, Stano M (2013) The economics of health and healthcare, 7th edn. Routledge, New York

Gigerenzer G, Hoffrage U (1995) How to improve Bayesian reasoning without instruction: frequency formats. Psychol Rev 102(4):684

Gosling JP (2014) Methods for eliciting expert opinion to inform health technology assessment. Medical Research Council. https://www.mrc.ac.uk/documents/pdf/methods-for-eliciting-expert-opinion-gosling-2014/. Accessed 05 June 2017

Gosling JP (2018) SHELF: the Sheffield elicitation framework. In: Dias LC, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgment. Springer, New York. (Chapter 4 in this book)

Griffin S, Claxton K, Palmer SJ, Sculpher MJ (2011) Dangerous omissions: the consequences of ignoring decision uncertainty. Health Econ 20(2):212–224

Grigore B, Peters J, Hyde C, Stein K (2016) A comparison of two methods for expert elicitation in health technology assessments. BMC Med Res Methodol 16:85

Grigore B, Peters J, Hyde C, Stein K (2013) Methods to elicit probability distributions from experts: a systematic review of reported practice in health technology assessment. PharmacoEconomics 31(11):991–1003

Haakma W, Steuten LM, Bojke L, IJzerman MJ (2014) Belief elicitation to populate health economic models of medical diagnostic devices in development. Appl Health Econ Health Policy 12(3):327–334

Iglesias CPTA, Rogowski WH, Payne K (2016) Reporting guidelines for the use of expert judgement in model-based economic evaluations. PharmacoEconomics 34(11):1161–1172

Morris DE, Oakley JE, Crowe JA (2014) A web-based tool for eliciting probability distributions from experts. Environ Model Softw 52:1–4

O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, David J, Jenkinson DJ, Oakley JE, Rakow T (2006) Uncertain judgements: eliciting experts' probabilities. Wiley, Chichester

Romanow RJQ (2002) Building on values: the future of health care in Canada, Final Report

Sculpher MJ, Claxton K, Drummond M, Mccabe C (2006) Whither trial-based economic evaluation for health care decision making? Health Econ 15(7):677–687

Soares MO, Bojke L, Dumville J, Iglesias C, Cullum N, Claxton K (2011) Methods to elicit experts' beliefs over uncertain quantities: application to a cost effectiveness transition model of negative pressure wound therapy for severe pressure ulceration. Stat Med 30(19):2363–2380

Soares MODJ, Ashby R, Iglesias C, Bojke L, Adderley U (2013) Methods to assess cost effectiveness and value of further research when data are sparse: negative pressure wound therapy for severe pressure ulcers. Med Decis Mak 33(3):415–436

Soares MO, Dumville JC, Ades AE, Welton NJ (2014) Treatment comparisons for decision making: facing the problems of sparse and few data. J R Stat Soc 177(1):259–279

Sullivan W, Payne K (2011) The appropriate elicitation of expert opinion in economic models. PharmacoEconomics 29(6):455–459

Chapter 19
Expert Judgment Based Nuclear Threat Assessment for Vessels Arriving in the US

Jason R. W. Merrick and Laura A. Albert

Abstract We demonstrate the use of extended pairwise comparisons for estimating the relative likelihood that a vessel approaching US waters contains a nuclear threat. We demonstrate an expert judgment based method consisting of a designed set of extended pairwise comparisons and parameter estimation for a predictive probability model using log-linear regression. Results are based on a proof-of-concept questionnaire completed by eight experts in port security. The model and parameter estimates obtained are used to demonstrate the type of predictions that can be obtained.

19.1 Introduction

Our nation's economic well-being is intrinsically linked with the success and security at our ports. International trade accounts for more than thirty percent of the United States economy. Ninety-five percent of international goods that enter this country come through one of our nation's maritime ports of entry, adding up to more than nine million containers every year (Ebeling 2009). We know that terrorist groups are trying to obtain radiological material to attack targets inside the United States with radioactive dispersal devices, or dirty bombs (Gardner 2003). Our maritime borders are one potential route for smuggling such a device into the country. Providing a multi-layered approach for port security, while not disrupting the flow of international trade, is challenging. There are enormous economic consequences when our nation's maritime port security system is compromised. The need to protect our nation from nuclear attacks, as well the release of radiological materials is an issue of vital national interest. Physical inspection and scanning with radiation portal monitors is one part of the solution (Merrick and McLay 2010). However, threat assessments from intelligence analysts and customs and

J.R.W. Merrick (✉)
Virginia Commonwealth University, Richmond, VA, USA
e-mail: jrmerric@vcu.edu

L.A. Albert
University of Wisconsin-Madison, Madison, WI, USA

L.C. Dias et al. (eds.), *Elicitation*, International Series in Operations Research & Management Science 261, https://doi.org/10.1007/978-3-319-65052-4_19

border patrol personnel adds an additional layer and allows inspection and scanning resources to be used more effectively.

Intelligence analysts are often asked to forecast significant effects and must do so based on limited data. This makes probabilistic reasoning the ideal mechanism for representing the level of uncertainty and for combining multiple assessments into a single forecast (Paté-Cornell 2002). Such events are contingent on either previous events or factors that describe the situation in which the event may occur. Clemen and Winkler (1999) review several models for combining experts' judgments of probabilities with the decision maker's prior information under the Bayesian aggregation framework developed in Morris (1974, 1977, 1983). It would seem natural to extend one of these techniques to incorporate the relevant factors. However, empirical research has shown that experts overestimate probabilities near zero (Cooke 1991). Research in judgment and decision-making suggests that decision makers do not weight rare events according to their actuarial chances of occurring. This is partly due to the use of a mental heuristic that we commonly apply when forming probability judgments.

The availability heuristic can be used when a probability of an event's occurrence is assessed based on the ease with which one can retrieve similar events from memory (Tversky and Kahneman 1973). It is often easier to recall instances of large classes than those of less frequent classes. External events and influences can have a substantial impact the availability of similar incidents, such as the media or particularly emotional events (Combs and Slovic 1979). Tversky and Kahneman (1983) suggest that the witnessing an accident will have a greater effect on a person's judgment of the probability of an accident than reading about it in a newspaper. Tversky and Kahneman (1973) describe an experiment where subjects were read lists of names of famous men and women and asked to assess whether there were more men or women on the list. Some subjects were given lists where the men were more famous and tended to respond that there were more men on the list. Some subjects were given lists where the women were more famous and tended to respond that there were more women on the list. Thus, the ease of recall affected the judgment. This is also true, for instance, in searching memory and imagining possible events. In the case of intelligence analysis, this heuristic could mean more recent information could be over weighted in forming the judgment even though it is less relevant. It could also mean that more dramatic and stimulating information could be recalled more easily and weighed more heavily than less exciting, but more relevant information.

Instead, in this work we use pairwise comparisons. To avoid asking for over-weighed probability judgments, we ask for relative judgments. In our previous risk assessment work, we have found that experts are more comfortable assessing the relative probability of an event in two situations when each probability is low. Thus, we ask experts to assess the ratio of the probabilities of the event for the two scenarios. Por and Budescu (2016) show that pairwise assessments are significantly more accurate than direct assessments. Multiple factors describe two scenarios to the expert in a meaningful manner and in each comparison one factor is changed between the two scenarios. The method is akin to that in Bradley and Terry (1952),

but the aim is to estimate the effect of the multiple factors rather than developing a ranking scale. In this case, the factors describe a vessel inbound to the United States and the comparison asks the experts to assess which vessel is more likely to be used to smuggle a nuclear threat into the country. Through several such questions with varying vessel descriptions, we can form a model that predicts the probability that any given vessel contains a nuclear threat and combines the judgments of numerous experts.

In the next section, we describe the question format and how multiple questions are used to assess threat probabilities across a range of possible vessels. We then define the probability model and the analysis used to aggregate the judgments and form a predictive probability forecast. Finally, we demonstrate the approach using questionnaire responses from eight experts in the domain.

19.2 Questionnaires

The questions that make up our questionnaire are paired situation comparisons, meaning that each situation differs by only one important element. Figure 19.1 shows an example. The columns in Fig. 19.1 are labeled Vessel 1, Vessel Description, and Vessel 2. Each expert is asked to compare Vessel 1 to Vessel 2 in using the descriptions provided in their respective columns. The columns Vessel 1 and Vessel 2 differ by only one entry that is written in *bold*. Researchers and experts in the field compose a series of factors that could potentially be linked to determining whether a vessel is a threat. The first three rows of each question show factors involving the countries in which the vessel has docked most recently, specifically the last three countries: *Last Country Docked*, *2nd to Last Country Docked*, and *3rd to last Country Docked*. We consider five countries that have been rated in the academic

Vessel 1	**Vessel Description**	**Vessel 2**
Hong Kong	Last Country Docked	Hong Kong
US	2nd to Last Country Docked	US
Singapore	3rd to last Country Docked	Singapore
Monthly	Frequency of US Calls	Monthly
Major Company	Vessel Ownership	Major Company
> 10 years	Age of Ownership Company	> 10 years
Container	Type of Vessel	Container
Mixed products	Type of Cargo	**Lead Materials**
Company Employees	Crew	Company Employees
More?: 9☐ 8☐ 7☐ 6☐ 5☐ 4☐ 3☐ 2☐ 1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐ 8☐ 9☐ More?		

Fig. 19.1 An example question from the questionnaire

literature as higher risk for container-based threats, Bangladesh, Indonesia, Pakistan, Turkey, and Yemen. The other countries considered have high cargo volumes and are not considered high-risk, Hong Kong, Singapore, the Netherlands, and the US. The next factor is the *Frequency of US calls*, which is broken down in to four levels: monthly calls to the US, semi-annual calls to the US, annual calls to the US, and the first call to the US. The next two factors are the type of *Vessel Ownership* (major company, small or regional company, leasing company, or uncertain (multi-layered) ownership) and the *Age of Ownership Company*. The *Type of Vessel* is one of container vessel, bulk carrier, roll-on/roll-off vessel, or a general freighter. The *Type of Cargo* is one of mixed products (such as a Wal-Mart cargo), ceramic tiles, lead materials, or scrap metal. The last factor is *Crew*, that can be company employees, labor union hall, short-term contracts, or crew of opportunity (picking up crew members available at port). We do not consider all possible values of the factors, just enough to assess whether the factor is important.

In Fig. 19.1, the *Last Country Docked* is Hong Kong, *2nd to Last Country Docked* is the United States, *3rd to last Country Docked* is Singapore, *Frequency of US calls* is Monthly, *Ownership* is a Major company (such as Maersk, Evergreen, or Hunan), the *Age of Ownership Company* is more than 10 years, the *Type of Vessel* is Container, and. The only thing that changes between the two vessels is the *Type of Cargo*. For the purposes of this survey, Vessel 1 contains Mixed Products, which are considered to be items common to Walmart or K-mart goods and products. Vessel 2 contains lead materials, such as auto parts, brake drums, or manifolds. Each expert is asked to determine whether the single change between the two vessels elevates the threat.

Now let us turn to the response scale below each question. There is a scale at 122 the bottom of the question. The ranking system ranges from 9 to 1 to 9. For the 123 example, if the expert considers that the change from Mixed Products in Vessel 1 124 to Lead Materials in Vessel 2 makes the expert feel that Vessel 2 has four times 125 the likelihood of containing a threat than Vessel 1, then the expert would check 126 the 4 to the right, closest to Vessel 2, as shown in Fig. 19.2. Alternatively, if the expert 127 thinks Vessel 1 is five times more likely than Vessel 2 to have a nuclear device 128 or illicit radiological materials smuggled onboard then the expert would check the 129 number 5 on the side of Vessel 1. If 9 times the risk is not a sufficient range and 130 the expert thinks Lead Materials on Vessel 2 is 12 times more likely to harbor a 131 threat compared to Vessel 1 then the expert can simply type “12” in the text box to 132 the right. If the expert sees no significant difference between the two vessels in the 133 likelihood of a threat, then the response would be to check the box marked 1.

Some questions in the questionnaire are similar. The key changes are highlighted in bold. However, minor changes may be easily missed. For example, the two questions in Fig. 19.3 are almost identical. Notice, though, that not only is the key comparison important (Hong Kong compared to Bangladesh as Last Country Docked), but also the 2nd to Last Country Docked is changed from the US as in question 1 to Pakistan in question 16. These secondary factors are highlighted in *bold italics*. This allows estimation of interactions between two factors. In this case,

Vessel 1	Vessel Description	Vessel 2
Hong Kong	Last Country Docked	Hong Kong
US	2nd to Last Country Docked	US
Singapore	3rd to last Country Docked	Singapore
Monthly	Frequency of US Calls	Monthly
Major Company	Vessel Ownership	Major Company
> 10 years	Age of Ownership Company	> 10 years
Container	Type of Vessel	Container
Mixed products	Type of Cargo	**Lead Materials**
Company Employees	Crew	Company Employees
More?: 9☐ 8☐ 7☐ 6☐ 5☐ 4☐ 3☐ 2☐ 1☐ 2☐ 3☐ **4☒** 5☐ 6☐ 7☐ 8☐ 9☐ More?		

Fig. 19.2 An example of an expert's response to the question in Fig. 19.1

Vessel 1	Vessel Description	Vessel 2
Hong Kong	Last Country Docked	**Bangladesh**
US	2nd to Last Country Docked	US
Singapore	3rd to last Country Docked	Singapore
Monthly	Frequency of US Calls	Monthly
Major Company	Vessel Ownership	Major Company
> 10 years	Age of Ownership Company	> 10 years
Container	Type of Vessel	Container
Mixed products	Type of Cargo	Mixed products
Company Employees	Crew	Company Employees
More?: 9☐ 8☐ 7☐ 6☐ 5☐ 4☐ 3☐ 2☐ 1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐ 8☐ 9☐ More?		

Vessel 1	Vessel Description	Vessel 2
Hong Kong	Last Country Docked	**Bangladesh**
Pakistan	2nd to Last Country Docked	***Pakistan***
Singapore	3rd to last Country Docked	Singapore
Monthly	Frequency of US Calls	Monthly
Major Company	Vessel Ownership	Major Company
> 10 years	Age of Ownership Company	> 10 years
Container	Type of Vessel	Container
Mixed products	Type of Cargo	Mixed products
Company Employees	Crew	Company Employees
More?: 9☐ 8☐ 7☐ 6☐ 5☐ 4☐ 3☐ 2☐ 1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐ 8☐ 9☐ More?		

Fig. 19.3 An example of two questions that combined allow the estimation of interaction effects between factors

the effect of *Last Country Docked* depends on *2nd to Last Country Docked*. In fact, we include a possible three-way interaction between *Last Country Docked*, *2nd to Last Country Docked*, and *3rd to last Country Docked*, as the exact combination of all three may be the cause of concern to the intelligence analyst.

The questionnaire takes between 30 and 45 minutes. We received questionnaires from eight personnel with significant expertise in intelligence analysis in the maritime domain.

19.3 Analysis

The occurrence and non-occurrence of a nuclear threat on board a vessel is modeled by exchangeable Bernoulli trials with an unknown probability that depends on the factors describing the vessel. The model assumed takes the form of a proportional probabilities model (Merrick et al. 2000), based on the idea of the proportional hazards model (Cox 1972). Let $X = (x_1, \ldots, x_q)^T$ denote the q factors describing a vessel that may contain a nuclear threat. The conditional probability of a nuclear threat, given the vessel description defined by X, is assumed to be

$$P\left(Threat|X, p_0, \beta\right) = p_0 \exp\left(X^T \beta\right) \tag{19.1}$$

where $\beta = (\beta_1, \ldots, \beta_q)^T$ is a vector of q parameters and p_0 is a baseline probability parameter. Consider a single question that asks the expert to compare two vessel descriptions defined by the factor vectors L (for vessel 1 on the left) and R (for vessel 2 on the right). The experts are asked to judge which vessel is more likely to contain a nuclear threat and by how much or

$$\frac{P\left(Threat|R, p_0, \beta\right)}{P\left(Threat|L, p_0, \beta\right)} = \frac{p_0 \exp\left(R^T \beta\right)}{p_0 \exp\left(L^T \beta\right)} = \exp\left((R - L)^T \beta\right) \tag{19.2}$$

where $(R - L)$ denotes the difference vector between the two factor vectors. This implies that the expert's response is related to the difference between the two vessel descriptions and the parameter vector β.

Multiple experts complete each questionnaire, so there are multiple responses to each question. Let the experts be indexed by $j(=1, \ldots, p)$ and the questions be indexed by $i(=1, \ldots, N)$, so the experts' responses can be denoted $z_{i,j}$. We now have that $z_{i,j}$ is the j-th expert's estimate of the ratio of probabilities for the i-th question, while the model gives this relative probability as $\exp((R_i - L_i)^T \beta)$. If we let $X_i = (R_i - L_i)$ then we can assume that $z_{i,j}$ and $\exp((R_i - L_i)^T \beta)$ are equal up to a random error, or

$$\ln\left(z_{i,j}\right) = X_i^T \beta + u_{i,j} \tag{19.3}$$

where $u_{i,j}$ is the residual error term representing the variation between the experts' responses around the model. Assuming that the errors $u_{i,j}$ are independent and normally distributed with zero mean and variance σ^2, this equation is a standard linear regression, where $y_{i,j} = \ln(z_{i,j})$ is the dependent variable, X_i is the vector of independent variables, β is a vector of regression parameters and $u_{i,j}$ is the error term (Press 1982).

The regression analysis provides an estimate of β, but does not provide an estimate of p_0. This means that we can estimate the ratio of the probability of a nuclear threat for any two vessels, but we cannot predict the probability for a single vessel without p_0. How then does a decision maker obtain the actual probability of a nuclear threat for a specific vessel for use in decision making? The decision maker can assess the probability for one reference vessel, with vessel description X_0. Suitable techniques for aggregation of probability assessments are reviewed in Clemen and Winkler (1999). The probabilities for any given other vessel with vessel description X_* can be found by multiplying

$$P\left(Threat|X_0,\beta\right) \times \frac{P\left(Threat|X_*,\beta\right)}{P\left(Threat|X_0,\beta\right)} = P\left(Threat|X_0,\beta\right) \times \exp\left((X_* - X_0)^T\beta\right) \tag{19.4}$$

However, we must remember that experts overestimate low probability events, which is our reason for using pairwise comparisons in the first place. This means that higher values of $P(Threat|\,X_0\,,\beta)$ are likely to be better calibrated. Thus, we must choose X_0 to be the vessel description with the highest possible chance of containing a nuclear threat in the expert's opinion.

19.4 Results

We obtained completed questionnaires from eight experts. Each expert had significant expertise in the maritime security domain ranging from 27 to 42 years and including commercial, US Coast Guard, Customs and Border Patrol, and Department of Homeland Security roles. Each expert completed 35 questions, allowing the estimation of all parameters.

Figure 19.4 shows the responses to the questions concerning the last three ports of call as box plots. For each country, the individual responses are shown as points on the plot and then the box includes a centerline for the median and an upper and lower end of the box for the 75th and 25th percentiles of the responses. The whiskers of the plot extend to the highest and lowest response value unless that value would make the whisker more than 1.5 times the length of the box. Values beyond this range are considered outliers. Figure 19.4 shows the same general pattern of response for the last and second to last port of call. There is more disagreement between the experts regarding the third to last port of call, although the ranking of the medians is the same.

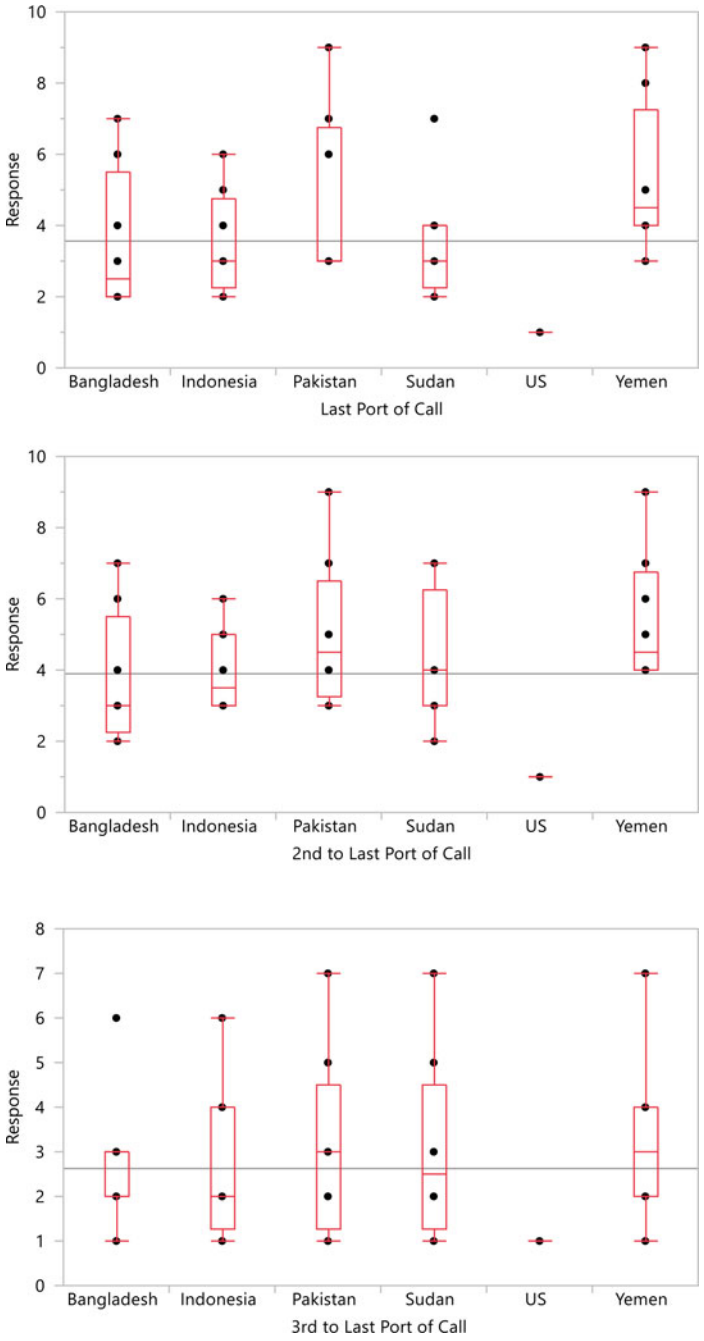


Fig. 19.4 Box plots of the experts' responses concerning the last three ports visited

Figure 19.5 shows the responses to the questions concerning the cargo on the vessel, the ownership of the vessel, and the manner of hiring crew for the vessel. There is some disagreement as to the effect of ceramic tile, lead products, and scrap metal on the likelihood of hiding a nuclear threat. This is because each of these cargoes can cause problems for radiation screening equipment. Ceramic tiles can be naturally occurring radioactive materials, or NORMs, causing false positive alarms. Lead can shield radiation and so could be used to hide nuclear material from screening equipment. There appears to be some disagreement between the experts whether terrorists would be likely to mask true nuclear threats with NORM sources or shield them in lead. However, the ranking of the median values is quite clear, so there is a partial consensus with only a few experts in disagreement. There is also some disagreement about the effect of a crew of opportunity. This essentially means that crewmembers are hired as needed when the vessel arrives in port. However, while the experts disagree as to how much worse this is than other hiring practices, they all agree that it is worse.

Figure 19.6 shows the responses to the questions concerning the frequency of calls to the US and the type of vessel. There is more general disagreement here. Some experts think that yearly calls and first calls to the US should be of great concern, whereas others think there is little difference. Furthermore, some experts are more concerned about container vessels, while others consider freighters a higher risk. We performed a stepwise regression on all factors in the model, including second and third order interactions between the factors representing the last three ports of call. As expected from Fig. 19.6, the disagreement about the frequency of US calls and the vessel type made those factors statistically insignificant. The interaction terms were also found to be insignificant, meaning that the pattern of the last three calls is not important, just whether they contain calls to specific countries. The final model is statistically significant (F-ratio $= 19.2367$, $n = 275$, and a p-value of 9.13×10^{-21}). The model had an R^2 value of 33.5%, a reflection of the disagreement of individual experts around the consensus represented by the model. A lack of fit test on the model was insignificant with a p-value of 0.3994.

Table 19.1 shows the parameter estimates for the final model, including estimates of the regression parameters, β, their 90% confidence intervals, and their individual p-values for significance ($\beta \neq 0$). All factors are scaled from 0 to 1, with 0 being the value of least concern and 1 being the value of most concern. Thus, the parameter values represent the range in the log-linear scale from the least to the most concerning values of that parameter. This means that we can compare which factor causes the largest change in likelihood of a threat when swung from its least concerning level to its most. We can see that ownership has the largest range from a major company to an unknown owner. The next largest range is the last port of call and then the type of crew hiring practices. In terms of the last three ports of call, the ordering is as expected from the last to the third to last port of call. The smallest range is for the type of cargo, probably because of the disagreement about cargoes that can mask or shield nuclear threats.

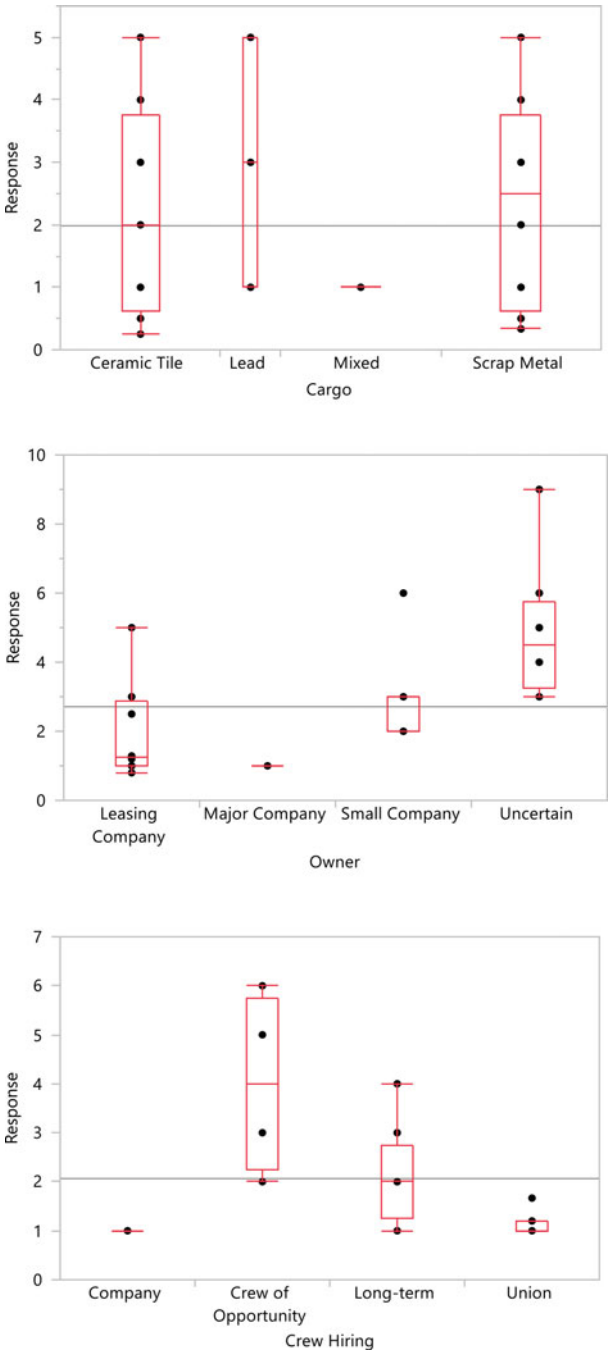


Fig. 19.5 Box plots of the experts' responses concerning the cargo, owner, and crew

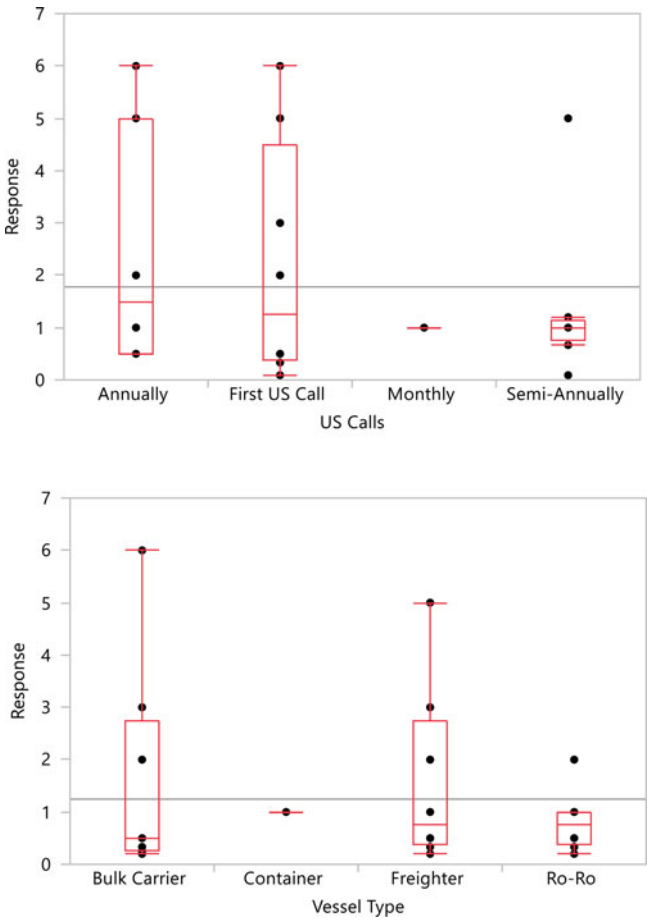


Fig. 19.6 Box plots of the experts' responses concerning the type of vessel and its frequency of calling at US ports

Table 19.1 Parameter estimates for the final model

Term	Estimate	Std. error	t ratio	p-value	Lower 95%	Upper 95%
POC 1	1.907	0.211	9.022	3.72E-17	1.491	2.323
POC 2	1.691	0.172	9.819	1.26E-19	1.352	2.030
POC 3	0.993	0.172	5.767286	2.22E-08	0.654	1.333
Owner	2.280	0.428	5.331	2.08E-07	1.438	3.122
Owner age	1.726	0.356	4.848	2.12E-06	1.025	2.427
Cargo	0.512	0.227	2.254	0.050	0.065	0.958
Crew	1.896	0.405	4.69E + 00	4.43E-06	1.099	2.692

19.5 Representative Threat Predictions

Consider the comparison of the two vessels in Fig. 19.7. Six factors have been changed between the two vessels. Could we simply ask the expert to compare these two vessels? The simple answer is no. The complexity of the task would make the judgment at least questionable, if not ill advised. Instead, we can use the model to predict the relative difference in likelihood between each vessel containing a nuclear threat. Substituting these two vessels in to Eq. (19.3) with the parameter values in Table 19.1, we obtain a best estimate of the log ratio of 3.89 with a 95% confidence interval of [3.19, 4.59]. However, taking the exponential transformation to get the ratio of the likelihoods, we obtain an estimate of 48.81 with a 95% confidence interval of [24.29, 98.01]. The most important difference here is that in the second to last country, which makes vessel 2 3.41 times as likely to contain a threat as vessel 1. While ownership has the largest parameter and so difference across its full range, the difference between a major company and a leasing company is small. Thus, it is the combination of the specific factor changes and the parameter values that determine the predicted ratio of the probabilities.

As a final analysis, let us consider the comparison of the least likely to the most likely vessel to contain a nuclear threat. Figure 19.8 shows such a comparison. Note that technically we should set vessel 2 to call in Yemen (ranked by all experts as the largest concern) for the last three ports of call, but instead we chose the sequence to be Yemen-Pakistan-Yemen-US. Substituting these two vessels in to Eq. (19.3) with the parameter values in Table 19.1, we obtain a best estimate of the log ratio of 5.15 with a 95% confidence interval of [4.14, 6.16]. However, taking the exponential transformation to get the ratio of the likelihoods, we obtain an estimate of 172.75 with a 95% confidence interval of [63.18, 472.33].

Vessel 1	Vessel Description	Vessel 2
Hong Kong	Last Country Docked	**Bangladesh**
US	2nd to Last Country Docked	**Pakistan**
Singapore	3rd to last Country Docked	**Turkey**
Monthly	Frequency of US Calls	Monthly
Major Company	Vessel Ownership	**Leasing Company**
> 10 years	Age of Ownership Company	> 10 years
Container	Type of Vessel	Container
Mixed products	Type of Cargo	**Ceramic Tile**
Company Employees	Crew	**Long-term**
More?: 9☐ 8☐ 7☐ 6☐ 5☐ 4☐ 3☐ 2☐ 1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐ 8☐ 9☐ More?		

Fig. 19.7 An example of a vessel comparison with many differences

Vessel 1	Vessel Description	Vessel 2
US	Last Country Docked	**Yemen**
US	2nd to Last Country Docked	**Pakistan**
US	3rd to last Country Docked	**Yemen**
Monthly	Frequency of US Calls	Monthly
Major Company	Vessel Ownership	**Uncertain**
> 10 years	Age of Ownership Company	**1 month**
Container	Type of Vessel	Container
Mixed products	Type of Cargo	**Lead Products**
Company Employees	Crew	**Crew of Opportunity**
More?: 9☐ 8☐ 7☐ 6☐ 5☐ 4☐ 3☐ 2☐ 1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐ 8☐ 9☐ More?		

Fig. 19.8 An example of a vessel comparison with largest possible differences

19.6 Conclusions

We have demonstrated the use of extended pairwise comparisons in predicting the relative likelihood that a vessel entering US waters contains a nuclear threat. We developed a questionnaire with 35 questions to populate the parameters of a probabilistic prediction model and demonstrated how log-linear regression can be used to estimate the parameters of the model by aggregating the questionnaire responses from multiple experts. We obtained responses from eight experts with significant knowledge of this problem domain and formed a predictive model based on their judgments. The predictive model was used to compare two example vessels and determine the relative risk associated with the most "risky" vessel and the least "risky" vessel.

This study stands as a proof-of-concept for the technique. The questionnaire can be designed to include additional factors, a larger set of possible ports of call, and finer granularity in important factors. We have also demonstrated how interactions between the factors can be included, although they proved not to be significant in our final model. The approach can be used to predict absolute probabilities using the threat probability for a reference vessel and then predicting the relative likelihood for any vessel of interest in comparison to the reference vessel. Clearly, this approach can be applied beyond nuclear threats and for situations beyond threats on a vessel. The same approach could be applied to individual containers on a vessel or even to air and land borders.

The frequentist regression analysis used here allows uncertainty to be represented as a confidence interval. However, there was considerable disagreement between the experts on the importance of the different factors, as indicated by the R^2 value of 33% and the wide confidence intervals obtained on the predictions. One might suggest that the calibration of the experts could be assessed—see Chapters 2: "Elicitation in the Classical Model" (Quigley et al. 2018) and 3: "Validation

in the Classical Model" (Cooke 2018) in this book—and included in the analysis through weighted least squares regression. However, the lack of experience with such specific threats, and the relatively low occurrence of terrorist threats overall, precludes such analysis. Thus, we are left with increasing the pool of experts and correctly expressing the remaining uncertainty. Bayesian analysis offers a more complete uncertainty analysis to be performed (Paté-Cornell 2002). Szwed et al. (2006) propose a fully Bayesian analysis of extended pairwise comparisons, while Merrick et al. (2005) propose an extended Bayesian analysis to allow for dependencies in the experts' responses due to overlapping information available to the experts. Thus, the application of these techniques could provide a more accurate expression of the remaining uncertainty in the threat probabilities predicted.

Acknowledgments Developed partially under grants from the U.S. Department of Homeland Security's Domestic Nuclear Detection Office under Grant Award Number 2008-DN-077 ARI001-02 and the National Science Foundation (CBET-0735735). The work was done at Virginia Commonwealth University. The views and conclusions contained in this document are those of the author and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security or the National Science Foundation. This research has been deemed to be exempt from federal regulations on treatment of human subjects, under approval VCU IRB #: HM11880.

References

Bradley R, Terry M (1952) Rank analysis of incomplete block designs. Biometrica 39:324–345

Clemen RT, Winkler RL (1999) Combining probability distributions from experts in risk analysis. Risk Anal 19(2):187–203

Combs B, Slovic P (1979) Newspaper coverage of causes of death. Journal Q 56(4):837–843

Cooke RM (1991) Experts in uncertainty: Expert opinion and subjective probability in science. Oxford University Press, Oxford

Cooke RM (2018) Validation in the classical model. In: Dias LC, Morton A, Quigley J (eds) Elicitation: the science and art of structuring judgment. Springer, New York. (Chapter 3 in this book)

Cox DR (1972) Regression models and life tables. J R Stat Soc Ser B 34:187–220

Ebeling CE (2009) Evolution of a box. Invent Technol 23(4):8–9

Gardner F (2003) Al-Qaeda "was making dirty bomb". Available at BBC News Online: http://news.bbc.co.uk/2/hi/uk_news/2711645.stm. Accessed 3 Mar 2017

Merrick JRW, van Dorp JR, Harrald J, Mazzuchi TA, Spahn J, Grabowski M (2000) A systems approach to managing oil transportation risk in Prince William Sound. Syst Eng 3(3):128–142

Merrick JRW, van Dorp JR, Singh A (2005) Analysis of correlated expert judgments from pairwise comparisons. Decis Anal 2(1):17–29

Merrick JRW, McLay LA (2010) Is screening cargo containers for nuclear threats worthwhile? Decis Anal 7(2):155–171

Morris PA (1974) Decision analysis expert use. Manag Sci 20:1233–1241

Morris PA (1977) Combining expert judgments: a Bayesian approach. Manag Sci 23:679–693

Morris PA (1983) An axiomatic approach to expert resolution. Manag Sci 29:24–32

Paté-Cornell ME (2002) Fusion of intelligence information: a Bayesian approach. Risk Anal 22(3):445–454

Por H-H, Budescu DV (2016) Eliciting subjective probabilities through pairwise comparisons. J Behav Decis Mak 30(2):181–196. doi:10.1002/bdm.1929

Press SJ (1982) In: Robert E (ed) Applied multivariate analysis using Bayesian and frequentist methods and inference, 2nd edn. Krieger Publishing Company, Malabar, FL

Quigley J, Colson A, Aspinall W, Cooke R (2018) Elicitation in the classical model. In: Dias LC, Morton A, Quigley J (eds) Elicitation: The science and art of structuring judgment. Springer, New York. (Chapter 2 in this book)

Szwed P, van Dorp JR, Merrick JRW, Mazzuchi TA, Singh A (2006) A Bayesian paired comparison approach for relative accident probability assessment with covariate information. Eur J Oper Res 169(1):157–177

Tversky A, Kahneman D (1973) Availability: a heuristic for judging frequency and probability. Cog Psychol 5:207–232

Tversky A, Kahneman D (1983) Extension versus intuitive reasoning: the conjunction fallacy in probability judgment. Psychol Rev 90(4):293–315

Chapter 20
Risk Assessment Using Group Elicitation: Case Study on Start-up of a New Logistics System

Markus Porthin, Tony Rosqvist, and Susanna Kunttu

Abstract This chapter presents a risk assessment for the start-up of a new logistics system within the pulp and paper manufacturer Stora Enso. The risk assessment was realised as a structured expert elicitation workshop using a computerised group support system. Experts representing different parts of the logistics system were invited to a one-day workshop to assess risks concerning the system start-up. The main topics of the workshop were hazard identification, risk estimation and risk control. Each identified risk scenario was assessed with regard to its likelihood and three consequence types related to logistics: timeliness, product quality and information quality. The top priority risks were identified and risk controls were outlined.

The computerised group support system made the workshop more efficient due to the possibility of simultaneous inputs from all participants to a shared environment, versatile processing possibilities of the inputs, voting features with instant results and automated documentation. An essential factor for the success of the workshop was thorough preparation in cooperation between the analysts and the problem owner. Each step of the workshop process was specified and special attention was given to ensure the elicitation questions were clear and unambiguous.

The risk assessment resulted in a prioritised list of realistic risk scenarios for the start-up of the logistics system and control ideas for the most important risks. The results helped the company structure their work to ensure a problem free start-up. In addition, the workshop participants found it valuable to meet representatives from other parts of the logistics chain.

20.1 Introduction

Risk analysis often deals with poorly documented complex systems which are hard to overview and for which data is difficult to obtain. Structured expert elicitation through expert workshops is one way to overcome the lack of information

M. Porthin (✉) • T. Rosqvist • S. Kunttu
VTT Technical Research Centre of Finland Ltd, Espoo, Finland
e-mail: markus.porthin@vtt.fi

L.C. Dias et al. (eds.), *Elicitation*, International Series in Operations Research & Management Science 261, https://doi.org/10.1007/978-3-319-65052-4_20

(Cooke 1991; Weatherall and Hailstones 2002). In the workshops the expertise and knowledge of different types of participants is exploited and interactively combined in order to produce new information which could not have been received by addressing the experts one by one. The use of computerised group support systems results in more efficient, controlled and comprehensively documented workshops.

This chapter presents an application of structured expert workshop techniques in change management through risk assessment of the start-up of a new logistics system. In 2005–2007, Stora Enso was reforming its logistics system for transporting paper products, mainly paper reels, from their mills in Sweden and Finland to Central Europe. The main objectives of the change were to harmonize the supply system for Nordic situated mills, increase resource and cost effectiveness and flexibility as well as to lower the level of transport damages, improve data quality and enable delivery tracking. A risk assessment was conducted by VTT Technical Research Centre of Finland in order to identify the main risks that may affect the start-up of phase 2 of the supply system in 2006 and to help the company define action plans to control the risks in advance before start-up. The first phase of the supply system did not have quite as smooth a start as hoped for and now the company wanted to ensure a problem free start-up for the second phase. The risk assessment was carried out through a structured computer assisted workshop with participants representing different parts of the supply chain as well as the supply system perspective.

According to the ISO 31000 (2009) standard risk assessment is a part of the risk management process which should go through all levels of the organisation (e.g. Hopkin 2014; Duijne et al. 2008; Arena et al. 2010). The risk management process is presented by five main phases as shown in Fig. 20.1. The main focus of this case study is on risk assessment.

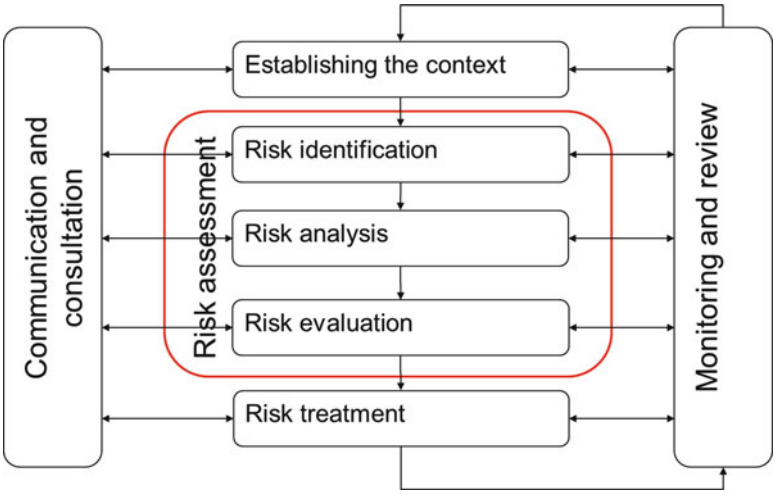


Fig. 20.1 Risk management process (ISO 31000 2009)

Networked computers and group support system software called GroupSystems, nowadays known as ThinkTank (2017), were used in the risk assessment workshop. The computerised system can be seen as an electronic substitute for conventional flip chart and post-it techniques. It enables simultaneous inputs from all participants to a shared environment, commenting on other participants' inputs, grouping, merging and editing of inputs as well as the use of various voting methods. The inputs can be anonymous. The system automatically generates a meeting record containing all inputs and voting results, thus reducing the secretarial work. Computer assisted meetings are usually more structured than conventional ones. A typical session may consist e.g. of exchanging of information and opinions, creation of ideas and actions, exploring and evaluating of the ideas as well as voting on priorities and building commitment. The reader is referred to Weatherall and Nunamaker (2000) for further information on group support systems.

20.2 North European Transport Supply System

In 2005–2007 Stora Enso was reforming its logistics system for transporting mainly paper reels from their mills in Sweden and Finland to Central Europe. The new supply system called NETSS (North European Transport Supply System) was based on transporting cargo in oversized containers, so called SECUs, (Stora Enso Container Unit) via a hub in Gothenburg to the designated ports on the European continent or the Great Britain. The purpose of the SECU containers, with dimensions optimised for paper reels, was to enable efficient loading and lower the level of transport damages through minimised direct handling of the reels. In addition, the SECUs were equipped with radio tags to support efficient tracking.

The new system was launched in three phases (Fig. 20.2): The first phase in 2005 included the Swedish and Southern Finnish mills, the hub in Gothenburg, as well as three discharge ports in Central Europe. In 2006, the Northern Finnish mills and the port of Antwerp were included. Finally in 2007, Lübeck was added to the transport system.

The risk analysis presented here focused on potential disturbances during the first six months after the start-up of the second phase in 2006. The assessment covered the system traffic from loading into SECU containers at the port of loading (in Finland) or at the mill (in Sweden) until the arrival to the inland terminal in Central Europe. Figure 20.3 shows the goods flow from the Finnish mills. From the mills that are not situated by the port, the goods are first transported by rail or truck to the port of loading where they are loaded into SECUs. The cargo continues to the hub in Gothenburg by ships dedicated for this traffic, where the SECUs are re-arranged according to the final destination and then shipped to the port of discharge. There the SECUs are unloaded and the goods are transported by rail to the inland terminal. Empty SECUs are shipped back to Sweden and Finland in returning ships.

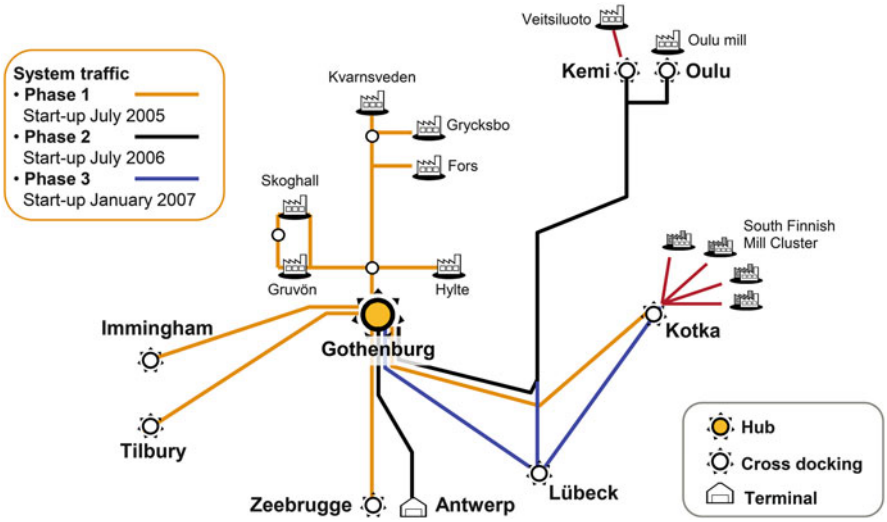


Fig. 20.2 The three phased start-up of NETSS system traffic

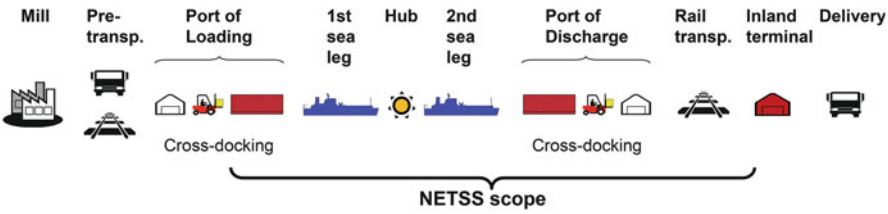


Fig. 20.3 Goods flow from Finnish mills to Central Europe

20.3 Pre Workshop Preparation

20.3.1 Planning of Workshop Process

The workshop was carefully planned in cooperation between the analysts and the client over a series of meetings. To ensure the success of the workshop, the process must be tailored to address the desired questions in an appropriate way. The theme and objectives of the workshop were discussed and defined. Based on this, the facilitators suggested a structure for the workshop, which was then refined throughout the planning stage.

Each step of the workshop process was clearly specified. The likelihood and consequences of the identified risks were to be assessed by voting. One important task was to determine and accurately define the consequence types to account for

Table 20.1 Likelihood scale

Scale	Description	Probability
1.	Very unlikely	0–5%
2.	Unlikely	5–25%
3.	Possible, "fifty-fifty"	25–75%
4.	Likely	75–95%
5.	Very likely	95–100%

and their associated voting scales. Three main quality attributes used in logistics were chosen to represent consequences of the risks and their qualitative scales were defined based on company specific conditions. The voting questions and scales should not only be unambiguous but also easy to internalise by the experts in order to aid the judgement process and prevent misunderstandings.

20.3.1.1 Definition of the Likelihood Scale

For the likelihood[1] assessment of the risks a five step scale was prepared (Table 20.1). The workshop participants were to assess how likely each scenario was to occur during the first 6 months of NETSS phase 2. The limits of the likelihood scale were defined using practical experience that people tend to have a good idea whether a scenario is likely or unlikely, but it may be harder to assess whether the probability for a possible scenario should be e.g. 40% or 60%. The likelihood judgements were subsequently converted into probabilities by using the middle points of the corresponding probability intervals in order to calculate the risk indices.

20.3.1.2 Definition of Consequence Types and Scales

The consequences were decided to be assessed using three criteria: Deviation of timeliness, Deviation of product quality and Deviation of information quality. The scales were developed together by the VTT risk analysts and the client (Tables 20.2, 20.3, and 20.4). In order to define the consequence scales in an as descriptive and intuitive way as possible for the workshop participants, each level on the scale was given a tailored verbal description. The consequence levels were also further clarified by expressing their effects on costs, customers and company internal credibility of NETSS.

[1]Here the term likelihood is used instead of probability, because the elicitation was done using a 1–5 scale whereas probability per definition is a number between 0 and 1. However, probabilities were used in the definition of the likelihood scale (Table 20.1).

Table 20.2 Scale for consequence 1: deviation of timeliness

Scale	Description	Extra costs (resources)	Effect on customers	Effect on company internal credibility of NETSS
1.	Problem is local and can be sorted out on local level	Negligible	No	No/negligible
2.	Problem is local but requires cooperation	Moderate	No	Negligible/moderate
3.	Problem is affecting subsystem performance and requires cooperation inside NETSS	Notable	Moderate	Notable
4.	Problem decrease NETSS system overall performance, requires cooperation and possibly alternative means for transportation/it messaging etc. for some time	Notable	Notable	Substantial
5.	Problem decrease NETSS system overall performance to a degree that it is difficult to compensate with alternative actions (transportation, messaging etc)	Substantial	Loss of credibility	Unrepairable

Table 20.3 Scale for consequence 2: deviation of product quality

Scale	Description	Extra costs (resources)	Effect on customer deliveries	Effect on company internal credibility of NETSS
1.	Problem is local and can be sorted out on local level (i.e. refurbishment). Corresponding to current damage levels	Moderate	No	No/moderate
2.	Problem is local but requires increased effort. Moderately increased damage level	Moderate/notable	No	Moderate
3.	Problem is affecting subsystem performance and requires cooperation inside NETSS. Increased damage levels affecting delivery reliability	Notable	Moderate	Notable
4.	Problem decreases NETSS system overall performance. Increased damage levels affecting delivery reliability	Notable/substantial	Notable	Notable/substantial
5.	Problem decreases NETSS system overall performance. Increased damage levels affecting delivery reliability and system reliability	Substantial	Substantial	Substantial

Table 20.4 Scale for consequence 3: deviation of information quality

Scale	Description of problem	Extra costs, manual work (resources)	Effect on info to external customers	Effect on info to internal customers
1.	Problem is local and can be sorted out on local level	Negligible/moderate	No	Negligible
2.	Problem is local but requires cooperation	Moderate	No	Moderate
3.	Problem is affecting subsystem performance and requires cooperation inside NETSS	Notable	Negligible	Notable
4.	Problem decrease NETSS system overall performance, requires cooperation and possibly alternative means for it messaging etc. for some time	Substantial	Moderate/notable	Substantial
5.	Problem decrease NETSS system overall performance to a degree that it is difficult to compensate with alternative actions	Substantial	Loss of credibility	Unrepairable

20.3.2 Selection of Experts

The experts for the workshop were selected by the client with assistance by the analysts. The experts were selected according to three criteria: (1) representation from parties exposed to the highest degree of change (2) expertise in operations, and (3) operational representation covering the whole NETSS chain including the NETSS project personnel.

The selected experts for the workshop consisted of 24 persons from Stora Enso and partners, representing operative and administrative key functions of the NETSS system, and one supply chain expert from VTT. The workshop was facilitated in cooperation by two risk analysts from VTT and a representative from the computer system provider MeetingSupport.dk.

20.4 Computer Assisted Expert Workshop

The risk assessment was carried out through a one-day structured computer assisted workshop consisting of hazard identification and risk estimation, identification of top priority risks and preparation of risk control ideas. The workshop was led by facilitators using techniques such as brainstorming, consensus building discussions and voting, following an agenda planned in cooperation by the analysts and the problem owner. The use of the computerised group support system gave all

participants the possibility to express their views simultaneously without anybody being able to dominate the discussion. The computer system enabled also further processing of the given inputs and instant voting based on them. In addition, all inputs were documented in the system. These features made the workshop more efficient than if using traditional post-it and flip chart techniques.

The targets of the workshop were to first find out potential hazards associated to the start-up of NETSS phase 2, then identify the most critical ones using likelihood and consequence criteria and finally consider ways to address the risks. The structure of the workshop is outlined below along with approximate duration of each task, excluding breaks.

Task 0: Introduction (45 min)

a. Goals and scope of the workshop
b. Short description of the NETSS system
c. Description of the work method

Task 1: Hazard identification (2 h)

a. Generation of risk scenarios
b. Commenting risk scenarios
c. Moderated merging of scenarios

Task 2: Risk estimation (1 h 30 min)

a. Estimation of likelihood of scenarios (voting)
b. Estimation of consequences of scenarios (voting)
   - Timeliness
   - Product quality
   - Information quality
c. Identification of top priority risks based on voting results

Task 3: Risk control ideas (1 h 40 min)

a. Small groups focussing on one risk at a time
b. Discussion about risk controls for top priority risks
c. Conclusions and closure of the workshop

20.4.1 Introduction of the Workshop

In the introduction, the goals and scope of the workshop as well as the NETSS system were presented by representatives of the NETSS management. The facilitators from VTT presented the workshop method. It was important to go through the system under examination in order to give the participants a common understanding.

20.4.2 Hazard Identification

The purpose of the hazard identification was to generate risk scenarios that might occur in the NETSS system. The participants were grouped into groups of 1–3 persons according to the part of the organisation they represented. The groups typed in scenarios into the computer system simultaneously and anonymously. To avoid duplicates, they could see what the other groups had already typed in. The instructions in Tables 20.5, 20.6, and 20.7 were presented and given to the participants.

The participants were instructed that a risk scenario should consist of a specific adverse event, which leads to some deviations from the normal operations (Table 20.5). To facilitate the scenario generation process, a list of general adverse events (Table 20.6) was given to the participants. The participants were also asked to categorize the scenarios according to where the adverse event occurs (Table 20.7).

Table 20.5 Instruction to the participants on the generation of risk scenarios

Generation of risk scenarios
1. Choose a supply chain function where the adverse event occurs (see Table 20.7)
2. Give title: [Concise title specifying (1) *adverse event* (see Table 20.6) and (2) resulting *deviation from normal operation* (3) at some *location*]
3. Comments: [Give further description of the hazard, risk control procedures already in place . . .]
Remember that the scenarios have to be related to NETSS step 2. Try to avoid very unrealistic scenarios

Table 20.6 List of adverse events

Category	Adverse event
Technical failures	Cargo capacity problems
	IT system performance capacity problems
	IT system breakdown
	Back up failure (IT & automation)
	Equipment breakdown, e.g. vessel, scanner (specify)
	Compatibility problems
	Other technical problems (specify)
Information/data errors	Delayed data
	Missing data
	Wrong or distorted data
	Other data errors (specify)
Human errors related to operational procedures	Error of commission (good intention, wrong outcome)
	Error of omission (neglect)

Table 20.7 Supply chain functions for categorizing the risk scenarios

Supply chain functions
Pre transportation
NETSS system traffic
Port of loading (SECU and vessel loading)
Hub operation
Port of discharge
Vessels/lines
Cross-docking
Capacity management (prioritization of SECUs, vessel capacity)
Order management (prioritization based on customer orders)
Distribution
Tracking & tracing within system traffic
Tracking & tracing from sales order to delivery
Other

This brought similar risks closer to each other and helped to process the results. To share the knowledge, the participants were asked to comment on the scenarios typed by the other groups. Before starting the scenario generation, examples of well-defined risk scenarios were presented to the participants.

During the hazard identification the participants generated approximately 100 risk scenarios. Guided by the facilitators, the group went through all the scenarios in order to merge overlapping scenarios, refine poorly defined ones, delete irrelevant ones and put aside scenarios out of scope. After this "cleaning" activity 58 mostly well-defined, specific enough and reasonable risk scenarios were left. Although the workshop aimed at identifying the top risks, all 58 scenarios should be regarded as relevant and at least checked whether they are already covered well enough for NETSS phase 2. After all, at least someone of the participating NETSS experts was concerned about each risk.

20.4.3 Risk Estimation

In the risk estimation task, each risk scenario was evaluated according to its likelihood and consequences in a reasonable worst-case scenario. Each group entered their judgements to each identified hazard into the computer system using predefined five-step scales, presented previously in Tables 20.1, 20.2, 20.3, and 20.4. After the evaluation, the scenarios were visualised by risk matrices where the most critical risks lie in the upper right corner (see Fig. 20.4).

As an example, the voting results of one of the risk scenarios ("Order amendment is not done correctly which leads to working based on wrong information") is given in Table 20.8.

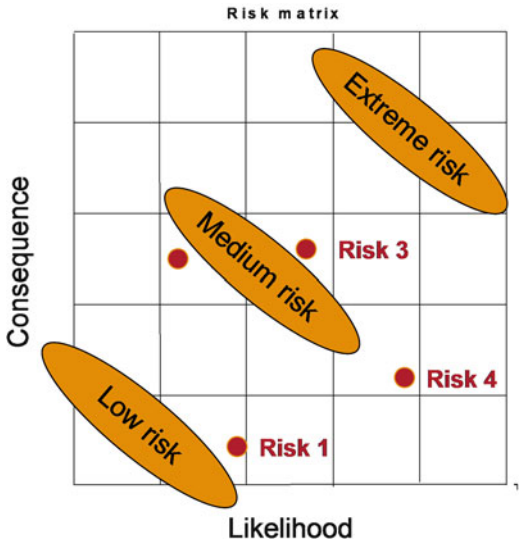


Fig. 20.4 Risk matrix with a five step likelihood and consequence scale as deemed adequate for the NETSS risk analysis

Table 20.8 Voting results for the example risk scenario "Order amendment is not done correctly which leads to working based on wrong information"

	1	2	3	4	5	Mean	STD	n
Likelihood		2	3	4	1	3.4	0.97	10
Timeliness		2	8			2.8	0.42	10
Product quality	7	2	1			1.4	0.7	10
Information quality	1	2	5	1	1	2.9	1.1	10

Columns labelled 1–5 indicate the number of votes given for each score on the voting scales

The mean values of the voting results were then used to produce risk matrices for the consequence categories, see Figs. 20.5, 20.6, and 20.7. The example risk scenario is marked in the figures with a red square. As the analysis is totally based on opinions and good guesses of the workshop participants, the results should be regarded as indicative only. No very critical risks were identified. However, the voting results clearly indicate which risks are regarded as more severe than others.

20.4.3.1 Risk Index

Risk indices combining the likelihood and consequence of the risk scenarios were calculated for each consequence category. The indices were obtained by multiplying the probability of occurrence of a risk with the consequences. To do this, the likelihood index voting results (on a scale 1–5) were converted to probabilities. As seen in Table 20.1, each number 1–5 on the voting scale corresponds to a probability interval. An estimate for the group's view of the probability of a risk scenario was attained by representing each vote (1–5) by the middle point of the corresponding probability interval and calculating the mean of the probabilities:

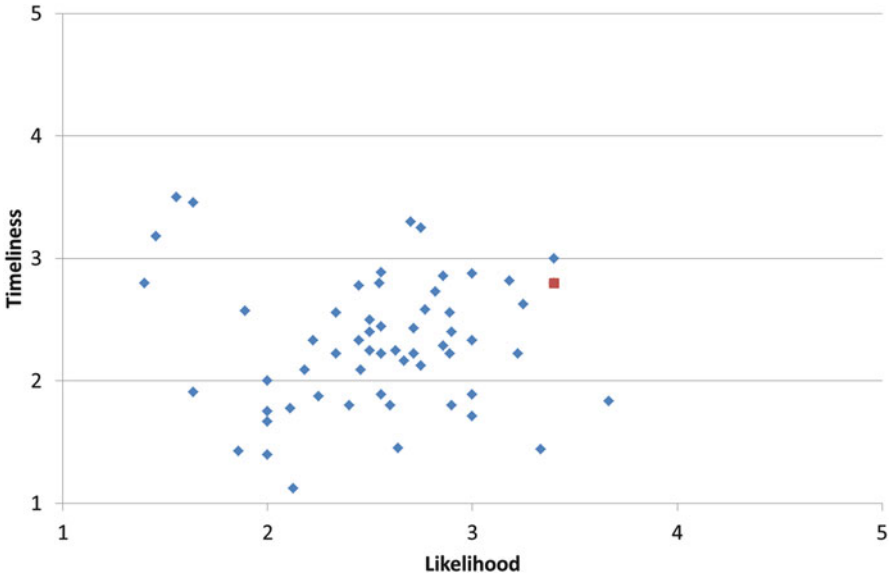


Fig. 20.5 Risk matrix showing the likelihood and timeliness consequence of the risk scenarios. The example risk scenario is marked with a *red* square

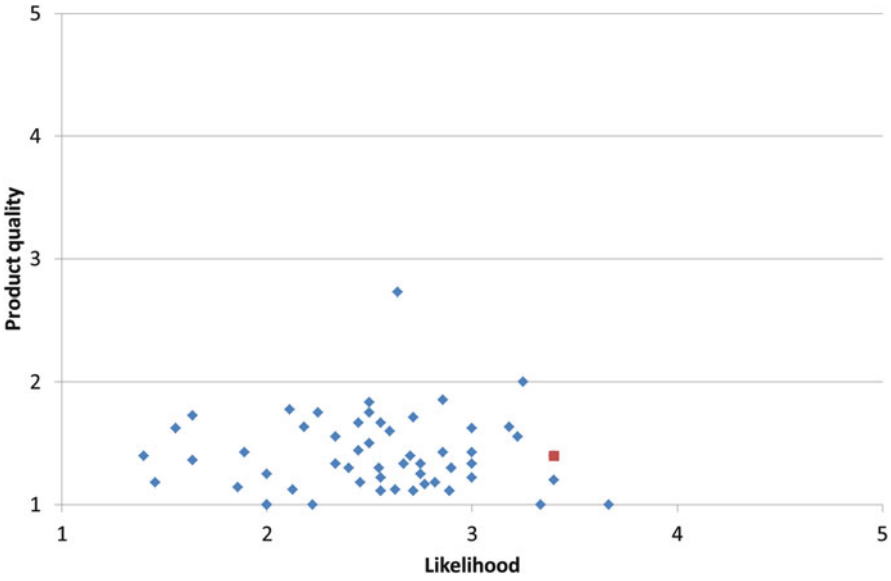


Fig. 20.6 Risk matrix showing the likelihood and product quality consequence of the risk scenarios. The example risk scenario is marked with a *red square*

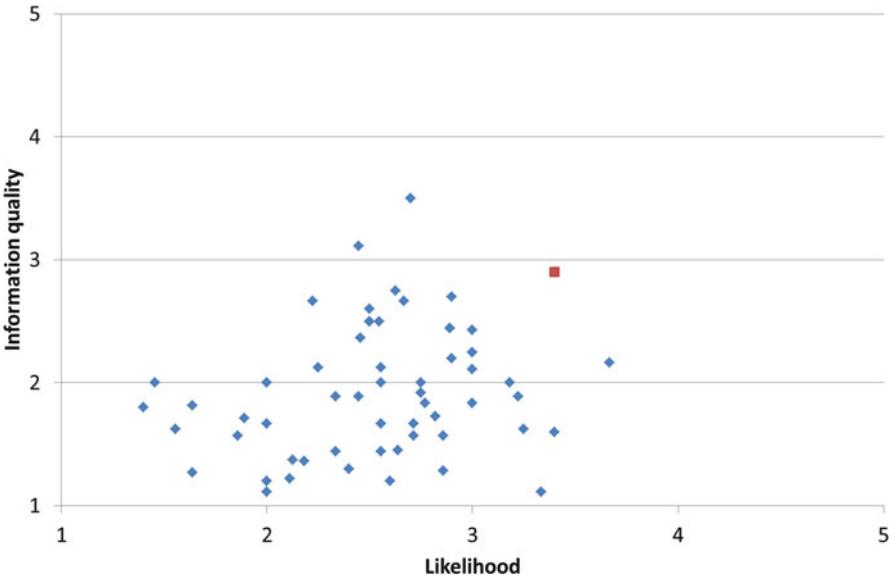


Fig. 20.7 Risk matrix showing the likelihood and information quality consequence of the risk scenarios. The example risk scenario is marked with a *red square*

$$\widehat{p}(x) = \frac{n_1(x) \cdot 0.025 + n_2(x) \cdot 0.15 + n_3(x) \cdot 0.5 + n_4(x) \cdot 0.85 + n_5(x) \cdot 0.975}{n(x)} \tag{20.1}$$

where $n_i(x)$ is the number of likelihood votes i on risk scenario x and $n(x)$ is the total number of votes on x. The risk index for scenario x for each consequence category j is then

$$risk_j(x) = \widehat{p}(x) \cdot consequence_j(x) \tag{20.2}$$

where $consequence_j(x)$ is the mean score of consequence category j for scenario x. To make the risks commensurable an overall risk index, accounting for all consequence categories, was also calculated:

$$risk(x) = risk_{Time}(x) + risk_{Prod}(x) + risk_{Info}(x). \tag{20.3}$$

The likelihood votes for the example risk scenario were $n_1 = 0$, $n_2 = 2$, $n_3 = 3$, $n_4 = 4$, $n_5 = 1$ (see Table 20.8). Using Eq. (20.1) the probability estimate for the scenario is thus 0.62. The risk indices for the scenario are shown in Table 20.9.

Table 20.9 Risk indices for the example risk scenario

Risk index	
Timeliness	1.73
Product quality	0.86
Information quality	1.79
Overall	4.38

20.4.3.2 Identification of Top Priority Risks

Out of the 58 identified risk scenarios, the ones nearest to the upper right corner of each risk matrix were selected as top priority risks for further examination. The following risk scenarios were selected:

- The 10 risk scenarios with highest overall risk index (Eq. (20.3)).
- For all three consequence categories, the risk scenario with the highest consequence specific risk index (Eq. (20.2)) out of the ones not in the top 10 overall risks.
- For all three consequence categories, the risk scenario with the highest consequence score.

20.4.3.3 Top Priority Risks

Generally speaking, the top priority risks had either to do with quality of information, transportation and handling or the logistics process.

The information in the transport system may be inaccurate or lacking due to technical or human errors. This may lead to working processes based on wrong information, e.g. ports of discharge are unable to work with units without unit information.

Local problems such as delay of a vessel or shortage of berthing places at ports as well as unclear work prioritisation may cause delays further down the logistics chain. Although the use of SECU containers was generally considered to lower the amount of reel damage, it was thought that too tight loading into the containers may also damage the reels.

The Stora Enso specific SECU containers which the NETSS cargo is transported in must be transported back to the ports of loading for reuse. NETSS uses also an early release principle, which means that cargo is sent forward in the transport system as soon as possible after becoming available for transporting. If not properly managed, these features can lead to logistic problems such as crowded warehouses or lack of SECU containers at specific locations.

20.4.4 Risk Control Ideas

Once the top priority risks were identified, the next task was to decide how to address them. The participants were divided into small groups of 2–3 participants, now mixed from the earlier tasks in order to bring different backgrounds together. Each group could choose which risk to focus on first, and then move on to another one. Finally the risk control or avoidance ideas were briefly discussed together in the large group.

The risk control ideas consisted both of preventive measures concentrating on the root causes of the disturbances and of contingency plans to be executed in case a risk is realised. The main ideas can be summarised as follows:

- Proper training and instructions to the personnel throughout the NETSS system and motivating the personnel by explaining the effects of their work on the system.
- Optimisation of the system as a whole, avoiding sub-optimisation.
- Safety margins to make the system resilient.
- Key performance indicators and clear early warning signals on developing disturbances.
- Premade backup plans in case of e.g. delays or full warehouses. E.g. alternative routes, warehouses and vessels.
- Clear communication procedures to sort out problems.

20.4.5 Conclusion of the Workshop

In the conclusion of the workshop, the client representative informed the participants on the next steps in the process to manage the NETSS start-up risks and named the responsible persons for each task. A feedback channel was established using which Stora Enso staff could give comments and remarks concerning the upcoming NETSS phase 2 start-up. A workshop feedback questionnaire was also filled in by the participants.

20.5 Post Workshop Actions

After the workshop, a report was compiled by VTT describing the workshop process and the results. The top priority risks were briefly analysed and recommendations on follow-up activities were given. Preparatory actions to be done before the NETSS start-up as well as contingency plans to be executed in case a risk is realized were suggested. The report contained also all written and numerical inputs recorded by the computer system during the workshop.

Stora Enso used the identified risk scenarios to structure their preparation work for the start-up and bring attention to important questions. The scenarios were utilised to follow up the performance throughout the NETSS start-up phase. According to Stora Enso representatives, they benefitted greatly from the risk assessment. They faced some minor disturbances in the eventual start-up, but to most part it went very well.

20.6 Lessons Learned

Computerised workshops make it possible for people to participate not only face-to-face, but also over the Internet from different locations or even at different times, if taken into account in the workshop settings. In this case it was however decided to hold a face-to-face meeting where the group and work process are easier to manage than in distributed settings, despite the apparent challenge to arrange for all the desired experts to be at the same place at the same time. The participants in face-to-face workshops are usually highly motivated and can focus on the common task without external disruptions. One of the challenges in workshops realised entirely over the Internet is to ensure active participation by all parties. Remote workshops are also demanding with regard to the clarity of the process and quality of the instructions, when face-to-face instructions cannot be given.

The whole risk assessment workshop was carried out during one day, because it was considered too challenging to gather all the needed experts for a longer time period. This lead to an intense workshop and it turned out that the discussion on risk control ideas in the afternoon could not be given quite as much time as some of the participants would have desired. An optimal duration of the workshop could have been one and a half days, moving the risk control ideas session to the morning of the second day. Alternatively, the risk control ideas could have been covered in a separate session over the Internet about one week after the workshop, giving also the participants the possibility to provide input in between. Remote sessions are easier to realise in follow-ups when the group is already familiar with each other than if pursued with new groups.

Although the workshop itself took only one day, it gives only a part of the picture. A successful expert workshop requires thorough preparation, as described earlier. It is important that the workshop agenda meets its objectives. Each task must be feasible in the workshop setting and allocated enough time. It is also important that the participants can easily understand the purpose and requirements of the tasks. After the workshop, resources must also be reserved for reporting, editing of the gathered data and analysis of the results.

In the workshop, the participants worked in pairs sharing the same computer (with some exceptions with one or three persons). In the two first tasks people with similar background sat together, whereas the groups were mixed in the third task. The setting invited for small discussions within the groups while working with the tasks, resulting in probably a somewhat smaller number of inputs but also

inputs of high quality and fewer duplicates. People with similar background could support each other when brainstorming risk scenarios together, whereas groups with different backgrounds were best prepared to come up with new and innovative risk controls.

The workshop participants felt it was both useful and interesting to meet the other parties of the NETSS chain and go through the possible risks well in advance before the system start-up. The structured brainstorm methodology was well-received as it enabled addressing of a broad scope in an efficient way. Despite a quite large group, everyone had a chance to share their concerns on all topics discussed and quite a lot of discussion about the main topics was possible. A broad participation of main stakeholders made also the risk analysis results easily acceptable within the company.

References

Arena M, Arnaboldi M, Azzone G (2010) The organizational dynamics of enterprise risk management. Acc Organ Soc 35:659–675

Cooke RM (1991) Experts in uncertainty; opinion and subjective probability in science. Oxford University Press, New York, Oxford

Duijne F, van Aken D, Schouten EG (2008) Considerations in developing complete and quantified methods for risk assessment. Saf Sci 46:245–254

Hopkin P (2014) Fundamentals of risk management: understanding, evaluating and implementing effective risk management, 3rd edn. Kogan Page, London

ISO 31000 (2009) Risk management–principles and guidelines. International Organization for Standardization, Geneva

ThinkTank (2017) http://thinktank.net/. Accessed 14 Feb 2017

Weatherall A, Hailstones F (2002) Risk identification and analysis using a group support system (GSS). In: Proceedings of the 35th Hawaii international conference on system sciences

Weatherall A, Nunamaker JF (2000) Getting results with electronic meetings. Chandlers Ford, UK, EMSL

Chapter 21
Group Decision Support for Crop Planning: A Case Study to Guide the Process of Preferences Elicitation

Pavlos Delias, Evangelos Grigoroudis, and Nikolaos F. Matsatsinis

Abstract The land of Paggaio, Kavala, Greece although very rich, has been cultivated in ways that affected both local environment and economies disadvantageously giving rise to the crucial problem of strategic crop planning. However, because of the many actors involved, and of their conflicting interests, reaching a consensus about what the objectives of such a planning should be, is a complex and challenging task. So as a first, preparatory step for strategic crop planning, the interested parties should acquire a clear view about what are the differences in the preferences of the involved actors. In this chapter, we present the steps that we followed in order to execute an end-to-end process for a client that needed to unveil what are the criteria that guide the preferences of the actors and which actors converge (or diverge) the most, with respect to the evaluation on these criteria. Following a prescriptive approach (that assumes that a preference model exists), we sketched the relevant problem situation and problem formulation, constructed an evaluation model based on a multiple criteria technique, and eventually reached some recommendations. The case study we present in this work could help analysts to structure their own decision aid processes based on an established roadmap, as well as to become aware of the process pitfalls. Regarding the referenced case study, it showed that actors have strongly diverging preferences, so that it was not possible to discover a robust collective model. However, we were able to identify the points of major conflict in two criteria (environmental friendliness and economical performance) and amongst certain stakeholders.

P. Delias (✉)
Department of Accounting and Finance, Eastern Macedonia and Thrace Institute of Technology, Agios Loukas, Kavala, Greece
e-mail: pdelias@teiemt.gr

E. Grigoroudis • N.F. Matsatsinis
Decision Support Systems Laboratory, School of Production Engineering and Management, Technical University of Crete, University Campus, Kounoupidiana, 73100, Chania, Crete, Greece
e-mail: vangelis@ergasya.tuc.gr; nikos@ergasya.tuc.gr

L.C. Dias et al. (eds.), *Elicitation*, International Series in Operations Research & Management Science 261, https://doi.org/10.1007/978-3-319-65052-4_21

21.1 Introduction

Crop planning is about deciding how to allocate a finite amount of agricultural land among various competing crops that could be grown. Because it involves multiple actors (potentially with conflicting interests), and it affects and is affected by multiple criteria, it is a particularly complex problem (de Groot et al. 2010; Dury et al. 2012). A common tactic in crop planning is to optimize the allocation plan according to a certain set of objectives (Chetty and Adewumi 2014; Janová 2012), nevertheless in order to be able to optimize, the involved actors must have first decided about their objectives. This is exactly the aim of this paper: to try to unveil how the involved actors are making their decisions and to figure out what are the points of consensus or conflict in a process involving multiple actors.

Crop planning is a core problem in agriculture. Its implications are far-reaching and related to many aspects such as land-use (Dai and Li 2013), environmental impact (Núñez et al. 2013), market concerns (e.g., employment, food prices) (Li et al. 2015), etc. Therefore, it is expected that the actors involved have conflicting preferences. Acquiring an understanding of the various preferences, becomes therefore a vital issue for key stakeholders. However, eliciting preferences in agricultural economics problems has been identified as a hard task itself (Adamowicz et al. 1998; Carson and Louviere 2011).

In this work, we describe an analytical, end-to-end process, to reach some answers concerning both the above research questions (how decisions are made and points of conflict). We build our plan on the multiple criteria decision aid paradigm that dictates some concrete steps for the problem structuring, as well as for the evaluation model construction. In particular, we present the process steps in tandem with a case study in an effort to exhibit several pitfalls, and highlight the critical milestones.

In the next section, we analyze the contents of different "problem situations" hoping to support decision analysts to anticipate how their analysis would change assuming a specific problem situation. A rigorous perception of the problem situation should promote defining a fitting set of problem formulations among which the analyst and the decision maker could select the most applicable one. In Sect. 21.3 we present an instantiation of a problem formulation, based on the case study. The details of the end-to-end process are presented in Sect. 21.4, while a brief discussion concludes the paper.

21.2 Problem Structuring

In this section, and following (Bouyssou et al. 2006), we will take the view that a decision aiding process is a process "in which different agents endowed with cognitive capabilities have to share some information and knowledge in order to establish some shared representation of the process object". During the first steps

of the process, these pieces of information and knowledge take the shape of two major deliverables, namely the problem situation and the problem formulation. The former boils down to a representation that will ultimately aid the client to better arrange herself regarding the decision procedure for which she asked the analyst's recommendation. This representation has mainly a descriptive, elucidative nature. The latter (problem formulation), is actually a task of translating the client's interest into a format that decision support techniques and methods can address. This is reached by using a formal decision support language, however since this will inevitably lead to a reduced reality, we ought to point out the following pitfalls: A problem formulation is not neutral to the final recommendation (solution), indeed a different formulation is very likely to lead to a different recommendation. The analyst's defense of this is that following a problem formulation, the client will eventually be able to anticipate the possible conclusions and check whether these are compatible with her expectations. It is quite clear that the analyst shall not continue the decision aiding process, unless the problem formulation is validated by the client.

In this work, and in accordance with (Bouyssou et al. 2006; Morisio and Tsoukiàs 1997; Stamelos and Tsoukiàs 2003), we define a problem situation $\mathscr{P}$ to be a triplet $\mathscr{P} = \langle \mathscr{A}, \mathscr{O}, \mathscr{S} \rangle$, where $\mathscr{A}$ are the actors involved in the process (as per the client's as well as the analyst's points of view), $\mathscr{O}$ are the objects (problems, interests, opportunities, stakes) introduced by each actor (for instance in choosing a crop a farmer may be concerned with the corresponding profit, while a local citizen may be concerned by the environmental impact the crop may bring), and $\mathscr{S}$ are the resources (monetary or not) committed by each actor to each object of her concern. Another triplet representation is employed for the problem formulation. In particular, we define a problem formulation Γ to be a triplet $\Gamma = \langle \mathbb{A}, V, \Pi \rangle$, where $\mathbb{A}$ is the set of alternatives, the set of potential actions that the client may undertake, V is a set of points of view (dimensions) from which the potential actions are observed, analyzed, evaluated, compared, etc, and Π is the problem statement (what is expected to be done with the elements of $\mathbb{A}$- some common problem statements are choice, ranking, rejection, etc.).

21.2.1 Problem Types

21.2.1.1 Crop Choice (Farm Level)

Probably the most fundamental situation in agriculture is having to decide about what to grow at a particular farm. This situation is historically important, yet still currently relevant (Asrat et al. 2010; Gal et al. 2011; Kassie et al. 2011; Manos et al. 2013). We selected this problem mainly because of its illustrative power and its intuitiveness. Following the definitions of Sect. 21.2, we consider that:

$\mathscr{A}_{CH}$ Is the set of actors that will assume the consequences of the decision, namely the farm owner(s), as well as the actors that influence the decision:

labour workers (because of their availability, working rate, and knowledge), commission agents (because of their charging rates and their access to market channels), purchasers, and the agronomist.

$\mathscr{O}_{CH}$ Are the farm owners' objectives (commonly the monetary profit, but can include other objectives like risk mitigation, and commitment to networks), the bargaining power of buyers and commission agents.

$\mathscr{S}_{CH}$ Are the farm, the labour, the knowledge, the machinery and operation timing.

$\mathbb{A}_{CH}$ Is a set of possible decisions, i.e., a crop list (e.g., wheat, barley, etc.)

V_{CH} The evaluation dimensions include the farm's characteristics (soil, irrigation); physiological characteristics (fertilizers, tolerance to invaders, etc.); knowledge about the crop (variety, cultivation practices); market elements (prices, channels)

Π_{CH} The scope of the evaluation is rather clear in this situation. It is to choose the crop to be grown.

21.2.1.2 Crop Acreage

The *crop acreage problem* boils down to deciding what percentage of land to devote to every crop (the assignment of a particular crop to each plot in a given piece of land is known as the *crop allocation problem*, and is left out of the discussion of this section) (Dury et al. 2012). This problem situation assumes a higher-level authority that has the capacity and the interest to plan at a macro-level.

$\mathscr{A}_{AC}$ In addition to the actors of $\mathscr{A}_{CH}$ (farm owners, workers, commission agents, purchasers, agronomist), in this set we shall include local communities, analysts/facilitators, and policy makers.

$\mathscr{O}_{AC}$ We assume that the higher-level authority has an interest for the greater good, so the objective here is sustainable development (expressed over dimensions like economic growth, social cohesion, positive environmental impact, minimal conflicts (e.g., agriculture vs. tourism)).

$\mathscr{S}_{AC}$ In addition to the resources $\mathscr{S}_{CH}$ (farm, labour, knowledge, machinery and operation timing), in this set we shall count the political capital (of the high-level authority) and the social capital (as induced by the local population and its societal development).

$\mathbb{A}_{AC}$ The possible solutions include different crop portfolios, i.e., the total areas of land allocated for every crop.

V_{AC} The most common evaluation dimensions used in the literature (Chetty and Adewumi 2014) are water requirements, irrigation cost, lower and upper bounds for every crop, profit, environmental impact, proximity constraints, employment potential, and common agricultural policy (CAP) compatibility.

Π_{AC} The final evaluation is expected to rank the different portfolios. It is also possible to require for a classification of the portfolios into predefined categories (e.g., qualified, non-qualified).

21.3 Case Study

21.3.1 Background

The land of Paggaio, Kavala, Greece although very rich (after reclaiming a dried lake in 1930) has been cultivated in ways that affected both local environment and economies disadvantageously. The case study presented in this work aims to reveal the preferences of local stakeholders, and is based on the work of Delias et al. (2013). The need for a strategic crop planning for that land has emerged because of efficiency has declined significantly. Although no official scientific research on the reasons for this decline has been conducted, it is empirically known such a decline can be mainly justified by: the over-intensive farming that contributed to soil erosion, rendering the land non-productive; the habit of stubble-burning; the governmental and European Union subsidies that disrupted farmers' planning process. In addition, the turn towards more environmental friendly practices and in general towards sustainable development demands for updates in the planning.

Nevertheless, reaching a consensus about what the objectives of such a planning should be is far from being an easy task, mainly because local stakeholders have conflicting interests. Indeed, interests may vary even among neighboring farmers. So as a first, preparatory step for a strategic crop planning, the interested parties should acquire a clear view about what are the differences in the preferences of the involved actors. In the next section, we present in detail the end-to-end process that we followed in order to make this first step.

21.3.2 Process Overview

The starting point for analysis is the client, who even though has high domain knowledge, has a major concern and is actively looking for support. In our case, the client is a key stakeholder, namely the president of the coalition of the local agricultural cooperatives. He is evidently interested in acquiring some knowledge about the preferences of the involved actors for a crop planning, so he initiated a relevant project.

The next steps are not exactly linear, in the sense that although an ordering of steps is suggested (aiming at a consistent and progressive deployment), this ordering is not strict. Switching between steps is an expected as well as an essential part of the decision aid endeavor. This is why we avoid presenting the steps visually (e.g., with a flowchart). Anyway, expectedly, the next step is to define the *problem situation* and *formulation.* Therefore, we proceed by defining:

$\mathscr{A}_{PM}$ Farmers, agronomist, local stakeholders (authorities), citizens, downstream businesses (purchasers), decision analyst

$\mathscr{O}_{PM}$ $\mathscr{O}_{CH} \cup \mathscr{O}_{AC}$

$\mathscr{S}_{PM}$ The resources $\mathscr{S}_{CH}$, as well as those of $\mathscr{S}_{AC}$, plus the social image of actors.

$\mathbb{A}_{PM}$ Crops to be grown, belonging to categories such as energy crops-biofuels, aromatic or medicinal plants and herbs, forages-feeding stuffs, and horticulture plants. In addition, because of a regulation that was introduced in Greece at that time and which promoted the installation of photovoltaic parks in agricultural lands, we counted this option too.

V_{PM} Similar to V_{AC}.

Π_{PM} Ranking the alternatives, with the ultimate goal to disaggregate the preference model.

Having established the problem formulation, and in order to provide a formal answer for Π_{PM}, we need a representation for an *evaluation model* that will eventually guide the selection/construction of the appropriate options (Sadok et al. 2008). In Bouyssou et al. (2006), authors propose a 5-tuple to define an evaluation model, and here we adopt that proposal. The particular assignment of these variables for the case study is presented in Sect. 21.4. However, it is often the case that there is not a single method that could be applied to the defined evaluation model. In Roy and Słowiński (2013), authors present several key questions that can help the analyst choose a right method. The first question refers to the problem statement. In our case as Π_{PM} dictates, the method should be able to rank the set of alternatives either by assigning a numerical value (utility) to each of them, or without associating any numerical values. Then, the method should take into account the imprecision (or even the lack) of some data, that will eventually make the definition of performances profoundly subjective. Therefore, the method should provide a sensitivity and/or robustness analysis, that will respond to the great need of comprehension of the final solutions, and of explanations of the technique's functioning.

Moreover, a major characteristic of the problem's context is its hardness to collect preference information, since stakeholders are not willing to answer to numerous and protracted questions. Considering the convenience to get the preference information, as well as the level of imprecision in the definition of performances, in this work, we choose the method introduced in Delias et al. (2013), Delias and Matsatsinis (2013). That method accepts a compensation of a poor performance on one criterion by an exceptional performance on another criterion, and it assumes that there is not any form of interaction among the criteria, two elements that are acceptable within the problem context.

Then, and in order to collect the data, we prepared an evaluation table, namely a dashboard with the profiles of the alternatives, to show it to the involved actors. That was perhaps the most controversial step of the process, so in order to highlight the relevant pitfalls, we provide some explanations in the next section. Following the data collection, we applied the multi-criteria evaluation method, and got the initial results. Because the first round of results were not satisfactory, we performed a second round of data collection and re-ran the method. Finally, we were able to provide some recommendations to the client.

21.4 The Process Instantiation

As mentioned earlier, it is the client that triggers the process. In the previous section, we tried to summarize his objectives as "acquiring some knowledge about the preferences of the involved actors" and "disaggregate the preference model". To make things clearer, we present two particular questions that were the client's major concerns: (1) Are some criteria globally (or by an extended majority) considered as important/unimportant? (2) Which actors seem to converge/diverge the most? The client validated that the problem situation and formulation presented in Sect. 21.3.2 reflects his concerns, so we were able to proceed to the evaluation model definition. In particular, we represent the evaluation model $\mathscr{M}$ as $\mathscr{M} = \langle A, \{D, \mathscr{E}\}, H, \mathscr{U}, \mathscr{R}\rangle$, where A are the alternatives under evaluation, D is the set of dimensions (attributes) under which the elements of A are observed, described, measured etc, $\mathscr{E}$ can be considered as the "scales" of D, H are the ultimate evaluation criteria, $\mathscr{U}$ is the uncertainty (if any) associated to the available information, and $\mathscr{R}$ are the aggregation operators. The explicit list for A is:

- Cultivation of colza (to extract oil and exploit the cake left)
- Cultivation of white poplar (Populus alba – Salicaceae) for the paper industry and biofuels
- Sugar beets (cultivated Beta vulgaris) for biofuels and the food industry
- Helianthus (sunflower) to mainly be used as a biofuel
- Stevia for pharmaceutical or food industry
- Photovoltaic parks
- Barley for mash production
- Wheat for the same purpose
- Soybean also for mash production
- Maize
- Pomegranate for the food industry as well as for pharmaceuticals.

Then for $\{D, \mathscr{E}\}$ there are plenty of elements that we could observe (see for instance Bohanec et al. 2008). At this point, and in order to reach the set H, we had vigorous discussions with the client about the aggregation operators that we should apply. The client insisted that the issues of completeness and uncertainty for the set of observed elements are burdensome, and that he could not validate any formal aggregation method of D into H, so he advocated an empirical aggregation into a linguistic ordinal scale. The evaluation criteria set that was suggested comprised: Environment friendliness, exploitation of natural resources, land reuse potential, economic performance, available information, investment attractiveness. All of them were characterized by a 5-level ordinal scale with descriptive labels such as: {Negative outlook, Poor performance, Moderate performance, Satisfactory performance, Outstanding}. It is evident that such an empirical measurement withholds inevitable imprecisions for the criteria performance evaluation. Let us illustrate this with two examples: First, by enforcing 5-level scales for every criterion allows manipulating the evaluations that are close to the bounds. For example, "Wheat

for mash production" was ranked as "Outstanding" with respect to the "land reuse potential". The client let us know that wheat (as well as barley and colza) are the top-performers among the alternatives, yet he was unsure if the term "Outstanding" fit the description of their performance (there existed drawbacks in every alternative). However, to differentiate their performance from the other alternatives, he assigned the top-level evaluation to them. If we had preferred a scale with many levels (e.g., 10), then we wouldn't have put the top mark. A second example of inevitable imprecisions is the following: "Maize" and "Pomegranate" were both ranked as "Outstanding" (the top level) with respect to the "investment attractiveness" criterion. However, pomegranate had greater potential to attract investors than maize (due to special circumstances, such as a local, famous investor who had at that time an established interest to promote pomegranate to the food industry). Nevertheless, since the criterion scale was bounded from above, this superiority could not be reflected. In other words, since scales are bounded above and below (by the highest and the lowest points), equality in a characterization of one criterion for two alternatives does not necessarily imply equality in performance. Nevertheless, this tactic was fully endorsed by the client, who insisted on having simple, jargon-free scales, and claimed that imprecision of performances are an inherent element of the problem. After all, it allowed us to fill the evaluation table. The evaluation table is actually a list of profiles of performances, i.e., every alternative was evaluated over every criterion. The data of the evaluation table originated from client's domain knowledge, and were considered as certain.

That table was shown to interviewees during the interviews. In particular, interviews were conducted at the work (or home) places of the interviewees. We managed to conduct six interviews by picking actors of the $\mathscr{A}_{PM}$ set. More specifically, the interviewees were an agronomist, the president of local cooperatives, a feed mill owner, two farmers, and a local resident. The particular individuals were selected on a convenience basis through a shortlist provided by the client. During an interview, the corresponding stakeholder and the analyst first discussed the general importance of a strategic crop planning. Then the analyst explained the objectives of the study and demonstrated the evaluation table to the interviewee. The interviewee had to declare her holistic preferences in terms of pairwise preference relations (e.g., sugar beets is an alternative *at least as good as* colza) and in terms of intensities of those preferences (e.g., my preference of sugar beets to colza is more intense than my preference of stevia to soybean). For instance, an interviewee (the agronomist) declared the following preferences:

- *Considering the holistic preference,* "Photovoltaic parks" is at least as good as "Maize"
- "Photovoltaic parks" is at least as good as "Soybean for mash production"
- "Photovoltaic parks" is at least as good as "Wheat for mash production"
- "Maize" is at least as good as "Wheat for mash production"
- "Maize" is at least as good as "Soybean for mash production"
- "Wheat for mash production" is at least as good as "Soybean for mash production"

- My preference for "Photovoltaic parks" to "Soybean for mash production" is more intense than my preference for "Photovoltaic parks" to "Wheat for mash production"
- My preference for "Photovoltaic parks" to "Wheat for mash production" is more intense than my preference for "Maize" to "Soybean for mash production"
- My preference for "Maize" to "Wheat for mash production" is more intense than my preference for "Wheat for mash production" to "Soybean for mash production"

An important aspect for data collection is that the interviewee did not need to declare explicitly his preferences (or intensities of preferences) for all pairs, but just for the pairs for which she could determine such a relationship. We should also stress that the stated preferences refer to the alternative as a whole, and that it is not required to ask a separate question for each criterion. This is an intrinsic part of the method, which proved to be very useful, because interviewees were willing to answer some questions, but not too many.

The evaluation method that we applied (Delias et al. 2013; Delias and Matsatsinis 2013) can be classified in the family of aggregation-disaggregation methods (Siskos et al. 2005). It solves a set family of linear programs to assess additive utility functions for all the criteria, and consequently, the criteria significance weights. Ultimately, the method assumes that the preference model of the stakeholders can be represented by the pertinent utility functions, a common assumption in the aggregation-disaggregation paradigm. In addition, among the outputs of the methods are the average stability indices (a global index, as well as partial indices for every criterion), and some metrics related to how divergent the involved actors are. More specifically, during the post-optimality stage, as many linear programs as the number of evaluation criteria are formulated and solved, which maximize repeatedly the significance weight of each criterion. The mean value of the weights of these linear programs is taken as the final solution, and the observed variance in the post-optimality matrix indicates the degree of instability of the results.

To give an impression of the first round results, we present them in Table 21.1. We shall note that weights are relative trade-offs and must sum to one, while for the ASI index, a value of 1 means perfect stability. The smaller the ASI index, the larger the deviation of the estimated weights during the post-optimality analysis. In practice, any value less than 0.6 signifies that stability of the results cannot be accepted (i.e., the deviation of the estimated weights is too large). That was the case during the

Table 21.1 First round results (adapted from Delias et al. 2013)

Criterion	Weight	ASI
Environment friendliness	0.17	0.44
Exploitation of natural resources	0.21	0.49
Land reuse potential	0.18	0.52
Economical performance	0.11	0.44
Available information	0.15	0.45
Investment attractiveness	0.18	1

first round. By examining in detail the linear program constraints, we were able to discover that this instability was mainly due to the preferences of the president of cooperatives, and the feed mill owner, who had the most divergent attitudes.

Hence, we started a second round of interviews with those two actors, plus the resident because her preferences set included a small amount of relations. During the second round of reviews we announced to the interviewees that results were not satisfactory, but we didn't disclose any details of convergency or divergence. In particular, we explained to the interviewees (in natural language) that the variation of the weights was too large to let us trust in them. We pointed out that probably the main reason for that was the insufficient preferences statements, and that a richer set of statements would hopefully produce more reliable results. Nevertheless, we emphasized that another reason for the low stability of results could be the inherent disagreement between the stakeholders, in order to be able to justify the non-improvement of the results, if that happened. Ultimately, we didn't ask from interviewees to change their minds, but to elaborate more on their statements, to reveal more of their preferences. All interviewees accepted that they could enrich their initial set, and duly did so. Interviewees could make any modifications to their initial set (add new ones, eliminate old ones, modify an existing preference or intensity, etc.). However, the second round input data resulted in an even worse performance of ASI (the criteria weights were only slightly modified).

The main implications from these results are twofold: First, the intuition of the president that there is no consensus among involved actors was confirmed. Indeed, the ASI indices were low after both rounds of interviews. We shall note that we did not disclose the preferences of the interviewees to each other after the first round, because we cared about assessing their preference model, and not about reaching a consensus. What happened during the second round was collecting additional declarations, which intensified the personal opinions of the interviewees. Should we have cared about reaching a consensus, this procedure would not have been fruitful, since we should have paid special attention in motivating each stakeholder to adjust her preferences with the rest.

Second, after inferring the criteria weights, we observe some kind of convergence just in the "Investment Attractiveness" dimension, while actors' preferences are really divergent concerning the " Environmental friendliness" and the "Economical Performance" dimensions. This is a popular pattern in agricultural communities (how to trade-off the economic performance with the ecological performance of a crop). We shall remind that significance weights are assessed through the LPs, and that there is no requirement for the interviewees to declare preferences for the same pairs of alternatives.

Last, because of the low stability, it is not safe to draw any conclusions for the significance of the criteria. A more thorough presentation of the case study results is out of the scope of this work, however we want to make clear that the final step (making recommendations to the client) is an essential step of the process. During this step, the analyst is expected to provide decision aid by participating in the final decision legitimization (Roy and Damart 2002). In particular, the analyst should be

able to enlighten and scientifically support decision-making, notably by Roy (1993): making the objective stand out more clearly from the less objective (e.g., point-out the limitations of the evaluation criteria scales); separating robust from fragile conclusions (e.g., explain what the implications of a low ASI are); and by avoiding the pitfall of illusionary reasoning by bringing out certain counter-intuitive results (e.g., a criterion with extremely high weight).

21.5 Discussion

The decision process is a complex process that starts with the definition of a problem situation, and eventually ends with the analyst's recommendation to the client. These recommendations should be based on formal models which by their turn, use data that are collected and manipulated during the process. It should be acknowledged that the construction of the formal model as well as data collection and manipulation are tasks that will inevitably lead to a simplification of a complex reality. We should therefore be realistic about our expectations for the precision of these models. This fact should not render the models useless in the analysts' minds, but rather increase their cautiousness in the elaboration of the recommendations. Regarding the current work, we presented the problem situation and formulation that we used. We also presented two relevant (and popular) problem formulations to emphasize the importance of this step, as well as to demonstrate to readers an established roadmap to structure their own formulations.

The end-to-end procedure, counting from the moment that the client announced definitely his desire to kick-off the project, was 4 months. The first month was dedicated to multiple meetings with the client to reach a validated problem formulation. Then, we needed 3 months to conduct the two rounds of interviews and analyze the data. The bulk of this period was spent in communication activities (e.g., contacting interviewees, scheduling interviews). It is worth noting that due to the special (agricultural) nature of the profession of some of the interviewees, there were several weeks during which their availability was very low. The first contact was performed by the client, who was endorsing the project to his contacts. As a result, all interviewees were positive in providing us with their preferences, but some of them were quite guarded in giving a large set of declarations. All of them claimed that their time was limited, and asked for short interviews.

It's worth mentioning two points that we vigorously debated with the client: The initial set of alternatives was very large. It comprised literally every possible crop that was ever cultivated in the area. We tried to convince the client to keep the set of alternatives small, because a large set would demand for a large set of declarations. Our argument was that the reference set should contain some indicative alternatives, and that we could extrapolate the results to every possible crop at a later phase. Finally, the client was convinced to drop some alternatives, yet in our opinion, the remaining set was still large. Some of the alternatives (colza, poplar, and stevia)

were not included in any of the pairwise comparisons. If we had the resources for additional iterations, we should have tried to repeat the interviews with a truncated set. Without this test, we can not be sure for the impact of the non-used alternatives on the final results.

A second point of debate was the scales for the evaluation criteria. At first, we took the view that the labelling of the scales should include descriptive sentences and not just a single word (for example, for the land reuse criterion, use a sentence such as "needs crop rotation or a set-aside scheme" instead of "Moderate"). The client's point of view was that since not all interviewees are experts, they will be daunted by a jargon-full labelling, especially by economic terms. In addition, he made the point that the imprecision of performance of several alternatives (or his inadequate knowledge about it) would render the filling of the evaluation table not possible, unless a simplified scale was employed. Those were two compelling arguments, that made us accept the client's point of view.

In the core of our procedure is the multiple criteria technique, inspired from the aggregation-disaggregation paradigm, which was applied to unveil the involved actors' preference model. We cannot stress enough that the process should not be built around the method, and that the method should not be linked to the problem formulation that has been adopted. In this case study, that particular technique was chosen because of its best fit to the client's requirements. Of particular importance for the technique selection was its intelligibility. The linear programs that are created, contain not only the utility functions variables, but variables for every pairwise comparison, and for every intensity declaration as well. This fact allows for a thorough analysis, since the magnitude of these variables indicate the stakeholders with the greatest divergence, and allow analysts to trace back the reasons for any inconsistencies.

Finally, it is clear that the model as well as the method are only suitable for the given client in the described, particular context. In that particular context, we assumed that all involved actors possessed a system of values independent of the process, so the prescriptive approach, as conducted by the proposed steps and techniques, appeared eminently suitable. However, it is within our future plans to enhance the technique by allowing different evaluation tables, one per interviewee. Last, we are considering how we could integrate the described procedures into a broader, longitudinal study.

References

Adamowicz W, Boxall P, Williams M, Louviere J (1998) Stated Preference Approaches for measuring passive use values: choice experiments and contingent valuation. Am J Agric Econ 80(1):64–75. doi:10.2307/3180269. doi:http://ajae.oxfordjournals.org/cgi/doi/10.2307/3180269

Asrat S, Yesuf M, Carlsson F, Wale E (2010) Farmers' preferences for crop variety traits: lessons for on-farm conservation and technology adoption. Ecol Econ 69(12):2394–2401. doi:10.1016/j.ecolecon.2010.07.006. doi:http://dx.doi.org/10.1016/j.ecolecon.2010.07.006

Bohanec M, Messéan A, Scatasta S, Angevin F, Griffiths B, Krogh PH, Žnidaršič M, Džeroski S (2008) A qualitative multi-attribute model for economic and ecological assessment of genetically modified crops. Ecol Model 215(1–3):247–261. doi:10.1016/j.ecolmodel.2008.02.016. doi:http://dx.doi.org/10.1016/j.ecolmodel.2008.02.016

Bouyssou D, Marchant T, Pirlot M, Tsoukiàs A, Vincke P (2006) Evaluation and decision models with multiple criteria: stepping stones for the analyst. In: International series in operations research & management science, vol 86 Springer, New York

Carson RT, Louviere JJ (2011) A common nomenclature for stated preference elicitation approaches. Environ Resour Econ 49(4):539–559. doi:10.1007/s10640-010-9450-x. http://link.springer.com/10.1007/s10640-010-9450-x

Chetty S, Adewumi AO (2014) Comparison study of swarm intelligence techniques for the annual crop planning problem. IEEE Trans Evol Comput 18(2):258–268. doi:10.1109/TEVC.2013.2256427. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6492247

Dai Z, Li Y (2013) A multistage irrigation water allocation model for agricultural land-use planning under uncertainty. Agric Water Manag 129:69–79. doi:10.1016/j.agwat.2013.07.013. http://linkinghub.elsevier.com/retrieve/pii/S0378377413001947

de Groot R, Alkemade R, Braat L, Hein L, Willemen L (2010) Challenges in integrating the concept of ecosystem services and values in landscape planning, management and decision making. Ecol Complex 7(3):260–272. doi:10.1016/j.ecocom.2009.10.006. http://linkinghub.elsevier.com/retrieve/pii/S1476945X09000968

Delias P, Matsatsinis N (2013) Multiple criteria decision aid and agents: Supporting effective resource federation in virtual organizations. In: Links, theory and applications. Wiley-Blackwell, New York, pp 273–284. doi:10.1002/9781118522516.ch11. doi:http://dx.doi.org/10.1002/9781118522516.ch11

Delias P, Manitsa P, Grigoroudis E, Matsatsinis N, Karasavvoglou A (2013) Robustness-oriented group decision support. a case from ecology economics. Procedia Technol 8:285–291, doi:10.1016/j.protcy.2013.11.038. doi:http://dx.doi.org/10.1016/j.protcy.2013.11.038

Dury J, Schaller N, Garcia F, Reynaud A, Bergez JE (2012) Models to support cropping plan and crop rotation decisions. A review. Agron Sustain Dev 32(2):567–580. doi:10.1007/s13593-011-0037-x. http://link.springer.com/10.1007/s13593-011-0037-x

Gal PYL, Dugué P, Faure G, Novak S (2011) How does research address the design of innovative agricultural production systems at the farm level? a review. Agr Syst 104(9):714–728 doi:10.1016/j.agsy.2011.07.007. doi:http://dx.doi.org/10.1016/j.agsy.2011.07.007

Janová J (2012) Crop planning optimization model: the validation and verification processes. Cen Eur J Oper Res 20(3):451–462, doi:10.1007/s10100-011-0205-8. http://link.springer.com/10.1007/s10100-011-0205-8

Kassie M, Shiferaw B, Muricho G (2011) Agricultural technology, crop income, and poverty alleviation in Uganda. World Dev 39(10):1784–1795. doi:10.1016/j.worlddev.2011.04.023. doi:http://dx.doi.org/10.1016/j.worlddev.2011.04.023

Li J, Rodriguez D, Zhang D, Ma K (2015) Crop rotation model for contract farming with constraints on similar profits. Comput Electron Agric 119:12–18. doi:10.1016/j.compag.2015.10.002. doi:http://dx.doi.org/10.1016/j.compag.2015.10.002

Manos B, Bournaris T, Chatzinikolaou P, Berbel J, Nikolov D (2013) Effects of CAP policy on farm household behaviour and social sustainability. Land Use Policy 31:166–181. doi:10.1016/j.landusepol.2011.12.012. doi:http://dx.doi.org/10.1016/j.landusepol.2011.12.012

Morisio M, Tsoukiàs A (1997) IusWare: a methodology for the evaluation and selection of software products. IEE Proc - Softw Eng 144(3):162

Núñez M, Pfister S, Antón A, Muñoz P, Hellweg S, Koehler A, Rieradevall J (2013) Assessing the environmental impact of water consumption by energy crops grown in Spain: LCA of water for energy crops in Spain. J Ind Ecol 17(1):90–102. doi:10.1111/j.1530-9290.2011.00449.x. doi:http://doi.wiley.com/10.1111/j.1530-9290.2011.00449.x

Roy B (1993) Decision science or decision-aid science? Eur J Oper Res 66(2):184–203. doi:10.1016/0377-2217(93)90312-B. http://linkinghub.elsevier.com/retrieve/pii/037722179390312B

Roy B, Damart S (2002) L' analyse coûts-avantages, outil de concertation et de légitimation? Metropolis 108/109:7–16

Roy B, Słowiński R (2013) Questions guiding the choice of a multicriteria decision aiding method. EURO J Dec Process 1(1-2):69–97. doi:10.1007/s40070-013-0004-7. doi:http://dx.doi.org/10.1007/s40070-013-0004-7

Sadok W, Angevin F, Bergez JÉ, Bockstaller C, Colomb B, Guichard L, Reau R, Doré T (2008) Ex ante assessment of the sustainability of alternative cropping systems: implications for using multi-criteria decision-aid methods. a review. Agron Sustain Dev 28(1):163–174, doi:10.1051/agro:2007043. doi:http://dx.doi.org/10.1051/agro:2007043

Siskos Y, Grigoroudis E, Matsatsinis N (2005) UTA methods. In: Multiple criteria decision analysis: state of the art surveys. International series in operations research & management science, vol 78. Springer, New York, pp 297–344

Stamelos I, Tsoukiàs A (2003) Software evaluation problem situations. Eur J Oper Res 145(2):273–286. doi:10.1016/S0377-2217(02)00534-9. http://linkinghub.elsevier.com/retrieve/pii/S0377221702005349