GUIDANCE OF EFSA

Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment[1]

European Food Safety Authority[2,3]

European Food Safety Authority (EFSA), Parma, Italy

ABSTRACT

Quantitative risk assessments facilitate the decisions of risk managers. In the EU, risk assessment in food and feed safety is the responsibility of the European Food Safety Authority (EFSA). Quantitative risk models should be informed by systematically reviewed scientific evidence, however, in practice empirical evidence is often limited: in such cases it is necessary to turn to expert judgement. Psychological research has shown that unaided expert judgement of the quantities required for risk modelling - and particularly the uncertainty associated with such judgements - is often biased, thus limiting its value. Accordingly methods have been developed for eliciting knowledge from experts in as unbiased a manner as possible. In 2012, a working group was established to develop guidance on expert knowledge elicitation appropriate to EFSA's remit. The resulting Guidance first presents expert knowledge elicitation as a process beginning with defining the risk assessment problem, moving through preparation for elicitation (e.g. selecting the experts and the method to be used) and the elicitation itself, culminating in documentation. Those responsible for managing each of these phases are identified. Next three detailed protocols for expert knowledge elicitation are given - that can be applied to real-life questions in food and feed safety - and the pros and cons of each of these protocols are examined. This is followed by principles for overcoming the major challenges to expert knowledge elicitation: framing the question; selecting the experts; eliciting uncertainty; aggregating the results of multiple experts; and documenting the process. The results of a web search on existing guidance documents on expert elicitation are then reported, along with case studies illustrating some of the protocols of the Guidance. Finally, recommendations are made in the areas of training, organisational changes, expert identification and management, and further developments of expert knowledge elicitation methodology within EFSA.

KEY WORDS

expert knowledge elicitation, Delphi, Sheffield method, Cooke’s method, food safety, feed safety, risk assessment

---

[1] On request from EFSA, Question No EFSA-Q-2011-00850, approved on 5 July 2011.
[2] Correspondence: amu@efsa.europa.eu
[3] Acknowledgement: EFSA wishes to thank the members of the Working Group on Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment: Fergus Bolger, Anca Hanea, Anthony O’Hagan, Jeremy Oakley, Gene Rowe and Meike Wentholt for the preparatory work on this scientific output and EFSA staff: Elisa Aiassa, Fulvio Barizzone, Eugen Christoph, Andrea Gervelmeyer, Olaf Mosbach-Schulz, Sara Tramontini for the support provided to this scientific output.

Suggested citation: EFSA (European Food Safety Authority), 2014. Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment. EFSA Journal 2014;12(6):3734. [278 pp.] doi:10.2903/j.efsa.2014.3734

Available online: www.efsa.europa.eu/efsajournal

TABLE OF CONTENTS

Abstract 1
Table of contents 2
Executive Summary 6
Background as provided by EFSA 13
Terms of reference as provided by EFSA 14
Context of the scientific output 15
Evaluation 18

Part I: Introduction 18
1. The relevance of expert knowledge elicitation methods to food and feed safety risk assessment 18
 1.1. Introduction to food and feed safety risk assessment 18
 1.2. The evidence base for risk assessment 20
2. Introduction to expert knowledge elicitation (EKE) 23
 2.1. What is EKE? 23
  2.1.1. Knowledge 23
  2.1.2. Uncertainty elicitation 23
  2.1.3. Uncertainty, variability and subjective probability 25
  2.1.4. Practical elicitation 26
 2.2. Why expert knowledge elicitation? 27
  2.2.1. Expert knowledge elicitation for risk assessment 28
  2.2.2. The benefits of quantifying uncertainty 28
 2.3. Challenges 30
  2.3.1. A diverse literature 30
  2.3.2. Psychological factors when eliciting probability judgements from an expert 30
  2.3.3. Some important heuristics and biases 30
  2.3.4. Psychological factors when working with several experts 32
  2.3.5. Practical challenges 32
 2.4. Choices 32
  2.4.1. What to elicit 32
  2.4.2. Choice of experts 33
  2.4.3. Managing the experts 33
  2.4.4. Eliciting judgements 34
  2.4.5. Fitting and aggregation 34
  2.4.6. Documentation and reporting 35
 2.5. Meeting the challenges 36
  2.5.1. Choices and challenges 36
  2.5.2. Protocols 36

Part II: Procedures and Methods 38
3. Initiation 39
 3.1. Problem definition 39
 3.2. Background information 40
 3.3. Identifying parameters for formal expert elicitation 41
  3.3.1. Parameters for which elicitation is not necessary 41
  3.3.2. Minimal assessment 42
  3.3.3. Sensitivity analysis 43
 3.4. Conditions and resources for the elicitation 44
 3.5. Proposal for the Steering Group 44
 3.6. Deliverables 45
 3.7. Evaluation 45
4. Pre-elicitation 45
 4.1. Specification of the elicitation question 45
 4.2. Definition of the expertise profiles 47

4.3. Selection of the expert panel ... 50
4.4. Decision on the expert elicitation method ... 57
4.4.1. Generic considerations ... 57
4.4.2. Context-specific considerations ... 58
4.5. Proposal for the Elicitation Group (elicitor) ... 60
4.6. Deliverables ... 61
4.7. Evaluation ... 61
5. Elicitation ... 61
5.1. Invitation of experts for elicitation ... 61
5.2. Training on probabilistic judgements ... 64
5.2.1. Delivery of training ... 67
6. Execution of the elicitation protocol ... 67
6.1. Sheffield method ... 67
6.1.1. Overview of the Sheffield method ... 67
6.1.2. Preparation for the workshop ... 68
6.1.3. Completing the SHELF1 form ... 70
6.1.4. The quartile method ... 75
6.1.5. Conclusion and post-workshop reporting ... 84
6.1.6. Qualities of the elicitor ... 84
6.2. Cooke's method ... 85
6.2.1. Preparation ... 87
6.2.2. List of participants/roles ... 89
6.2.3. Equipment ... 90
6.2.4. Seed variables ... 91
6.2.5. Timeline of the workshop ... 93
6.2.6. Elicitation ... 94
6.2.7. Analysis ... 94
6.2.8. Robustness ... 96
6.2.9. Feedback and documentation ... 97
6.3. Delphi method ... 101
6.3.1. Step 1. Choose survey media ... 102
6.3.2. Step 2. Develop the survey ... 102
6.3.3. Step 3. Pilot survey ... 106
6.3.4. Step 4. Estimated timeline for expert involvement ... 108
6.3.5. Step 5. Execute Delphi survey ... 109
6.3.6. Step 6. Collection of results and analysis ... 110
6.3.7. Step 7. Subsequent Delphi rounds: repeat steps ... 113
6.3.8. Post-elicitation reporting ... 116
6.4. Deliverables ... 117
7. Post-elicitation ... 117
7.1. Technical documentation ... 117
7.2. Summary on the use of the elicitation results ... 118
Conclusions and Recommendations ... 119
8. Implementation of EKE in EFSA ... 119
8.1. Training and enabling ... 119
8.2. Organisational structure ... 120
8.3. Expert identification and management ... 120
8.4. Extensions of the EKE process ... 121
References ... 122

Appendices 130

A. Principles and Practical Recommendations 131
A.1. Principles and practice of problem, process and protocol definition 131
A.1.1. Search for evidence (identification of information deficit) 131
A.1.2. Defining the question for the experts 131
A.1.3. Standard scales 134
A.1.4. How to select the most suitable method for this problem 136
A.2. Principles and practice of selecting and motivating experts 138
A.2.1. Rationale for chapter 138
A.2.2. Defining expertise 138
A.2.2.1. Common definitions 138
A.2.2.2. Expertise continuum 140
A.2.2.3. Granularity and scope of expertise 140
A.2.2.4. Types of expertise 141
A.2.2.4.1 Procedural versus declarative 141
A.2.2.4.2 Theoretical versus practical 141
A.2.2.4.3 Substantive versus normative 141
A.2.2.5. Principles and practice of defining expertise 142
A.2.2.6. Identifying experts 142
A.2.2.7. Screening experts 151
A.2.2.8. Limits of expertise (what we can reasonably expect) 151
A.2.2.9. Principles and practice of identifying experts 152
A.2.3. Expert recruitment and retention 153
A.2.3.1. Where to find experts (e.g. databases) 153
A.2.3.2. How to approach experts (importance of identity of messenger) 154
A.2.3.3. Motivating experts 154
A.2.3.4. Principles and practice of recruitment and retention 156
A.2.4. Constructing groups of experts 157
A.2.4.1. Heterogeneity of opinion 157
A.2.4.2. How many experts? 159
A.2.4.3. Principles and practice of constructing expert groups 160
A.2.5. Conclusions 160
A.3. Principles and practice of eliciting a probability distribution 163
A.3.1. Elicitation roles 163
A.3.2. Representing a probability distribution 163
A.3.3. The strategy for eliciting a probability distribution. 165
A.3.4. General principles 166
A.3.5. What questions to ask 167
A.3.5.1. Eliciting points on the CDF: the options 167
A.3.5.2. Practicalities 170
A.3.5.3. Using the fixed interval and roulette methods 170
A.3.5.4. Using the variable interval method 170
A.3.6. Fitting and feedback 171
A.3.6.1. Fitting a distribution 171
A.3.6.2. Choice of distribution 173
A.3.7. Checking the fitted distribution 174
A.3.8. Dependence and multivariate elicitation 174
A.3.9. Elicitation software 175
A.3.10. Imprecision 175
A.4. Principles and practice of handling multiple experts – Interactions and aggregations 176
A.4.1. Introduction 176

A.4.2. Behavioural aggregation ... 177
A.4.3. Mathematical aggregation ... 178
A.4.3.1. Bayesian methods ... 178
A.4.3.2. Weighted combinations of probabilities ... 179
A.4.3.2.1 Cooke's Model ... 179
A.4.4. Mixed techniques ... 181
A.4.5. Final remarks ... 183
A.5. Documenting the method applied and the results ... 184
A.5.1. The result report ... 184
A.5.2. The technical support document ... 186
A.5.3. The expert feedback ... 190

B. Extensive Web search on Guidance Documents on expert knowledge elicitation ... 191

C. Case Study in Animal Health and Welfare ... 200

D. Case Study in Plant Health ... 202

E. Case Study in Biological Hazards ... 257

Glossary ... 275
Abbreviations ... 278

EXECUTIVE SUMMARY

In the European Union, risk assessment in food and feed safety is the responsibility of the European Food Safety Authority (EFSA) - the outcomes of EFSA's risk assessments facilitate the decisions of risk managers. The use of comprehensive quantitative models for risk assessment is well established, and the approach will be increasingly adopted by EFSA in the future. These models should be informed by systematically reviewed scientific evidence. However, empirical evidence is often limited, absent, conflicting, of doubtful relevance, open to alternative interpretations or not accessible via publicly available information sources (e.g. bibliographic databases, scientific journals, or websites). In such cases, reliable information can be obtained from knowledgeable experts using systematic and standardized methods for eliciting their knowledge. Even where there is sufficient empirical evidence, consultation with relevant experts can be beneficial, for instance, at the start of a process, to identify relevant information sources of which EFSA may be unaware, and to identify other experts who might be able to address deficits.

In 2012, EFSA established a working group to develop guidance on expert knowledge elicitation[4]: their task was to propose a practical process to elicit from knowledgeable experts quantitative parameters and their uncertainties in a probabilistic way.

This Guidance Document is organised in three parts:

- An Introduction that presents the risk assessment context and the motivation for the use of expert knowledge elicitation;

- Detailed Procedures and Methods. In particular, three protocols for expert knowledge elicitation are given in detail - these protocols can be used as guidance for the development of adapted protocols for real-life questions in food and feed safety. Some Conclusions and Recommendations are also presented on how expert knowledge elicitation could be implemented into EFSA's daily work.

- Appendices. Appendix A discusses different principles on framing the question, selecting the experts, eliciting uncertainty, and aggregating the results of multiple experts. Appendix B describes the results of a web search on existing guidance documents on expert elicitation. Appendices C and D are case studies illustrating some of the protocols of the guidance.

We will now summarize each part.

Introduction (Part I: Chapters 1 and 2)

Risk assessment in food and feed safety is a method to estimate the likelihood of occurrence of adverse effects on human health, animal health and welfare, plant health and the environment resulting from exposure to hazards. For instance, a risk assessment for contaminants in the food chain might seek to answer the following questions: What is the potential of the contaminant to induce cancer in humans (i.e. hazard identification)? What is the toxicity of the contaminant in its target organ in humans (hazard characterisation)? How likely is the contaminant to be consumed in food (exposure assessment)?

Quantitative risk assessment involves developing a model (i.e. a simplified representation) of how a hazard might cause adverse effects, estimating values of model parameters, and identifying attendant uncertainties related to exposure or to a particular hazard; models can be very complex, with many parameters, or rather simple with just a few parameters. The types of scientific information available to answer the questions generated by any risk assessment model can be classified into three main categories, which represent the evidence base for risk assessment: (1) empirical evidence from primary

[4] See Glossary for specific terminology used in this Guidance.

research studies available in the open literature; (2) raw data from suitable national and international databases; and (3) expert knowledge. The primary focus of this Guidance is the last of these three.

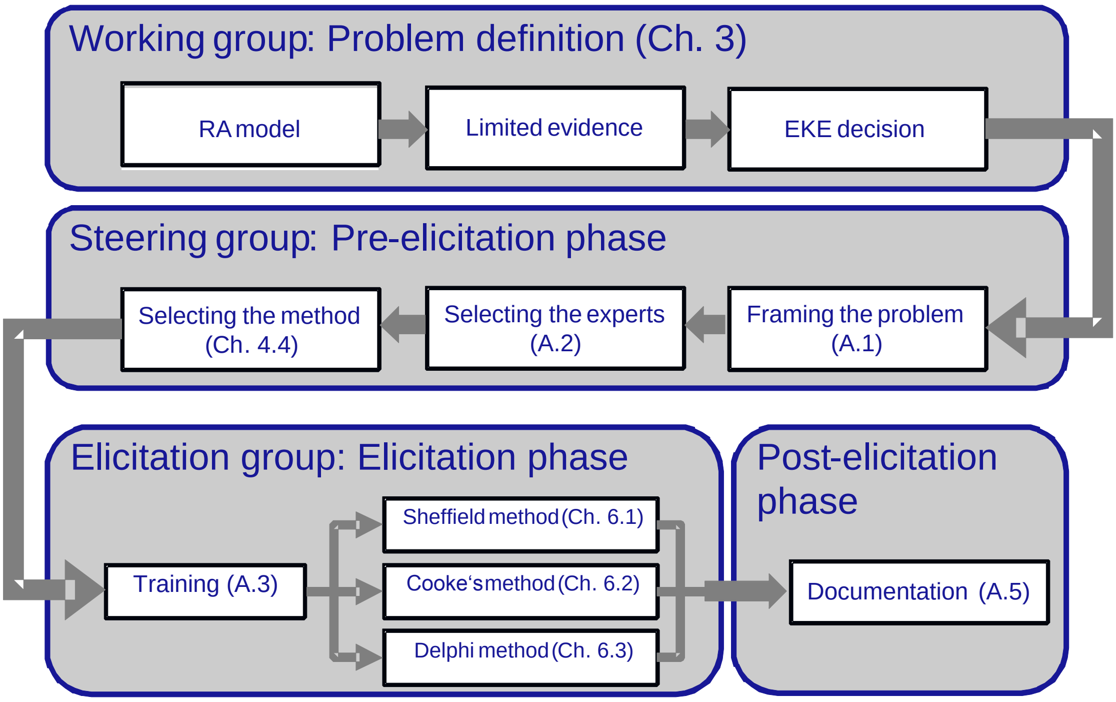
Expert knowledge elicitation (EKE) refers to the drawing out of knowledge from one or more experts. Experts can be asked for specific information (facts, data, sources, requirements, etc.) or for judgements about things (preferences, utilities, probabilities, estimates, etc.). Elicitation of specific information is relatively simple; the expert either does or does not know the answer, so two experts who both know the answer should give exactly the same values. Eliciting judgements is more challenging as we wish the expert to use his or her expertise to, for instance, make an estimate of an uncertain quantity; now it is clear that different experts can give different answers. In the case of eliciting information we must select experts who know the right answer, whereas in eliciting judgements we need experts who not only have skill at estimation, but who can also give realistic judgements as to the accuracy of their estimates. In this Guidance we are primarily interested in expert estimation of *uncertain quantities* (as facts can generally be found in the literature); in this case we do not simply want the 'best guess' of the quantity, but also a representation of the uncertainty surrounding any guess.

Uncertainty can arise for many reasons (e.g. natural variability or lack of information), and also be expressed in many ways (e.g. linguistically or numerically), which are briefly discussed. In this Guidance Document we concentrate on uncertainty from whatever source expressed as *probabilities* that are *personal* or *subjective* to individual experts (i.e. an expression of their beliefs). In particular, we focus on the elicitation from experts of *probability distributions* surrounding quantities being assessed (e.g. toxicity of a contaminant). Probability distributions are built up from a series of expert judgments about ranges of the uncertain quantity containing the true value with a particular probability (e.g. the toxicity values exceeding the actual toxicity value with 95%, 50% or 5% probabilities).

The pros and cons of eliciting probabilities in general, and probability distributions in particular, relative to other methods of representing uncertainty, are discussed and justification is given for both the quantification of uncertainty, and engagement in expert knowledge elicitation in the first instance. A number of challenges to expert knowledge elicitation are also enumerated, for instance, the existence of well-documented psychological biases in expert judgement, and social-psychological processes affecting the quality of outcomes from interacting groups: solutions to these challenges are proposed in this Guidance. Further, a number of practical challenges to the effective use of expert knowledge elicitation are introduced here in the form of a set of choices that need to be made: What to elicit? Which experts? Which elicitation method? How to aggregate expert judgements for policy makers? How to document and report the process? Again solutions to these challenges are proposed in this Guidance.

In order that the reader can better understand this Guidance we think it is useful to first present the expert knowledge elicitation process from beginning to end: (see Figure 1.).

The next part of the Guidance provides protocols for specific paths through the process, each tied to a concrete example, while the Appendix A focuses on the principles of good practice at five key points in this overall process.



(Ch. 4.4 = Chapter 4.4; A.1 = Appendix A, chapter 1 of the Guidance etc.)

Figure 1: The process of expert knowledge elicitation

Protocols

This part of the guidance provides the reader with concrete examples, and describes in detail three different protocols which incorporate all the steps of an EKE exercise:

1. the Sheffield protocol with group interaction of experts (behavioural aggregation);

2. the Cooke protocol with use of seed questions for the calibration of experts (mathematical aggregation);

3. a Delphi protocol on written individual expert elicitation (i.e. remote) with feedback loops (mixed behavioural and mathematical aggregation).

Each protocol is illustrated using the same hypothetical example of a risk assessment of a non-toxigenic foodborne bacterial pathogen, which was newly identified at a border inspection.

This section of the guidance can be considered a “cookbook” containing three different recipes that can be used to plan and conduct an elicitation process: it describes the elicitation process step by step, defining milestones, sub-tasks, responsibilities and outputs. The elicitation process can be structured into four phases (see Figure 2): initiation (Chapter 3); pre-elicitation (Chapter 4); elicitation (Chapters 5 and 6); post-elicitation (Chapter 7). The three different Protocols are specified In detail at the elicitation phase only. In the other three phases we give directions that are common to all the protocols. The phases are conducted consecutively by different project teams with well-defined deliveries. This structure should allow EFSA to outsource parts of the elicitation process to specialised contractors.

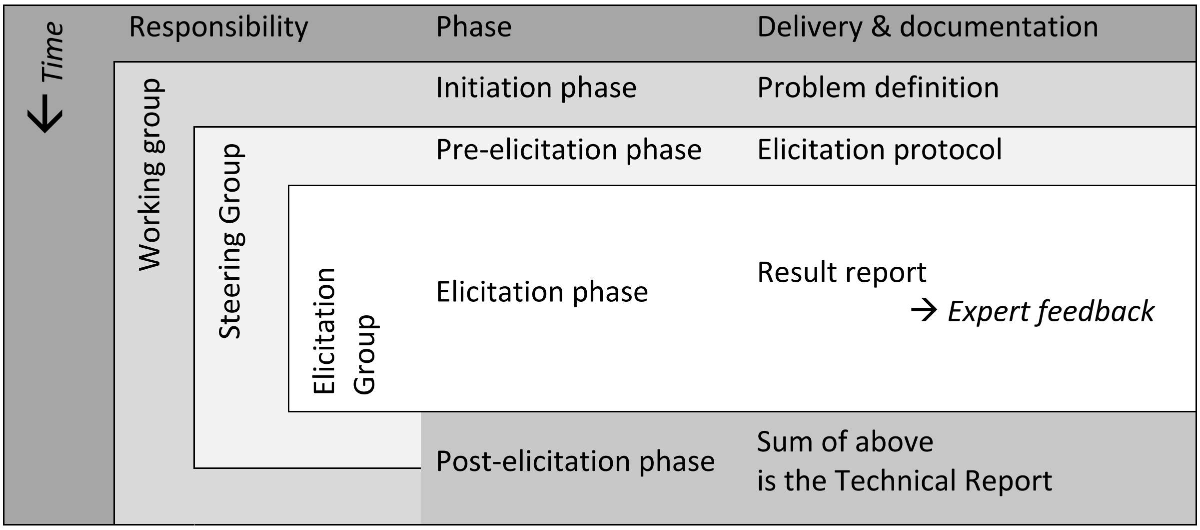


Figure 2: Phases of the elicitation process and related responsibilities

At each phase detailed examples of the EKE process and instruments/materials are given including:

- specific questions to be put to experts and examples of expert profiles
- letters and questionnaires to be sent to experts for recruitment purposes
- example training and elicitation materials appropriate to each elicitation method (e.g. Delphi - questionnaire; Sheffield - workshop protocols; Cooke - seed variables)
- sample documentation and feedback to experts.

Chapter 6 also details existing free software solutions to enable readers to carry out parts of their own EKE exercise, especially elicitations and calculations.

Chapter 8 presents the conclusions and recommendations in the areas of training, organisational changes, expert identification and management, and further developments of expert knowledge elicitation methodology in EFSA.

The conduct of an expert knowledge elicitation exercise needs both education on the techniques and also experience in concrete applications, which cannot be given by this guidance: the Working Group therefore recommends further training for Steering Group members and EFSA staff in Chapter 8.

Finally, the appendices present concrete case studies of expert elicitation already performed or planned within the current EFSA remit in:

- the Animal Health and Welfare team ; and
- the Plant Health team within EFSA's ALPHA unit.

Principles (given in Appendix A)

First a problem (i.e. a potential threat to food and/or feed safety) must be identified by EFSA and a Working Group (WG) established to explore this problem (see Figures 1 and 2). Initial tasks of the WG are to search for evidence regarding the nature and extent of the problem and to develop a model relating hazards to harmful effects; during the course of their investigations the WG may identify a lack of empirical data (or perhaps gaps in expertise on the WG, for instance, with regard to formulation of the model), indicating a need for expert knowledge elicitation.

Once a need for expert knowledge elicitation has been established then we recommend the establishment of a Steering Group (SG) to manage the expert knowledge elicitation. The SG will initially be composed of some members of the WG but later will also have amongst its members the person or persons responsible for carrying out the expert knowledge elicitation (i.e. the *elicitor(s)*). The SG then must determine which model parameter(s) (or other aspects of the model) will benefit from expert knowledge elicitation. It will often be the case that the model contains a large number of parameters whose uncertainty should be quantified. Full and careful EKE may be impractical for all of these parameters, but is necessary for at least some of them. Full EKE should then be employed for a subset of the most important and influential parameters. A "minimal assessment" method is proposed to provide a preliminary quantification of uncertainty for all the parameters, so that a sensitivity analysis can be carried out to identify the parameters to be the subject of full EKE. The minimal assessment is retained for the other, less influential, parameters.

The SG must also construct a set of questions to be posed to the experts regarding the chosen parameters. An important aspect to constructing the questions is to determine the form in which any quantities will be expressed (i.e. the measurement scales): scales should be familiar to the chosen experts yet easily converted to standard metrics required for the models. The SG must also determine the general methods to be used for expert knowledge elicitation (e.g. single vs. multiple experts, remote vs. face-to-face elicitation, mathematical vs. behavioural aggregation - see below). Issues involved in the EKE process up to this point are addressed in Chapter A.1 of the Guidance: 'Principles and practice of problem, process and protocol definition'.
Once the SG has decided on the form of the questions to be asked of experts, a search for experts must be carried out (possibly the general or specific nature of the expert knowledge elicitation methods to be used may be determined at this point, although in some cases these might be decided later). The form of questions to be put to experts guides the search process through the definition of *expert profiles*; how this is done, and other issues involved in expert selection, such as identifying appropriate experts, recruiting and motivating them, and constructing groups of experts are discussed in Chapter A.2 of the Guidance: 'Principles and practice of selecting and motivating experts'.

Selection of the expert knowledge elicitation *method* will normally be performed by the SG once the experts have been identified, as then the number of experts, their physical location, their availability etc. will be known. However, if for some *a priori* reason a particular method is preferred, then this could help guide the selection process. Further, experts may provide input that feeds back into the decision making processes of the WG and SG, thereby impacting on the model, parameter selection, expert profiles and so on. Once an expert knowledge elicitation method is selected then a qualified person (or persons) must be selected to act as elicitor(s); these will be persons who are themselves experts in the particular elicitation method. Elicitors should become a sub-part of the SG - known as the *Elicitation Group* - and may help redefine the parameters and thus the questions to be put to experts (and perhaps inform the search for additional experts).

The next stage in the expert knowledge elicitation process is the elicitation itself; Chapters A.3 and A.4 of the guidance respectively address two important aspects of the elicitation, namely: the principles and practice of eliciting a probability distribution; and of dealing with multiple experts. As we have already indicated, the elicitation of probability distributions is performed by a person that we refer to as an 'elicitor' - this is a person who has training in specific methods for eliciting probability distributions: some specific methods are described in Chapter A.3 (these are the methods elaborated in Part II of the Guidance: 'Protocols ').

An elicitor is needed because many experts are unfamiliar with expressing uncertainty in the form of distributions and because, even when an expert understands the concept of a distribution, it can be very difficult for him or her to produce a distribution without help. Three common ways of representing probability distributions are described: (1) Specifying the cumulative distribution function (CDF); (2) Specifying the probability density function (PDF); or (3) Choosing a standard distribution and specifying the parameters (see chapter A.32).

One significant problem for elicitation is that an expert cannot provide the *entire* distribution. The solution, common to most elicitation methods, involves a combination of all three representations of a probability distribution. First, the expert provides values of the CDF at a small number of points. Second, the elicitor constructs a probability distributions that complies with the experts assessments. Third, the elicitor plots the PDF corresponding to the choice in step two, to show the result of the elicitation to the expert (see chapter A.3.5).

Another problem for elicitation is that an expert's judgements can be biased by the way the elicitation is performed. A number of recommendations are made for preparing experts for an elicitation and structuring the elicitation so as to assist experts to produce unbiased distributions (see chapters A.3.5.2 and A.1.2).

If only one expert is used then his or her estimates of quantities and probability distributions can then be used directly to fill-out the model and thereby inform policy making; however, usually more than one expert is polled so as to increase the reliability and validity of judgements. Chapter A.4 'Principles and practice of handling multiple experts: Interactions and aggregations' discusses issues arising when several experts take part in an elicitation exercise.

If there are several experts it is nevertheless generally desirable to have just one set of estimates and distributions to enter into the model which raises the problem of how to *aggregate* over the estimates and judgements of individual experts. There are two basic approaches to aggregation: *behavioural* and *mathematical*: in the former case the aggregate is obtained by interacting experts reaching a consensus; in the latter, although experts may interact in some methods, aggregation is obtained by the elicitor, for instance, by using a linear combination of judgements. Mixed methods, such as the Delphi technique, described briefly here in Chapter A.4, and in more detail in Part II, involve some controlled interaction between experts and also some mathematical aggregation.

Behavioural aggregation presents various challenges in terms of managing the interaction in order that several well-documented biases in group processes (e.g. conformity effects, risky shift, dominating individuals etc.) do not detract from the obvious advantages of having multiple perspectives and pooled knowledge. Another problem for behavioural aggregation that is addressed in Chapter A.4 is what to do if experts cannot reach consensus. A well-designed protocol for the elicitation should help the elicitor in what otherwise would be a daunting task: The Sheffield Elicitation Framework (SHELF) outlined in this chapter and described in detail in Part II provides such a protocol for behavioural aggregation.

For mathematical aggregation to be performed it is not necessary for experts to interact (in fact, it may be preferable that they do not, so as to remove potential biasing effects of others); the main challenge here is then to select the most appropriate method for combining judgements. A controversial issue in mathematical aggregation is whether or not to give each expert's judgments equal weight. If differential weighting is performed then it is necessary to have some firm basis for giving some experts' opinions greater weight than others: Cooke’s method, introduced here and described in detail in Part II, uses experts' performance in judging 'seed variables' (variables related to the target variables, but for which the true answers are known) as the basis for differential weighting.

Once a single set of judgments has been produced the EKE process is nearly finished: all that remains is to document what has been done and what has been found in line with EFSA's policy on transparency. This Guidance has shown that EKE is rather a full process than a single method. The documentation therefore has to summarise all steps and decisions taken from the initiation until the final result. Documentation is the responsibility of different participants in the process rather than any one individual: reasoning of the need/use of expert elicitation is carried out by the Working Group, definition and evaluation of the protocol by the Steering Group, and documentation of conformity with the protocol and the results by the Elicitation Group (see Table 1).

Table 1: List of reports to document an expert knowledge elicitation

Type of report	Content/audience	Author
Result report	Summarises the results and will be used and published in the risk assessment procedure	Elicitation Group
Technical support document	Includes a full description of the process and enables the public to review the study	Working Group
	Decision for expert knowledge elicitation	Working Group
	Definition of the elicitation protocol and selection of experts	Steering Group
	Execution and documentation of the elicitation process	Elicitation Group
Expert feedback	Confidential documentation for the individual expert summarising the input from each expert	Elicitation Group

BACKGROUND AS PROVIDED BY EFSA[5]

In order to achieve the general objective of a high level of protection of human health and life, Regulation (EC) No 178/2002[6] recommends that food law be based on risk analysis and that risk assessments are undertaken by the European Food Safety Authority (EFSA) in an independent, objective and transparent manner, on the basis of all available scientific information and data. EFSA shall "(a) provide the Community institutions and the Member States with the best possible scientific opinions in all cases provided for by Community legislation and on any question within its mission", and "(b) promote and coordinate the development of uniform risk assessment methodologies in the fields falling within its mission".

Comprehensive food and feed safety risk models must be informed and implemented using reliable scientific evidence. All data used for the assessments, including unpublished data and personal communications, must be referenced and the data must be evaluated to determine their relevance to the assessment and quality. These should be reflected in the relative weight given to them in the assessment and taken into account in the overall evaluation of uncertainty (EFSA SC, 2009).

EFSA has analysed the use of systematic reviews to retrieve, appraise and synthesise publicly available and accessible scientific evidence (EFSA, 2010a) and, where possible, is implementing this method to inform food and feed safety risk assessment models.

The use of comprehensive models informed with systematically reviewed scientific evidence is especially recommended when the problems include parts with essential uncertainties (uncertainty analysis) and/or ambiguities (scenario analysis). Sometimes however empirical evidence may be limited (if not even absent) or not accessible via publicly available information sources (e.g. bibliographic databases, scientific journals or websites). For instance, for performing comprehensive risk assessments, often empirical evidence is neither available nor publicly accessible to estimate:

- specific geographical and temporal conditions;
- industrial handling and processing techniques;
- actually used processes in the food chain;
- specific risk assessment parameters (e.g. for prospective sample size calculation);
- animal or plant health pathways (e.g. *Tilletia indica* quantitative pathway analysis—EFSA PLH Panel, 2010b);
- exposure data assessment (EFSA SC, 2006b; Algers et al., 2009);
- future developments (e.g. emerging risks).

In such cases, reliable information can be obtained for instance from national authorities (e.g. internal reports, guidelines or practices); opinions of consultants or stakeholders, good practices in industries; specialists involved in the process; or other knowledgeable, skilled or trained persons[7].

The above is also supported by the International Risk Governance Council. It classifies the risk assessments as simple, complex, uncertain, ambiguous and recommends the involvement of stakeholders when the problem is uncertain or ambiguous (IRGC, 2005).

---

[5] Question No EFSA-Q-2011-00850, approved on 5 July 2011.

[6] Regulation (EC) No 178/2002 of the European Parliament and of the Council of 28 January 2002 laying down the general principles and requirements of food law, establishing the European Food Safety Authority and laying down procedures in matters of food safety. OJ L 31, 1.2.2002, pp. 1-24.

[7] These persons are also called "external experts" to distinguish the experts involved in expert knowledge elicitation from, for example, EFSA panel members, ad hoc working group experts or hearing experts.

Formal techniques for eliciting knowledge from specialised persons were introduced in the first half of the 20th century (e.g. Delphi method in 1946 or Focus groups in 1930—Ayyub Bilal, 2001) and after the sixties they became popular in risk assessments in engineering. Since then, research has been done to optimise the techniques and results obtained (Cooke and Goosens, 1999, 2000; Jones et al., 2001; Brown et al., 2001; Knol et al., 2010; and O'Hagan et al., 2006).

TERMS OF REFERENCE AS PROVIDED BY EFSA[8]

In view of the above, the EFSA guidelines on systematic review methodology should be complemented by a Guidance document on methodology for eliciting knowledge from experts (i.e. specialists, professionals or other knowledgeable persons) in food and feed safety risk assessment.

The Guidance should be practical and applicable to the different relevant food and feed safety fields. In particular, the EFSA Guidance should include:

- a review and discussion (including strength and limitations) of the existing approaches to eliciting knowledge from experts;
- a practical guidance on:
  - how to ensure neutrality and comprehensiveness in the choice of experts in a way such as to reduce biases and guarantee that the relevant expertise is covered.
  - how to conduct the process of elicitation of knowledge using a consistent and reproducible approach. The work should focus on quantitative questions and rankings. The method proposed in the Guidance should focus on how to minimise and/or analyse sources of bias and uncertainty (e.g. disagreements between the experts, different interpretations and/or variation within groups of experts).
  - how to extract reliable information (e.g. considering correction of bias, calibration, information on disagreement, variation and uncertainties).
  - how to document and present the results and the method applied to gather expert opinion, to allow peer review.
- a glossary of relevant terms.

The Guidance may also include recommendations for future research and/or improvement of existing methods.

A draft version of the Guidance should be made available for public consultation, to ensure all relevant information is taken into account to guarantee the reliability and consistency of the method described in the final document.

Part of this project should be a workshop, where a draft version of the EFSA Guidance on expert knowledge elicitation methods will be presented to and discussed with EFSA staff and panel members. Feedback from the workshop participants should be considered to finalise the Guidance and, in a subsequent phase, support the implementation process within EFSA.

8 Question No EFSA-Q-2011-00850, approved on 5 July 2011.

CONTEXT OF THE SCIENTIFIC OUTPUT

Approach to the Mandate

For the preparation of this guidance document, the Scientific Assessment Support (SAS) Unit of EFSA has set up a Working Group composed of six scientific experts specialised in elicitation methodology and psychology and six EFSA scientists covering a broad range of possible applications, e.g. in biological and chemical contamination and in plant or animal health and welfare.

During the initiation phase the members of the Working Group were introduced to the different risk assessment procedures applied in EFSA, followed by intensive discussion on the Terms of Reference. This resulted in a preliminary list of topics to be handled by the Working Group. It was discussed if the expertise of the Working Group was sufficient to cover all topics. This process was iterated until the final composition of the Working Group was completed.

The intended content of the guidance was evaluated

- by comparing it with the content of key references already identified during the initiation phase of the Working Group (ACERA, 2009; Ayyub, 2001; Cooke, 1991; Cooke and Goossens, 1999; Meyer and Booker, 2001; O'Hagan et al., 2006; RIVM, 2008; US-EPA, 2011, 2009; cp. appendix B);
- by referring to the results of an extensive web search on existing guidance documents for expert knowledge elicitation performed in spring 2012 (see appendix B).

The final structure of this guidance document comprises three parts:

I. an introduction to expert knowledge elicitation, giving an overview and motivation;

II. three protocols of expert knowledge elicitation, giving a step-by step description of three concrete processes (in the main part of the Guidance);

III. the principles of expert knowledge elicitation, discussing the advantages and disadvantages of different approaches (in appendix A).

Several means to ensure the applicability of the guidance were applied. Scientists from EFSA units on Biological Hazards (BIOHAZ), Plant Health (PLH) and Animal Health and Welfare (AHAW) were continuously involved in the preparation of the draft guidance. Additionally, from the beginning, the scientific expert members of the Working Group were also involved as hearing experts in other working groups on current risk assessments to discuss and perform case studies of expert knowledge elicitation under the restrictions of daily work at the EFSA. The results were incorporated into the draft version of the guidance document. The case studies have been completed and were discussed at a workshop in 2014.

Furthermore, draft versions of the guidance were reviewed by different internal and external audiences:

- The draft version was discussed by the Scientific Committee of EFSA to ensure a harmonised approach in EFSA and give strategic advice on the use of expert knowledge elicitation in EFSA.
- A public consultation was organised for summer 2013, to gather comments from stakeholders and other scientific communities, as well as from the public.
- At the beginning of 2014, EFSA held a workshop at which Panel members and scientists from EFSA discussed the guidance document and its applicability with a view to its future implementation.

- A final check by two external reviewers was conducted prior to finalisation of the guidance document.

The final guidance takes into account all the comments made during these four phases of review.

Objectives and intended users of the guidance

This document intends to provide guidance on the process of expert knowledge elicitation in the context of risk assessments within EFSA. The objective is restricted to the elicitation of quantitative parameters and rankings. It uses the elicitation of probabilities to express the parameters and the uncertainty of the process in a quantitative manner. A purely qualitative approach to elicitation is not part of this guidance.

The guidance presents particularly (Table 2)

in part I	a general introduction and motivation;
in part II	a step-by-step description of concrete processes.
in appendix A	a review and discussion of existing principles.

The document intends to guide the reader in the selection and adoption of different methods to conduct an expert knowledge elicitation on quantitative values for use in safety risk assessment for food and feed. Therefore, the applicability of the described methods is the main criterion in the selection of the content.

The guidance is principally intended for EFSA staff, scientific experts (Members of the Scientific Committee, Scientific Panels and their working groups) and external contractors[9], who are providing input to EFSA's work, and who want to apply expert knowledge elicitation, but who have relatively little knowledge of the process. Experienced elicitors may find additional information on how an elicitation process should be adapted to the specific conditions of risk assessments in the area of food and feed safety.

Furthermore, the guidance should help external experts who are contacted by EFSA for an elicitation process to understand EFSA's approach to expert knowledge elicitation. The guidance should explain the role of the external experts in knowledge elicitations[10], the needs and restrictions of the methods used by EFSA and the analysis of answers obtained in the elicitation process.

Finally, in line with the EFSA commitment to transparency, the guidance should contribute to informing stakeholders and the general public of the working procedures of EFSA in food and feed safety risk assessment.

[9] In the context of EFSA, an external contractor is an entity that is awarded with a procurement or grant.
[10] In this document the term "external expert" refers to external experts in the knowledge elicitation process. It is not used for members of EFSA's scientific committee, scientific Panels or their working groups.

Table 2: Structure of the guidance

Part I: Introduction	**Chapter 1:** The relevance for food and feed safety ***The context of this guidance***
	Chapter 2: Introduction to expert knowledge elicitation ***The motivation for this Guidance***
Part II: Exemplary protocols / implementation	**Chapter 3:** Initiation ***Preparations of the Working Group***
	Chapter 4: Pre-elicitation ***Preparations of the Steering Group***
	Chapter 5: Elicitation ***Preparations of the Elicitation Group***
	Chapter 6: Execution of the elicitation protocol ***The "Cookbook" for beginners***
	Chapter 7: Post-ElicitationI ***Finalising the elicitation***
	Chapter 8: Implementation in EFSA ***Recommendations for implementation***
Appendix A: Principles	**Chapter A.1:** Principles of problem definition ***How to frame the problem?***
	Chapter A.2: Principles of selecting external experts ***How to select the external experts?***
	Chapter A.3: Principles of eliciting a distribution ***How to elicit uncertainty?***
	Chapter A.4: Principles of handling multiple external experts ***How to combine several answers?***
	Chapter A.5: Documenting the results ***What to document?***
App. B	**Appendix B:** Web search on EKE guidance documents
Appendicies C-D: Case studies	**Appendix C:** Case study by the Animal Health and Welfare Unit
	Appendix D: Case study by the Plant Health Unit

EVALUATION

PART I: INTRODUCTION

1. The relevance of expert knowledge elicitation methods to food and feed safety risk assessment

1.1. Introduction to food and feed safety risk assessment

Risk analysis typically comprises three parts: risk assessment, risk management and risk communication (Figure 3). Risk assessment is a specialised field of applied science that involves reviewing scientific information in order to evaluate risks associated with certain hazards. The overall scope of risk assessment is to provide, as far as possible, a complete set of information to risk managers—so that the most systematic, comprehensive and accountable decision can be made concerning a potentially hazardous situation (Asante-Duah, 2002). Essentially, risk assessment provides information on the risk, and risk management is the action based on that information. The outcomes of risk assessment will typically facilitate decisions about the allocation of resources for safety improvements and hazard/risk reduction and will generally provide decision-makers with a more justifiable basis for determining risk acceptability, as well as aid in choosing between possible corrective measures developed for risk mitigation programmes. "*Risk communication means the interactive exchange of information and opinions throughout the risk analysis process as regards hazards and risks, risk-related factors and risk perceptions, among risk assessors, risk managers, consumers, feed and food businesses, the academic community and other interested parties, including the explanation of risk assessment findings and the basis of risk management decisions*" (EC 178/2002[11]).

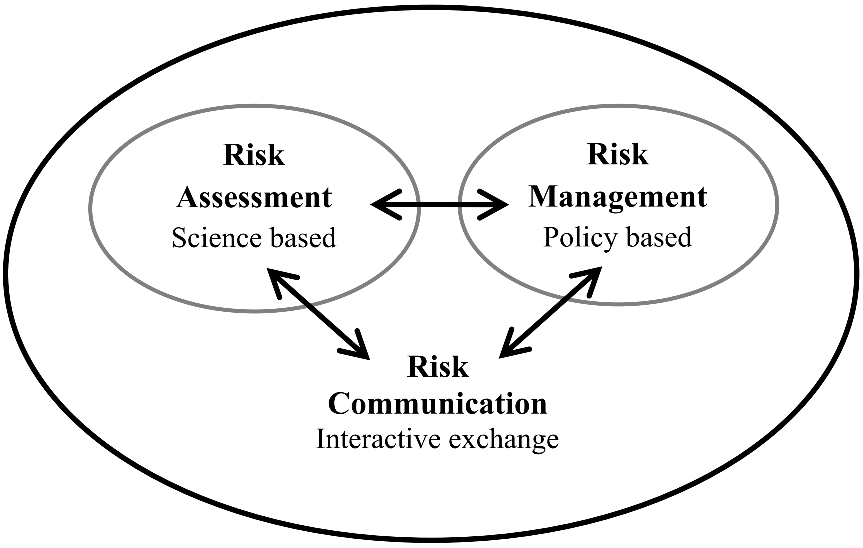


Figure 3: Risk analysis (CAC, 2011; EC 178/2002[12])

In the European Union (EU), risk assessment in food and feed safety is the responsibility of the European Food Safety Authority (EFSA[13]), while the European Commission (EC) is mainly in charge of risk management. EFSA and EC are each responsible for risk communication in their own areas.

[11] See footnote 5, page 13
[12] See footnote 5, page 13

Risk assessment in food and feed safety is primarily a method to estimate the likelihood of occurrence of adverse effects on human health, animal health and welfare, plant health and the environment resulting from exposure to hazards. Within the EU, a "hazard" has been defined as a biological, chemical or physical agent in, or condition of, food, and a "risk" is defined as a function of the probability of an adverse health effect and the severity of that effect, consequential to a hazard (EC 178/2002[14]). The International Program on Chemical Safety (IPCS) of the WHO has defined hazard as "*the inherent property of an agent or situation having the potential to cause adverse effects when an organism, system or (sub)population is exposed to that agent*" and risk as "*the probability of an adverse effect in an organism, system or (sub)population caused under specified circumstances by exposure to an agent*" (IPCS, 2009). In the case of *chemical risk assessment* (whether in the field of ecology, environmental protection or animal or human health) the hazard is a chemical compound, whereas in *microbial risk assessment* the hazard is a biological hazard such as a pathogen.

A model can be defined as a (simplified) representation of the essential components (scenarios, parameters, relations, processes or mechanisms) of a system, which incorporates existing knowledge and/or assumptions about the relationship between all system components in an explicit form (EFSA AHAW Panel, 2009). In general, food and feed safety risk assessment models follow an accepted methodology consisting of four fundamental pillars: (i) hazard identification (which is a prerequisite of any risk assessment); (ii) hazard characterisation; (iii) exposure assessment; and (iv) risk characterisation (Figure 4) (cp. NRC, 1983).

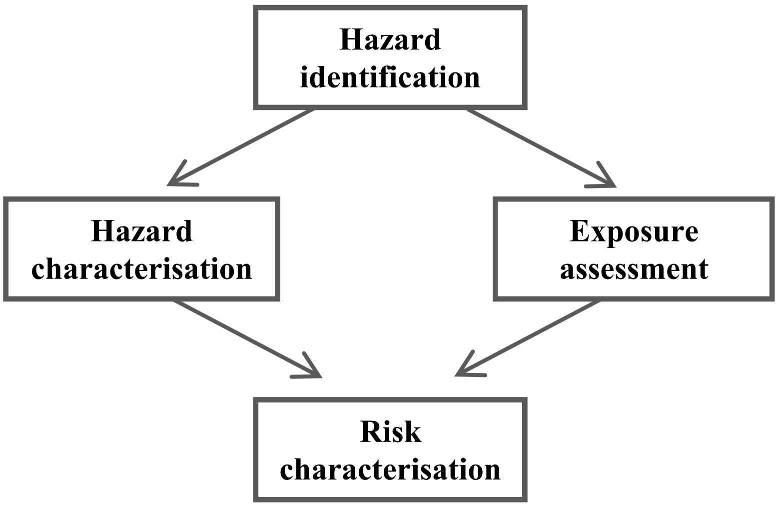


Figure 4: Steps in risk assessment (cp. NRC, 1983)

The concepts of risk assessment are similar whether they are applied to human and animal nutrition, animal health and welfare, plant health or the environment, but terminology and specific procedures may differ (Figure 5).

More formally, risk assessment is the process of calculating or estimating the risk to a given (sub)population, including the identification of attendant uncertainties, related to exposure to a particular factor (i.e. hazard) taking into account the inherent characteristics of the factor of concern as well as the characteristics of the specific target system (IPCS, 2004; EC 853/2004[15]).

---

[13] http://www.efsa.europa.eu/
[14]See footnot 5, page 13.
[15] Regulation (EC) No 853/2004 of the European Parliament and of the Council of 29 April 2004 laying down specific hygiene rules for on the hygiene for food of animal origin. OJ L 139/55, 30.4.2004, p. 1-151.

The method of *risk assessment for animal welfare* is based on a similar approach as chemical and microbial risk assessment. Problem formulation, including factor identification, is a prerequisite of the process and is equivalent to hazard identification, which considers whether particular factors (i.e. any aspect of the environment of the animals in relation to housing and management, animal genetic selection, transport and slaughter) have the potential to improve or impair directly or indirectly the animal welfare in the target population (EFSA AHAW Panel, 2012). The next stage after factor identification is formal animal welfare risk assessment, which comprises three steps: (i) exposure assessment; (ii) consequence characterisation; and (iii) risk characterisation.

Animal import risk assessment is used to evaluate the likelihood and relevance of adverse consequences upon entry, establishment or spread of a pathogenic agent in an importing country. The World Organisation for Animal Health (OIE) has defined risk as as the "likelihood of the occurrence and the likely magnitude of the biological and economic consequences of an adverse event or effect to animal or human health." (OIE, 2012).

Similarly, the method of *plant health risk assessment*, which follows International Standards for Phytosanitary Measures (IPPC, 2004), consists of the evaluation of the probability of entry, establishment and spread of hazards (i.e. harmful organism such as any species, strain or biotype of plant, animal or pathogenic agent injurious to plants, plant products or biodiversity in the EU) (EC 2000-29[16]) and the magnitude of the associated potential consequences (EFSA PLH Panel, 2010a,b).

Environmental risk assessment is a process of predicting whether there may be a risk of adverse effects on the environment caused by the presence of a pathogenic agent, harmful organism, pest or invasive plant (EFSA PLH Panel, 2011).

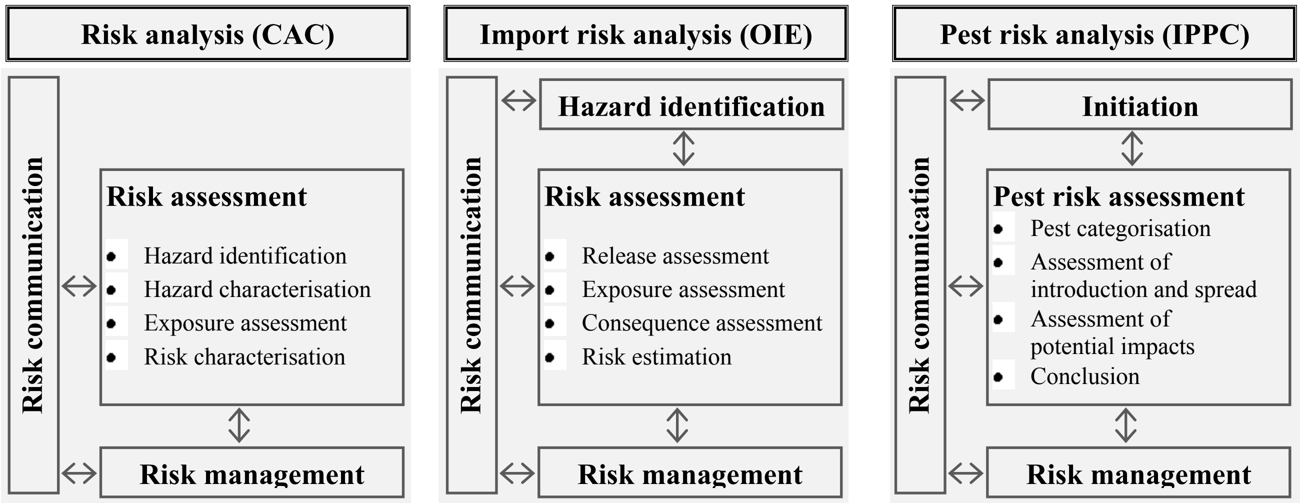


Figure 5: Comparison of risk assessment structures within risk analysis frameworks of CAC, OIE and IPPC (EFSA SC, 2012)

1.2. The evidence base for risk assessment

Whether it is chemical/microbial human health risk assessment, animal import risk assessment, animal welfare risk assessment, pest risk assessment or environmental risk assessment, the process breaks down the broad, overarching question into various sub-questions and answers to these sub-questions feed back into the overall assessment to provide the answer to the broad policy problem. For instance, a risk assessment for contaminants in the food chain might seek to answer the following questions: What is the mutagenicity of a chemical, i.e. its potential to induce cancer in humans and/or animals

[16] Council Directive 2000/29/EC of 8 May 2000 on protective measures against the introduction into the Community of organisms harmful to plants or plant products and against their spread within the Community. OJ L 169/1, p. 1-112.

(i.e. hazard identification)? What are the toxicokinetics of the compound in humans and/or animals and the toxicity (toxicodynamics) dose–response of the chemical in its target organ in humans and/or test species (hazard characterisation)? How widespread is the occurrence of the chemical, how much is likely to be consumed in food and what is the limit of detection of the analytical technique (exposure assessment)? In animal import risk assessment, the areas to address are the description of the pathogen under assessment (hazard identification), the definition of the possible pathogen introduction pathways (introduction assessment), exposure to the pathogen (including the assessment of the probability that a susceptible host in importing countries is exposed to the pathogen, once the pathogen has been introduced in the new country) (exposure assessment) and the likelihood of spread of the pathogen, once it has been introduced (consequence assessment).

A systematically review of scientific evidence is recommended in situations when quantitative models include parts with substantial uncertainties (uncertainty analysis) and/or ambiguities (scenario analysis). However, sometimes empirical evidence is limited (or even absent) or not accessible via publicly available information sources (e.g. bibliographic databases, scientific journals or websites). In such cases, information can be obtained from experts. In order to maximise its relaiability, systematic and standardised methods are recommended.The robustness of the evidence base for the risk assessment models is of paramount importance to ensure the reliability of the final outcomes for the assessed risk. It is highly dependent on the method applied for identifying and integrating such evidence in the risk assessment. Therefore, the method must be rigorous, systematic, standardised and transparently documented.

The types of scientific information on parameters available to answer the questions generated by any risk assessment model can be classified into three main categories, which represent the evidence base for risk assessment: (1) empirical evidence from primary research studies[17] available in the open literature; (2) raw data from suitable national and international databases (e.g. food consumption data; zoonoses, etc.); and (3) when empirical evidence or raw data are not available or not adequate, expert knowledge (Figure 6).

Accordingly, and in line with the 2012–2016 EFSA Science Strategy (EFSA, 2012), EFSA is defining best practices for gathering, selecting, appraising and (when applicable) synthesising the evidence base for risk assessment and integrating it into food and feed safety risk models. The best methods identified are (1) for empirical evidence: systematic review (EFSA guidance entitled “Application of systematic review methodology to food and feed safety assessments to support decision making”—EFSA, 2010a); (2) for raw data: well-defined data collection frameworks (EFSA Technical Report on “Data Collection: Future Directions” (EFSA, 2010b) and the *EFSA Data Collection Framework*); and (3) for expert knowledge, systematic and standardised methods for eliciting expert knowledge, which are described in the present EFSA guidance—Figure 6).

These methods should be seen from today’s perspective as the ideal ones able to provide the information needed for a model. For instance, in some cases a simple literature search may be the most appropriate method to gather the information needed for a model. Moreover, in many cases a risk assessment takes advantage of the application of different methods to provide information for the different components into which a model has been broken down. The application of the different methods for a given model should be decided on a case-by-case basis taking into consideration the pros and cons of each method. Expert knowledge elicitation is one possible source of evidence.

[17] A primary research study means an original study in which data were produced. The term is sometimes used to distinguish such studies from secondary research studies (e.g. reviews) that re-examine previously collected data (EFSA, 2010a).

	Type of evidence available to answer a sub-question		
Available evidence	Individual data	Aggregated data	Expert knowledge
Source of evidence	Primary research, databases	Sources of literature, databases	Knowledgeable external experts
Method for 1. Gathering 2. Validating	1. Perform experiment or survey / extract data 2. Validate /evaluate methodological quality	1. Extensive literature search, select studies, extract data 2. Critically appraise, summary data extracted	1. Frame problem, select experts, elicit judgements 2. Aggregate judgements
Method for 3. Analysing (incl. integrating)	Analyse and synthesise Data including integrating evidence		
Relevant EFSA project	EFSA Technical Report on "Data collection future directions" (EFSA, 2010b); EFSA Data Collection Framework	EFSA Guidance on "Application of systematic review methodology to food and feed safety assessments to support decision making" (EFSA, 2010a)	Present EFSA Guidance on "Expert knowledge elicitation in food and feed safety risk assessment"

Figure 6: Evidence base for risk assessment, methodological framework and relevant EFSA projects

2. Introduction to expert knowledge elicitation (EKE)

An expert knowledge elicitation (EKE) exercise requires careful planning. A number of important decisions have to be made about the experts, the format of the elicitation and the specific questions to be asked. These issues are addressed in detail in the following chapters, but we begin in this chapter with an overview to introduce the various topics. The primary purpose of the overview is to show how the detailed discussions in later chapters fit into the larger picture of a complete EKE exercise.

2.1. What is EKE?

The term "elicitation" has many meanings, representing different aspects of the general meaning of drawing out some information that is needed. EKE clearly refers to the drawing out of knowledge from one or more experts. However, even this is a broad term because it can refer to many different kinds of knowledge. In this section we discuss elicitation in a very general way and identify the specific sense in which it will be used in this document.

2.1.1. Knowledge

Elicitation is a broad field and in particular includes asking experts for specific information (facts, data, sources, requirements, etc.) or for expert judgements about things (preferences, utilities, probabilities, estimates, etc.). In the case of asking for specific pieces of information, it is clear that we are asking for expert knowledge in the sense of eliciting information that the expert *knows*. Elicitation of this kind of knowledge is relatively simple, comprising a direct question to the expert—the expert either does or does not know the answer. Two experts who both know the answer should give exactly the same values. Choosing the experts to use for the elicitation is a matter of finding people who know the answers to the questions we wish to ask.

When EFSA is eliciting judgements, there are many more challenges. EFSA is no longer simply asking for a piece of information, but wishes the expert to use his or her expertise to make an expert judgement, for instance an estimate of something. Now it is clear that different experts can give different answers. Some answers may be more accurate than others, and the choice of experts is much more complex.

Eliciting judgements is also knowledge elicitation. In order to answer questions, experts must draw on specific knowledge and also their expertise. The expertise which allows an expert to apply interpretation, analysis and synthesis to his or her specific knowledge to reach a judgement is a form of knowledge that we could call generic knowledge. A good expert will generally make more accurate judgements through having both good specific and good generic knowledge.

Elicitation techniques for different forms of judgement are studied in their own right, and in some cases have a substantial literature. For example, the elicitation of preferences is an important field, with in particular much interest in eliciting people's preferences between different states of health, in order to make evidence-based decisions about healthcare resource allocation (Torrance, 1986; Kharroubi et al., 2007).

2.1.2. Uncertainty elicitation

This document focuses primarily on the elicitation of expert judgements about uncertain events or quantities. The reason for this emphasis is explained in section 2.2.2 below. If there is no uncertainty about a quantity of interest, then its value is a fact that can be elicited as specific knowledge and no judgement is required. Where the value is not known by anyone, we seek to elicit the judgements of relevant experts because their judgements should be most useful, being based on more extensive knowledge and experience than the judgements of a non-expert. But it is important to recognise that the expert does not know the true value of the quantity. An expert will also be uncertain, although hopefully less uncertain than a non-expert.

In this context, a variety of judgements could be elicited. An obvious judgement is an estimate. However, simply eliciting an estimate barely scratches the surface of the expert's knowledge regarding the quantity of interest. In particular, it gives no indication of the expert's degree of uncertainty or confidence regarding this estimate. A simple estimate may be useful to a decision-maker such as a risk manager, but without any accompanying measure of uncertainty it has very limited value. In risk assessment, particularly, it is important to know how far from the estimate a quantity may realistically be. Estimates made by experts should be closer, on average, to the unknown true value than the estimates of non-experts, but we need to have an idea of how close; does the expert judge that the true value is most probably within ± 10 % of the estimate? or within ± 1 %?

There are strong theoretical and practical arguments to say that the proper representation of an expert's knowledge about an uncertain quantity is a probability distribution (e.g. Lindley, 1982; Cheeseman, 1985). A probability distribution quantifies the expert's knowledge and uncertainty in detail. From such a distribution we can compute the probability that the quantity will lie within any given range, or calculate precisely how much more or less probable some specific values are than others.

The remit for this Working Group and this report is the elicitation of expert knowledge about uncertain quantities *using probabilities*. However, as background to this remit, the reader may find it useful to note that some competing theories and representations of an expert's knowledge and uncertainty have been widely advocated.

- *Fuzzy membership functions* are claimed to be a less prescriptive and more realistic representation of an expert's knowledge (Zimmermann, 2001). Similar claims are made for *belief functions* (Shafer, 1976). Although superficially similar to probability distributions, these constructs follow different laws and so fuzzy membership or belief values behave differently from probabilities.
- Some psychological research shows that experts do not make choices that are consistent with having a probability distribution (such that decisions are made by maximising expected utility). It is claimed that alternative theories, such as *prospect theory*, better describe how people actually behave (Kahneman and Tversky, 1979).

Others argue that it is unrealistic for an expert to have or express a precise probability for the quantity lying in a given range and have developed theories of *imprecise probabilities.* An expert's judgement regarding whether the quantity lies in the given range would be described not by a single probability but by upper and lower probabilities, with the interpretation that the probability is only specified to be between these bounds (Walley, 1991).

The working group argues that the remit of this Guidance is right not to consider fuzzy membership functions, belief functions and similar alternatives to probability for two reasons. First, they do not have operational definitions, so it is not possible to provide a clear interpretation of what is meant by, for example, a fuzzy membership value of 0.4. In contrast, the meaning of a probability of 0.4 is clear and operationally defined through a reference standard. The probability of drawing a red ball from a bag containing two red and three white balls, when the balls are otherwise identical, the bag is shaken well and we cannot see inside the bag when picking a ball, is 0.4. Hence, to say that the probability that a quantity lies in a given range is 0.4 is a judgement that this event is equally likely as drawing a red ball in the above reference experiment. The reference standard provides an operational definition of probability in the same way that a tape measure, or more formally the official standard metre according to the International Bureau of Weights and Measures (BIPM) in Paris, provides an operational definition of length.

The second reason for not considering alternatives such as fuzzy membership functions is precisely that their values do not behave like probabilities, and thereby produce paradoxical results (Lindley, 1987).

We do not admit theories such as prospect theory because they are descriptive rather than normative. There are strong axiomatic arguments (DeGroot, 1970) to the effect that if a person is to make coherent judgements and decisions in the presence of uncertainty then his or her knowledge about the uncertain quantities should be described by a probability distribution (such that their decisions are then made by maximising expected utility). The fact that experts may not make perfect judgements that match and reveal their underlying probabilities is a practical challenge for the elicitation of probability distributions, not a fundamental reason for abandoning probability in favour of a description of imperfect judgements.

Closer to the use of probability distributions is the notion of imprecise probabilities, and we recognise that it is not realistic for an expert to be able to make precise judgements of probabilities. However, it is equally unrealistic to put precise upper and lower bounds on probabilities. We address the imprecision of probability judgements in a less formal way in appendix A, chapter A.3.

Having very briefly mentioned the alternatives, this report follows the Working Group's remit by considering the elicitation of expert knowledge about uncertain quantities only in the form of elicited probability distributions. Whilst it may not always be practical to elicit full probability distributions, this should always be the ideal.

We do consider some other forms of elicitation in some parts of this document because there are many choices and decisions to be made when planning and conducting an elicitation: often these can themselves be considered expert judgements, although not judgements regarding the value of some uncertain quantity. So some other forms of elicitation will be discussed indirectly in the report. Nevertheless, the primary focus is on elicitation of knowledge about uncertain quantities in the form of probability distributions, and on the principles and the nature of good practice for this kind of elicitation. Unless otherwise specified, the terms "expert knowledge elicitation (EKE)" and "elicitation" will always refer to this form of elicitation.

2.1.3. Uncertainty, variability and subjective probability

The word "uncertainty" is also used in different ways in different disciplines and contexts. In particular, it is sometimes interpreted more narrowly than the way it will be used in this guidance. The different usages can generally be understood in terms of the distinction between aleatory and epistemic uncertainty. This distinction is discussed briefly here, with particular reference to the meaning of probability.

Many things in nature exhibit random variability. The size of individuals of a certain species is an example: there is random variation in the population. Even if the distribution of sizes in the population is known completely, we will be uncertain about the size of an individual of that species randomly selected. This is uncertainty due to variability. It is also called aleatory uncertainty. The term "aleatory" is from the Latin and means "relating to randomness".

In reality, we will generally not know the true distribution of sizes for any given species. Uncertainty regarding properties of this distribution, such as the mean size, is not aleatory. The distribution itself does not vary randomly. It is fixed but usually unknown. Uncertainty is now due to simple lack of knowledge and is called epistemic uncertainty, from the Greek meaning "relating to knowledge". Some people interpret uncertainty narrowly to mean only epistemic uncertainty, and refer to aleatory uncertainty as "variability".

In practice, experts participating in EKE are often asked about quantities that take fixed values but whose values are uncertain simply because of our lack of knowledge. The uncertainty regarding such quantities is purely epistemic. It is sometimes necessary to elicit expert knowledge about a quantity that exhibits both kinds of uncertainty. An example would be the size of an individual of the species when we are also uncertain about the underlying population distribution of sizes.

These distinctions are sometimes useful when trying to understand the sources of uncertainty in an elicitation, but should not be taken too seriously because the boundaries are not defined precisely. Consider the prevalence of a particular disease of goats in a country that is divided into a number of districts. The prevalence is likely to vary between districts. Suppose that an expert knows the prevalence in the country as a whole (perhaps from having seen a published figure) but is asked for the prevalence in district A. Is the expert's uncertainty about the prevalence in district A aleatory or epistemic? Is it a matter of pure uncertainty or is there a component of natural variability? In a case such as this, it is useful (and indeed important) for the expert to be aware of these two ways in which the prevalence in district A is uncertain, but generally unhelpful to require the expert to understand them in terms of aleatory and epistemic uncertainty.

The different kinds of uncertainty are linked, however, to debates over the meaning of probability. The traditional definition of probability, which is still the version that almost every student first encounters, is that the probability of an event is the limiting relative frequency with which it occurs in an infinite sequence of instances. This definition obviously applies only to aleatory uncertainty, i.e. to uncertainty due to variability. It cannot apply to something which has a single fixed value because we cannot conceive of an infinite number of instances (or even of two!). For instance, the prevalence of disease in district A is a particular number and the event that it exceeds 0.01 is either true or false. The expert does not know whether the prevalence exceeds 0.01, and so is uncertain, but we cannot quantify that uncertainty with the traditional definition of probability because there is no sense in which the prevalence of the disease in district A can be repeated many times to see how often it exceeds 0.01. The limiting relative frequency definition is essentially useless in EKE because it cannot quantify epistemic uncertainty.

Instead probability is always interpreted in EKE using a definition known as *personal* or *subjective* probability. In this formulation, probability is a measure of a person's degree of belief in something. For instance, a probability of 0.4 that the quantity of interest will lie in a certain range expresses the expert's personal degree of belief in the proposition that it will lie in that range. Such probabilities clearly differ from one expert to another because experts all have different knowledge, both specific, factual knowledge and generic expertise. Although the subjective nature of such a probability definition is seen as a disadvantage in some quarters, it is entirely natural in EKE, where it is clear that the probabilities we elicit are personal judgements.

2.1.4. Practical elicitation

An elicitation exercise has a number of important components, all of which require careful thought and planning.

- *The objectives.* The purpose of the elicitation is generally to provide information in order to solve some problem, make a decision or to obtain greater understanding of some phenomenon. This external purpose dictates the quantities to be elicited. They are those quantities whose values are required in solving the problem, making the decision or understanding the phenomenon, but whose values are uncertain. The quantities of interest must be carefully defined to avoid ambiguity. (see appendix A, sections A.1.2 and A.1.3).
- *The format of the elicitation.* Elicitation can be conducted remotely via a questionnaire delivered by mail, by email or through a website, or by telephone, or in a meeting. (see part II, section 3.4).
- *The experts.* There will be one or more experts. The number may be influenced by factors such as availability, resources or the complexity of the quantities of interest. In EFSA's terminology, the experts whose knowledge is being elicited may be external experts or scientific expert members of the relevant Working Group (see appendix A, chapter A.2).
- *The elicitor.* Another necessary participant is known as the *elicitor*. In the case of elicitation by means of a meeting, the elicitor is an active participant, conducting the elicitation in a face-to-face dialogue with the experts and controlling the flow of the elicitation. The elicitor plays

a very similar role in a telephone elicitation, conducting one side of the discussion. However, when elicitation is conducted through a questionnaire, the elicitor's role is still one of controlling the flow of the elicitation but is limited to designing the questionnaire. The elicitor has expertise in the process of elicitation, and is an important component of the exercise (see appendix A, chapter A.4).

- *Other participants.* In a meeting there will often be others participating. For instance, there may be field experts with general knowledge of the field whose role could include providing background data, clarifying the objectives and quantities of interest for the experts, or helping the elicitor (who is often not an expert in the specific field) to communicate with the experts effectively. In EFSA's terminology, field experts will generally be members (experts or EFSA staff) of the relevant Working Group. Other participants may be keeping records or running software used in the elicitation process. In the background, although not usually a participant in a meeting, is the client for whose external purpose the elicitation is being conducted. In the EFSA context, the client may be considered to be the Working Group (see appendix A, chapter A.2).

- *Preparation and training.* To elicit carefully considered and informed judgements, it is usually important to give the experts preparatory material and training in the elicitation process. Training in the nature of probability judgements is a key part of this (see appendix A, chapter A.3).

- *The sequence of questions.* Elicitation of a probability distribution inevitably requires a number of judgements to be made. The precise wording of the questions and the order in which they are asked can contribute greatly to the quality of the resulting judgements (see appendix A, sections A.1.2 and A.1.3).

- *Aggregation and fitting.* Technical issues in eliciting probability distributions also require careful consideration. It is impractical to ask large numbers of questions about a single uncertain quantity, and in practice a probability distribution is obtained by fitting a suitable form of distribution to a relatively small number of expert judgements. Also, where there are two or more experts, the issue arises of how to aggregate their separate judgements into a single distribution (which is generally needed for the external purpose) (see appendix A, chapter A.4).

- *Documentation and reporting.* At the end of the elicitation, final elicited distributions for the quantities of interest must be reported to the client. It is good practice also to make a formal record of all the steps and judgements during the elicitation which culminated in these elicited distributions. (see appendix A, chapter A.3).

2.2. Why expert knowledge elicitation?

Expert knowledge about uncertain quantities, in the form of probability distributions, is widely used. One can find examples of expert elicitation in many fields.

- Expert elicitation has been used in the design of many large engineering projects. Examples include nuclear reactors and waste storage facilities (O'Hagan, 1998; Cooke and Goossens, 2000), dams (Brown and Aspinall, 2004) and defence systems.

- In the management of large infrastructure systems, such as water distribution networks and railways, expert elicitation has been used to supplement detailed engineering studies of small parts of the system (O'Hagan and Wells, 1993; Cooke and Jager, 1998).

- Elicitation is also used to quantify uncertainty in applications of statistical decision theory (Edwards et al., 2007).

- Uncertainty about parameters in large environmental models is also often addressed using expert elicitation (O'Hagan, 2012).

There are clearly good reasons for the use of EKE in these applications, but do these reasons also hold for EFSA? This section sets out the value of EKE to EFSA in the context of risk assessment.

2.2.1. Expert knowledge elicitation for risk assessment

Several of the examples in the preceding section involve risk. Risk assessments are typically complex, and it is usual to construct a model to describe and relate the various components of the problem, and to link them to available evidence. Such models generally contain many uncertain parameters[18].

In the EFSA context, consider, for example, an assessment of risk to human health from verotoxigenic *Escherichia coli* (VTEC). There are several pathways to consider from infection of cattle through to human consumption of dairy and meat products. Each pathway will be modelled as having several stages at which VTEC prevalence and concentration may be reduced or increased, by extents that are determined by parameters. All pathways depend on the prevalence of VTEC in cattle herds, which will be modelled in terms of parameters concerned with farm management, rates at which VTEC infection is passed through faeces (shedding), and so on. Even though VTEC has been heavily studied, most of these parameters are known only partially, particularly because there are many different strains of VTEC with different characteristics.

For the kinds of new and emerging diseases or hazards that EFSA is called upon to assess there is typically much greater uncertainty. Data relating to individual parameters in the model are often sparse or even non-existent. It may also be unclear how best to structure the problem with a model. In such a context, expert judgement is important and is recognised in the use of experts in EFSA Working Groups. Scientific external experts are frequently consulted to supplement the knowledge within the Working Group.

It should be noted here that the model itself is a judgement and could benefit from application of best practice in expert elicitation. Although this report focuses on elicitation of knowledge about uncertain quantities, the elicitation of a model and other judgements is discussed in appendix A, chapter A.1.

2.2.2. The benefits of quantifying uncertainty

In current EFSA practice, the elicitation of expert knowledge does not extend to eliciting probability distributions to quantify fully the experts' knowledge about the quantities of interest. Estimates of quantities required in a risk model may be elicited, but not necessarily in numerical form—qualitative estimates such as "high", "low" or "negligible" are also commonly elicited. The experts' uncertainty is described by a qualitative "degree of confidence/knowledge" or not at all. Such practices are simple to implement. They place modest demands on EFSA staff and experts. But in the context of this guidance they are inferior to a well-conducted elicitation of probability distributions and can lead to less accurate risk assessment (cp. EFSA SC, 2012).

> ***Example:*** *In order to discuss the deficiencies of current practice more fully, consider the following very simple example. In a certain risk model, one component is the proportion of mosquitoes (which act as a vector for the disease being studied in this risk model) that survive two applications of a spray. If X is the proportion surviving the first spray and Y the proportion of those that survive the second spray, then the proportion surviving both applications is XY (assuming time elapsed between the two applications is lower than needed for mosquito reproduction and assuming no development of resistance).*

First, qualitative estimates are particularly deficient. The formulae used for combining qualitative assessments in a risk model are arbitrary and can lead to misleading conclusions.

[18] Sometimes, the risk assessment may involve only some simple calculations, such that an explicit model may not be obvious, but a model is nevertheless implicit in the way that uncertain parameters are combined by those calculations to produce the final assessment.

> ***Example:*** *Suppose that X and Y are both estimated to be "small", meaning that few mosquitoes will survive a single application of spray. What now is the estimate of XY? A formula might say that XY should be given the qualitative description "very small", but what does that mean? And does it mean the same as "very small" when applied as a direct judgement to either X or Y? If we were able to give a numerical range to "small", such as saying that it means between 1 % and 15 %, whereas "very small" means less than 1 %, then the formula is misleading because when both X and Y are in the range 1 % to 15 % it is by no means necessary that XY is less than 1 %.*

Second, simply plugging estimates (whether qualitative or quantitative) into a risk assessment model gives only an estimate of the overall risk. Uncertainty in the parameters means that the true overall risk can be higher or lower than the estimate, and the magnitude of uncertainty in the final risk is important for decision-making. Even when qualitative assessments of uncertainty in parameters have been elicited, they cannot be accurately propagated through the model to determine the uncertainty in a final assessment.

> ***Example:*** *Suppose that X and Y are both given numerical estimates of 2 %. Putting these estimates into the model simply produces an estimate of 0.04 % for XY, but this is only an estimate and the true value could clearly be larger. It may be that in the risk model a value of XY less than, say, 0.05 % would not give cause for concern, but that a value larger than 0.05 % could lead to appreciable risk of undesirable consequences. Then the estimate of 0.04 %, without any acknowledgement of uncertainty, gives a false sense of security.*

Third, the estimate of overall risk obtained by plugging estimates into a model can actually be biased even when the individual parameter estimates are not.

> ***Example:*** *Suppose that the uncertainty about both X and Y is described by a uniform probability distribution over the range 0 % to 4 %. The estimates of 2 % are in the middle of this range and are unbiased. But consider the probability distribution of XY. X and Y are obviously related quantities, and it would be natural to assume that a single application of the spray kills the same proportion of mosquitoes each time, i.e. to assume that X = Y. Then* $XY = X^2$ *takes values from* $0.00^2 = 0$ *% to* $0.04^2 = 0.16$ *% with an expected value of 0.053 %. The "plug-in" estimate of 0.04 % is biased downwards.*

Finally, estimates obtained by simple questioning can be subject to many kinds of biases. Biases can be induced by the way the question is phrased or even by the demeanour of the questioner.

These deficiencies are minimised when we quantify expert knowledge and uncertainty fully by eliciting probability distributions, using EKE methods that conform to best practice in the field.

Once we have probability distributions for all the uncertain parameters in a model, that uncertainty can be propagated through the model to obtain a probability distribution for any relevant output of the model. If, for instance, the desired output is the number of new infections per year from a certain organism, then instead of a simple estimate we obtain a full probability distribution for the number, which quantifies uncertainty around the estimate. A simple example of propagation through the model is the Monte Carlo method[19]. The Monte Carlo method involves sampling many random values of all the parameters from their probability distributions and computing the risk output for each set of sampled values. The resulting large set of output values describes its probability distribution.

Using a well-structured EKE protocol also addresses the fourth deficiency above. Through best practice in EKE, as discussed in detail in this report, we obtain judgements from the experts that are as far as possible accurate, unbiased and reliable. Of course, we also need to acknowledge the limitations of "as far as possible". Expert judgements are inevitably subject to imprecision, as considered in

[19] Although the mechanics of propagating uncertainty through models is an important topic it is outside the scope of this guidance.

section 2.1.2. The resulting risk assessment is therefore not perfect but it is "as far as possible" an accurate basis for decision making given the prevailing state of knowledge, and should be considerably more reliable.

2.3. Challenges

The objective of elicitation is always to produce an outcome that is, as far as possible, reproducible, reliable and an accurate expression of the expert's knowledge and uncertainty. However, this is a challenging objective to achieve. In this section we review the challenges.

2.3.1. A diverse literature

In order to identify good practice for the purposes of risk assessment in EFSA, a literature search was employed. However, the fact that the applications of elicitation are so diverse means that the literature is also very diverse and spread over the journals and grey literature in numerous disciplines.

2.3.2. Psychological factors when eliciting probability judgements from an expert

It is tempting to think we can carry out EKE just by asking the experts to tell us some probabilities, or mean and variance, or even to tell us that their uncertainty about something is described by a normal distribution. Unfortunately, there is a substantial body of literature in the field of psychology which studies how people make judgements of uncertainty and probability, and demonstrates that naive approaches work badly.

Psychologists tell us that experts do not have probability distributions sitting in their brains ready to be drawn out by the elicitor's questions. The expert's judgements are constructed in response to questions, rather than being drawn ready-formed from the expert's brain (so elicitation is not a very appropriate term!). In constructing their judgements, people naturally employ short-cut heuristics, and this leads to biases. How we ask a question affects the answer, as does the sequence in which we ask things. The challenge is to conduct elicitation in such a way as to eliminate, or more realistically to minimise, these effects.

2.3.3. Some important heuristics and biases

A large part of the psychological literature on elicitation deals with *heuristics and biases* (Kahneman et al., 1982). The claim of the original researchers in this area was that our brains evolved to make quick judgements. Prehistoric man did not have time to think hard and carefully when faced with the everyday dangers of his existence, and so our brains developed short-cut mechanisms called heuristics. Evolution ensured that these heuristics worked well in the prehistoric world, but they can be a liability in the modern world. Our instinctive thought processes have not had enough evolutionary time to adapt to rapid technological and social change, so that the old heuristics may now mislead us and result in biases in our judgements. Some of the more important heuristics, in the sense of potentially biasing judgements in the elicitation of probability distributions, are considered briefly below.

2.3.3.1. Anchoring and adjustment

When asked to make two consecutive, related judgements, people tend to respond to the second question by adjusting the answer that they gave to the first. Experimental findings generally suggest that people do not adjust enough. The first answer is referred to as the anchor because it acts as a restraint on the response to the second question. This can happen even when the two questions are not obviously related—simply having the numerical answer to the first question in his or her head, the expert is influenced in a subsequent answer towards that first number. In general, answers to a sequence of questions may be influenced by the order in which the questions are posed.

Even numerical values given by the elicitor can influence the expert's subsequent responses in a similar way. For example, suppose that the quantity of interest is prevalence of a disease amongst animals and the elicitor asks the expert to give his or her probability of the prevalence being greater than 10 %. This already gives a cue to the expert that a value greater than 10 % is plausible. The expert might have been quite sure that the prevalence would be less than, say, 5 %, and there may be no reason to think that the elicitor knows anything about the topic, yet the expert is likely to respond with a non-negligible probability.

Choice of the sequence of questions in an elicitation procedure should aim to minimise anchoring, and should avoid introducing numerical values which might then serve as anchors.

2.3.3.2. Availability

In responding to a question, experts effectively search their memory for relevant information and processes the evidence to construct an answer. In doing so, information that is quickest and easiest to retrieve from memory is given priority and other information may not be retrieved at all. Anchoring and adjustment is connected with the availability heuristic in the sense that recent responses are highly available memories.

It is therefore important for experts to review all relevant evidence so that it is all equally available when they are responding to questions.

2.3.3.3. Range–frequency compromise

When experts are asked to provide probabilities for several possibilities, the number of options affects the probabilities that they give. Suppose, for instance, that we are interested primarily in the probabilities of options A and B and so ask the expert to provide probabilities for A, B and "anything else". Experiments show that we would, on average, obtain lower probabilities for A and B if we asked the expert instead to give probabilities for A, B, C and "anything else". This is related to one of the aspects of anchoring and adjustment, that by offering option C explicitly we make it appear more credible.

This is known as the range–frequency compromise because the expert's judgement of the probability (or frequency) of an option is compromised by the range of other options that are offered.

2.3.3.4. Other sources of bias

Several other heuristics have been discussed by psychologists as potentially biasing judgements of probability. According to the "representativeness" heuristic, judgements of the probability of an event are affected by how well its description fits some key piece of evidence, without reference to how likely the event would have been in the absence of that evidence (which is also known as "baseline neglect"). In particular, an event that is specified more narrowly can be judged more probable even though it must logically be less probable (the "conjunction fallacy").

Another example is "the law of small numbers", which refers to the fact that experts tend to give too much weight to small samples of data. This is one aspect of a general finding of "overconfidence"—when people are asked to specify an interval (or range of values) in which they are, say, 95 % confident (meaning that they have probability 0.95) that the unknown value of the quantity will lie, in general the true values are found to lie in such intervals far less than 95 % of the time. In general, the term "overconfidence" is used in elicitation to mean that the expert really has more uncertainty than is implied by the elicited probability distribution.

These and other kinds of errors and biases in judgements of probability are widely discussed in the psychology literature. However, there are also strong differences of opinion within the field (see Gigerenzer, 1991; Kynn, 2008).

2.3.3.5. Limitations of the psychology literature

Despite the wealth of psychological research, there are important limitations which mean that it is not always relevant to EKE in practice. Most of the experiments elicit judgements about uncertain quantities for which the subjects are not genuine experts. For instance, a large proportion of experiments have been conducted on psychology students. There is therefore rather little evidence of how real experts judge probabilities concerning questions in their domain of expertise.

Furthermore, most of the research has focused on individual probabilities of discrete events, rather than on eliciting probability distributions for uncertain quantities. In general, practical EKE typically involves asking experts to make more complex judgements than those which have been studied in any detail by psychologists. Elicitors must therefore have familiarity with the psychological literature on biases reviewed above to judge the extent to which any of the biases identified might apply in any practical elicitation task, and to decide how best to structure their questioning to minimise those biases.

2.3.4. Psychological factors when working with several experts

Additional psychological factors arise when elicitation is conducted with a group of experts.

An attractive feature of using a group of experts is that they can be invited to discuss and share their knowledge. However, it is important for the elicitor to ensure that all experts contribute to discussions and to the elicited judgements, according to their knowledge and expertise. Some experts are likely to be more naturally quiet and less likely to express opinions than others. Experts with stronger, more forceful, personalities will tend to dominate the group. The elicitor should be alert to these issues and seek to draw out contributions from all experts, whilst being aware that some may not be speaking on a given topic because they genuinely have nothing to say.

There is also evidence from the psychology literature on the performance of groups of experts when the objective is for them to reach consensus judgements. In particular, it has been found that probability judgements elicited from a group may exhibit overconfidence. It seems that the act of discussing and reaching consensus may, perhaps through a sense of common achievement, lead the group to feel more certain about the likely values of a quantity than is justified by the discussion and exchange of information. On the other hand, there may be irreconcilable differences of opinion between the experts that make them unwilling to consider any kind of consensus judgements.

2.3.5. Practical challenges

For various reasons we are in practice always limited in what we can do.

Resource limitations affect how many experts we can use and how we work with them. For instance, face-to-face meetings are more expensive than telephone interviews or unsupervised questionnaires.

Even if we had unlimited time and resources, we could not elicit unlimited amounts of detail because experts cannot make arbitrarily fine distinctions, become tired if the elicitation involves lengthy sequences of questions and are anchored by previous judgements. We also have to accept that the judgements of experts are inherently imprecise.

2.4. Choices

A complete elicitation exercise is a complex process, every step of which involves choices. This section briefly reviews the choices, with a view to alerting the reader to the issues which are addressed in later chapters.

2.4.1. What to elicit

EKE for risk assessment typically begins with identifying a risk model, a choice which should in principle be informed by the available evidence and available areas of expertise. The model will contain a number of *input* quantities which combine to produce one of more *output* quantities that together characterise the overall risk. Some of the input quantities may have known values, or values that can be found precisely from documents or experts (a process that can be viewed as another kind of elicitation). Others will be uncertain, and are therefore candidates for elicitation. To a large extent, therefore, the choice of model determines what to elicit.

The choice of model and how it determines the quantities to be elicited is discussed further in appendix A, chapter A.1.

2.4.2. Choice of experts

A risk assessment typically requires the elicitation of expert knowledge regarding several quantities. Different experts will have knowledge and expertise about different quantities, so for each quantity of interest we need to identify suitable experts.

But what is an expert and what constitutes relevant expertise? In simple terms, an expert is somebody who has knowledge that we consider to be worth eliciting; it is important to recognise that such a person may not be an "expert" in the everyday meaning of that word. Somebody who is an expert in this sense is a potential participant in the elicitation, but having identified experts there are some more choices to make. We need to select a subset of the identified experts to recruit for the elicitation exercise. Generally, we seek a group to cover the range of available knowledge and perspectives without unnecessary duplication. Then the chosen experts must be recruited and motivated to take part.

These choices are discussed more fully in appendix A, chapter A.2.

2.4.3. Managing the experts

The number of experts may not only be constrained by resources, but may also be influenced by how we choose to work with the experts—face to face (together or individually), over the telephone (individually or with a conference call), by internet (with a simple questionnaire or an interactive program) or by (e-)mail. The more intensive processes, under the active control of a elicitor, may be expected to produce more accurate elicitation of the experts' knowledge, but also require more skilled facilitation and more resources. These choices are considered in part II, section 4.4.

When multiple experts are to be used in eliciting a distribution for a given quantity of interest, a choice is required of whether, and if so how, to allow/encourage interaction between the experts. We can have free group discussion and group consensus judgements, free discussion but individual judgements, interaction only in restricted ways controlled by the elicitor, or no interaction. As mentioned in section 2.3.4, interaction between experts may bring benefits in the form of exchange of evidence and informed debate, but also introduces additional challenges in managing the experts.

The risk assessment context demands a single probability distribution representing best available knowledge about each uncertain quantity. Where experts will discuss and reach a single consensus distribution, then it should be made clear what that distribution represents. It does not have to be a consensus in the sense that all experts accept it as a representation of their individual beliefs. Indeed, we cannot expect that after a group discussion experts who begin with divergent views will converge on a common view. Rather, the final probability distribution should represent what it would be reasonable for a neutral but intelligent observer to think after assimilating the knowledge of the various experts. The elicitor may even find it useful to think of him- or herself as being that observer.

If experts are to make separate judgements leading to separate elicited probability distributions, then the question arises of how to combine them into a single final distribution. A number of algorithms for pooling several individual distributions into one have been proposed, notably the linear and multiplicative opinion pool formulae. It is also possible to choose between weighted and unweighted (or, equivalently, equally weighted) forms of these algorithms. In order to use a weighted form, however, a way must be found of constructing suitable weights. These issues are discussed further in appendix A, chapter A.4.

Another important aspect of the management of experts is the provision of briefing and training. Experts are expert in relation to the quantities of interest but will rarely be expert in making probability judgements. So training is essential to ensure that they understand what is required and know how to think about probabilities. Provision of briefing material helps to prepare the ground for training, but may also play important roles in ensuring that all relevant evidence is reviewed and ‘available’ and in obtaining the experts’ cooperation and commitment. These and other choices that are preliminary to the formal elicitation questions are considered in chapters 4 and 5.

2.4.4. Eliciting judgements

Consider a single uncertain quantity X. In order to elicit an expert’s knowledge about X in the form of a probability distribution, there are various kinds of judgements that we might ask the expert to make. Some may be qualitative, such as judgements about the general shape of the distribution; for instance a judgement that the distribution is unimodal simply says that there is a most probable value of X (the mode) and that possible values become less probable as we move away from the mode in either direction. However, since we are interested in quantifying uncertainty, we will always need to ask for some quantitative judgements. These generally are of the following types.

- Probabilities. For instance we may ask for a judgement of the probability that X exceeds 1.
- Quantiles. We may ask for a value x such that the probability that X is less than x is some stated value. For instance, the median m is the value of X for which the probability that X is less than m is 0.5 (and therefore the probability that it is greater than m is also 0.5).
- Probability intervals. Similarly, we may ask for a range of values of X such that there is a specified probability of X lying in this range.
- Moments or other more complex descriptors of a distribution. We might ask for the mean or standard deviation of the distribution.

Choices must also be made about the format in which the experts supply these judgements. Although quantitative judgements can be expressed as numbers, they can also be given implicitly as a mark on a suitable numerical scale (called a visual analogue scale) or using a physical device such as moving a pointer on a wheel or dragging a slider in an on-screen box. Qualitative judgements may be elicited as a free verbal response or constrained by some form of multiple-choice format.

In view of the psychological factors discussed in section 2.3, it is important to think carefully not only about how questions are phrased but also about the sequence in which they are posed.

These choices are considered in detail in appendix A, section A.3.5.

2.4.5. Fitting and aggregation

In this guidance we adopt the position that the goal of EKE is to produce a probability distribution for the uncertain quantity. However, elicitation produces a (usually rather small) number of individual judgements, and the question arises of how to convert these into a distribution.

The literature on elicitation exhibits a variety of approaches to this problem. One is simply to report the experts’ judgements without conversion to a distribution, but it is the view of the Working Group that EKE should result formally in a probability distribution. Some elicitation practitioners, having elicited some quantiles, assume uniform probability densities between the values of the uncertain quantity for which probabilities are given. Others view this as unsatisfactory because the resulting distribution’s probability density function has unrealistic discontinuities; instead they fit a smooth probability density function to the elicited judgements. The choice between these approaches is discussed in appendix A, section A.3.6.

It is important to recognise that a fitted distribution implies many judgements that the expert has not been asked to make. For instance, the expert might have expressed the judgement that the probability that *X* is less than 2 is 25 %, while the probability that it is less than 5 is 50 %. The fitted distribution will imply a probability for *X* being less than 3 (and indeed for every value between 2 and 5). It is good practice, therefore, to verify that the fitted distribution is a reasonable representation of the expert's views. The processes known as feedback and overfitting can be useful in this task and are also described in appendix A, section A.3.6.

A related challenge arises when judgements are elicited from multiple experts. We nevertheless require a single probability distribution, which means that the various experts' opinions must be aggregated in some way. This aggregation can be done by the experts themselves, through a process of interaction between experts that is designed to encourage them to reach a consensus view. Alternatively, it may be done by the elicitor applying an aggregation formula. The former method is referred to as *behavioural aggregation*, while the latter is known as *mathematical aggregation* or *pooling*. Choices in how to aggregate the judgements of multiple experts are discussed in appendix A, chapter A.4.

2.4.6. Documentation and reporting

The end result of the elicitation process is an elicited distribution (for each uncertain quantity of interest), and it is necessary to record and report this result.

Documentation is important. The client, for whom the elicitation is conducted, and to whose analyses and decisions it is intended to contribute, might in principle require only a statement of the distribution(s) obtained. Nevertheless, EFSA's commitment to transparency requires full documentation and reporting of the process, including at least a summary of the discussions, the experts' individual judgements (in anonymous format), and the fitting and aggregation stages. It is also good practice to solicit the experts' opinions of the elicitation exercise, as part of a process of ongoing improvement and quality assurance.

Elicitations conducted for EFSA risk analyses form just part of the risk assessment, and in particular they are inputs to the analysis rather than conclusions in their own right. The way in which expert elicitations are presented in EFSA's formal reporting is largely outside the remit of this guidance. For instance, it is important from the perspective of this guidance to provide detailed documentation of elicitation exercises for use within EFSA, but EFSA may choose to include only a brief summary of elicitation documentation in its own reporting of the risk assessments. EFSA also faces challenges of appropriately communicating risk and uncertainty in its reporting, but those are also outside the scope of this guidance.

However, one aspect of EFSA external reporting may impact on the elicitation exercise itself. Whether, and if so in what context, the opinions and judgements of individual experts are to be reported may influence the experts' willingness to participate in elicitations, and if they participate it may affect how openly and honestly they provide those opinions and judgements. Experts may be inhibited if they fear that what they say may later be used against them. On the other hand, they may not take sufficient care in their judgements if they feel there will be no consequence to them of making ill-considered statements.

These choices are discussed in appendix A, chapter A.5.

2.5. Meeting the challenges

2.5.1. Choices and challenges

There is a close relationship between choices and challenges, because good choices help us to address and overcome the challenges.

An important feature of elicitation is that in practice good elicitation requires the involvement of somebody with expertise in elicitation. This person will be referred to here as the *elicitor*, but elsewhere often as the facilitator. The elicitor is a different kind of expert. The experts whose knowledge is to be elicited have expertise regarding the uncertain quantities of interest, whereas the elicitor has expertise regarding the process of elicitation. It is the elicitor who will make the choices discussed above in order to conduct the elicitation effectively, and one purpose of this guidance is to help the reader to become a competent elicitor[20].

Psychological challenges are addressed through how the elicitor organises and manages the elicitation process, and particularly through how he or she designs the sequence of questions. Practical challenges affect the selection of experts and the way that the elicitor manages the process. This guidance document addresses those challenges in two ways.

First, chapters A.1 to A.5 form appendix A of this guidance, and provide detailed discussion of the principles and practice of the various kinds of choices. Specific guidance and recommendations are provided wherever good practice can be identified. The second way is through the development of protocols, below.

2.5.2. Protocols

Some of the choices will always be context specific, such as defining the quantities of interest and selecting particular experts. Others may be more generic, such as the choice of questioning sequence, or of whether to allow group interaction of experts. In such matters it is generally not possible to identify unique good or best choices, because there simply is not compelling evidence or agreement amongst leading practitioners of expert elicitation.

The situation is analogous to elicitation itself—nobody knows the ‘right’ answers. We therefore rely on expert judgements (in this case the judgements of experienced elicitors) and accept that experts’ opinions will differ. This guidance is based on the experience and judgements of the Working Group members and upon their knowledge of, and research into, the practices of other experts in the field of elicitation.

So although this guidance will identify bad choices in each of the areas presented in section 2.4, it will in many cases not recommend unique good or best options. In any given elicitation exercise to be conducted by EFSA it will still be necessary to choose from among the ranges of options that part II of this guidance recommends. We will refer to a set of choices that defines a complete method for conducting an elicitation as a protocol. Experienced elicitors in expert elicitation generally have their own preferred protocols which they use, with only minor variations, for each elicitation task.

We recognise that it would be unrealistic to expect somebody to make all the necessary choices, and so develop his or her own preferred protocol, simply on the basis of studying appendix A of this guidance. In part II, comprising chapters 3 to 7, we offer some fully specified protocols as exemplars of good practice in EKE. Finally, we present in Appendices C and D two case studies in which some of the ideas in this guidance have been carried through in support of specific EFSA mandates.

[20] It is important, however, to recognise that good elicitation skills are not acquired simply through reading a report such as this. They also need to be developed and refined through substantial practical experience.

Through applying the exemplar protocols to real-life EFSA risk assessments, we hope that relevant EFSA staff will acquire practical experience and develop their own expertise in elicitation. Variations on the exemplar protocols will no doubt be devised to suit the specific needs of EFSA, at which point it would be appropriate to revise this guidance document.

PART II: PROCEDURES AND METHODS

This part of the guidance describes in detail three different protocols which incorporate all the steps of an EKE exercise: the Sheffield protocol with group interaction of experts; the Cooke protocol with use of seed questions for the calibration of experts; and a Delphi protocol on written expert elicitation with feedback loops.

Each protocol is illustrated using a hypothetical example. The objective of this chapter is to inform the reader of the necessary steps in the protocols. All decisions on the process that are necessary to perform EKE for the hypothetical example were made according to the specific protocol. Each protocol represens one possible set of choices, which can be seen as the starting point of an EKE. Nevertheless, in practice, all decisions have to be justified with regard to the specific situation. Therefore, a concrete application will most likely differ from these protocols and will make use of some alternatives, as discussed in Appendix A. Appendix A gives detailed discussions of the main principles including recommendations which may allow the improvement of the three protocols for specific applications.

Appendices C and D present concrete case studies of expert elicitation performed within the current EFSA remit, in the Animal Health and Welfare unit, and the Plant Health unit. A WHO project including an expert knowledge elicitation is presented in Appendix E to cover also biological hazards.

The case studies may not cover the whole elicitation procedure, but discuss necessary adaptations and limitations of EKE in the daily work of EFSA.

Exemplary protocols

The following chapters describe three protocols applied to a hypothetical example of a risk assessment of a non-toxigenic foodborne bacterial pathogen, which was newly identified at a border inspection. One question is identified to be suitable for EKE.

The chapter describes the elicitation process step by step, defining milestones, sub-tasks, responsibilities and outputs. This part of the guidance can be considered a "cookbook" containing three different recipes that can be used to plan and conduct an elicitation process. Nevertheless, it provides the reader only with examples. In a specific application the process can be adapted using the alternatives discussed in detail in Appendix A.

On the other hand, the process for each protocol is described as fully as possible. In some cases existing free software solutions are identified to enable readers to carry out parts of their own EKE exercise, especially elicitations and calculations. The software solutions are specific to these protocols, but this does not mean that alternatives cannot be used instead. The objective is to make it feasible for readers to carry out an elicitation procedure and perform their own practical exercises.

As reported below, the conduct of an EKE exercise needs both education on the techniques and also experience in concrete applications, which cannot be given by this guidance. Therefore, the Working Group recommends further training for Steering Group members and EFSA staff in chapter 8.

The elicitation process can be structured into four main tasks (phases, Figure 7):

First phase: **initiation (see chapter 3)**

Second phase: **pre-elicitation (see chapter 4)**

Third phase: **elicitation (see chapters 5 and 6)**

Fourth phase: **post-elicitation (see chapter 7)**

The phases are conducted consecutively by different project teams with well-defined deliveries. This structure should allow EFSA to outsource parts of the elicitation process to specialised contractors.

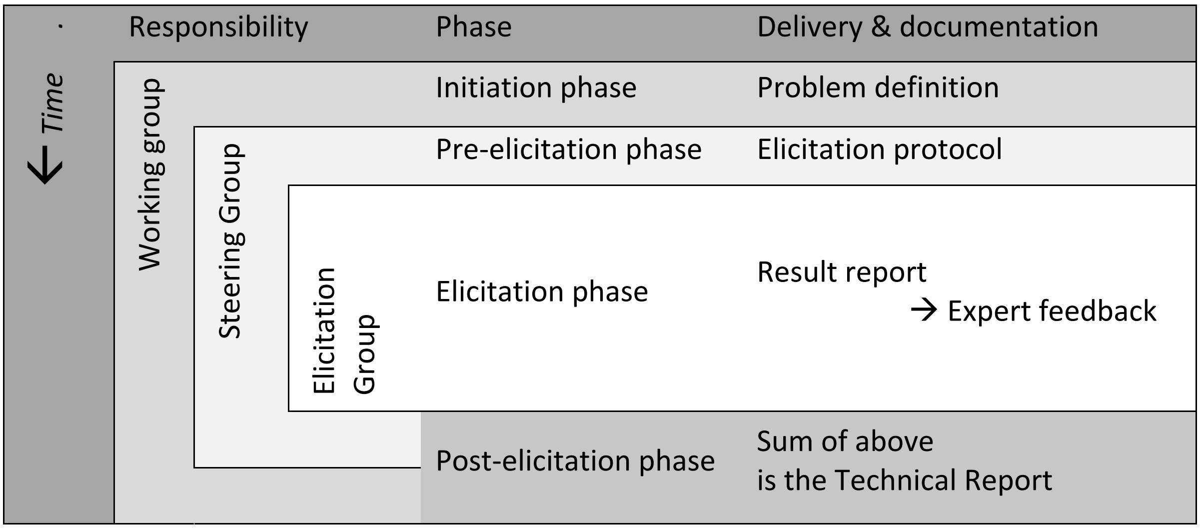


Figure 7: Phases of the elicitation process and related responsibilities

Additionally, the deliveries allow a review of each phase including a decision to go on to the next step.

The following chapters describe the phases and their sub-tasks in detail.

3. Initiation

The initiation of an EKE exercise is part of the regular risk assessment of EFSA decided by the responsible panel or unit. The actual performance is normally delegated to the Working Group which is tasked with fulfilling the mandate of EFSA under the responsibility of a panel.

The Working Group has to define the problem and justify the needs and conditions for an EKE, while EFSA has to provide the resources for the elicitation.

3.1. Problem definition

According to the objective of this guidance it should focus on the elicitation of quantitative questions. These questions typically arise when one or more parameters in a quantitative risk assessment cannot be obtained from regular publications (peer-reviewed or other scientific publications), public databases with a proper discussion of their uncertainties or competent authorities (see part I, section 1.2). This includes, for example, estimates of specific geographical or temporal conditions, industrial handling and processing techniques, processes actually used in the food chain or future risks.

Before starting the expert elicitation process the Working Group has to decide on the risk assessment model and the specification of all model parameters. Even at this stage it could be helpful to call on additional expertise, but this will normally be done within the Working Group by the recruitment of additional members or questions to hearing experts.

Please note that the hypothetical example is constructed for illustration of the process for all readers and does not reflect a real problem. Experts on non-toxigenic bacteria might find some aspects of the example unrealistic.

The hypothetical example: a non-toxigenic bacterial pathogen risk assessment

A Member State reported through the rapid alert system that a non-toxigenic bacterial pathogen was detected at a regular border control in a specific food item from a third country (outside the EU). Based on the conditions of production in the third country it is assumed that more consignments of this food may be contaminated, while food from other origins seems to be safe.

To determine the necessary detection level at border control a simple risk model is constructed. This model links the final contamination at the end user with the contamination at the border (point of entry) by a single parameter: the growth/survival/inactivation rate of this pathogen during transport from border to the end user. The transport conditions are too diverse and too divided into short passages for the application of a more stratified model to be feasible. Instead, experienced experts should form a judgement on the parameter taking into account their knowledge of transportation conditions and the pathogen characteristics.

3.2. Background information

The model will typically depend on various parameters whose true values are uncertain. The Working Group must review the existing information concerning these parameters which is accessible in books, peer-reviewed journals, other scientific publications or public databases. This will normally be achieved by a comprehensive search of the available information, for example by carrying out an extensive literature search using techniques described in the EFSA guidance on systematic review methodology (EFSA, 2010a).

The collection of background information is also helpful:

- to further specify the risk assessment model;
- to inform the experts of the known facts and existing information gaps;
- and to (re-)define the questions for the expert elicitation process.

Thus, the Working Group should prepare a short report summarising the existing background information on the risk assessment model and the parameters of interest.

The hypothetical example: report on background information

The working group summarised existing information in a background report:

- *The model and parameter definition:*

$$C_{\text{end user}} = C_{\text{entry}} \times R_{\text{transport}}$$

where

$C_{\text{end user}}$ is the contamination with the bacteria at end user (in colony-forming units (CFU)/g)

C_{entry} is the contamination with the bacteria at border (in CFU/g) and

$R_{\text{transport}}$ is the growing/survival/inactivation rate during transport and storage (dimensionless).

- All available information on model parameters or influencing factors:
  - growth, survival and inactivation parameters of the pathogen (under experimental conditions);
  - *characterisation of the food matrix, e.g. pH, water activity, constituents, etc.;*

  - *conditions of usual transport and storage, e.g. size of consignments, transportation means, storage condition (ambient temperature/chilled/frozen), shelf life, transport distances/times;*
  - *intracommunity trade regarding this food (no stratification by origin);*
  - *import of this food into each EU country from the specific origin;*
  - *levels of the pathogen detected at border control.*

- *The Working Group concludes that the conditions during transportation could be especially favourable for growth of the pathogen, but no specific information on conditions during transport that would enable the model to be refined or the global growth rate, $R_{transport}$, to be estimated is known. The Working Group therefore recommends the use of EKE to estimate this parameter.*

3.3. Identifying parameters for formal expert elicitation

3.3.1. Parameters for which elicitation is not necessary

Formal, rigorous EKE demands non-trivial resources in terms of time, work and money, and in general it is neither necessary nor feasible to apply the methodology to elicit distributions for every parameter in a risk model. Therefore, the next step is to identify which parameters should be the subject of a formal EKE process.

Following the background information review, formal elicitation will not be required for a parameter in the following cases.

- *Value known.* The information review may locate a precise true value[21] for the parameter (or one with negligible measurement error). This value can simply be inserted in the risk model and the parameter is removed from the list of uncertain model parameters.

- *Uncertainty well-quantified.* The review may locate an estimate for the parameter with known uncertainty. For instance, the estimate may be a measurement with known accuracy, or may be derived from a statistical analysis of data that provides a measurement of the estimate's standard error. A distribution may be assigned to this parameter with mean equal to the estimate and with standard deviation equal to the published standard error. The form of the distribution should be appropriate to the context of the parameter. Unless uncertainty is relatively large a normal distribution will be often a reasonable choice.

- *Uncertainty requires small inflation.* Frequently, a figure is found which may be considered as a case of either "value known" or "uncertainty known", but is a value for a different parameter. For example, the model may require the incubation period for a particular animal infection, and although there are no data for that infection the incubation period for a similar infection is known. Provided the quantity for which the value has been located is judged to be sufficiently similar to the parameter of interest, this case can be accommodated by a small increase in uncertainty. If we have a "value known" for the related quantity, then we can consider it to be an estimate of the parameter of interest with small uncertainty. The Working Group may then assign a suitable small standard error to this estimate based on their judgement of how large the difference might be between the two quantities. If the quantity of interest is X and the value of the related quantity is Y, then the Working Group should now consider how large the difference X – Y might be. They should assign a standard deviation s to this difference (for instance as suggested for "minimal assessment" below). If we have a "value known" for the related quantity Y, then a normal distribution can be assigned to X with mean equal to the (known) value of Y

[21] Parameters are considered in the Guidance as fixed values. When there is variability, parameters will typically be a mean or another expression of the variability.

and with standard deviation s. If we have an "uncertainty well-quantified" estimate for Y, with standard deviation t, then the distribution for X should be same as that for Y, but with the standard deviation increased from t to the square root of $(t^2 + s^2)$. In both cases, it is essential that s is small: the Working Group must consider that the uncertainty about X – Y is much smaller than the uncertainties about other parameters in the model.

However, these cases are relatively rare. In practice, it is very common for the background information to include some evidence relevant to the parameter but which does not fit any of the above three situations. There may be two or more estimates, but such that a formal statistical meta-analysis (which would reduce this to an uncertainty well quantified case) is not appropriate. Then expert judgement is needed to synthesise the information into a single probability distribution. It is very common for an estimate to be obtained of a quantity which is related to the parameter of interest but which is not similar enough to qualify for the "uncertainty requires small inflation" case. Or there may be data on the parameter (or on a closely similar quantity) whose quality is dountful, which may be subject to bias, or whose relationship to the quantity of interest is unclear. In all of these particular situations, and in general whenever the above three cases do not apply, the parameters will remain as candidates for formal EKE.

If sufficient resources are not available to conduct a full EKE for all of the remaining parameters, then it is necessary to prioritise the parameters in order to select those whose uncertainty will be elicited. Section 3.3.3 discusses how to prioritise parameters, but first it is convenient to consider how probability distributions will be assigned to the other parameters.

3.3.2. Minimal assessment

When a parameter is not to be addressed by full expert elicitation, it is still necessary to characterise its uncertainty by assigning a probability distribution. The solution then is for the Working Group to assign a distribution by a simplified process that we will refer to as minimal assessment.

The WG first considers the available evidence for this parameter and selects their own best estimate. This will be taken as the mean, m, of the probability distribution. The second step is to select a value s that is a reasonable margin of error for that estimate. The WG should judge that the true value of the parameter is likely to be found in the range from m – s to m + s (see 'Important considerations', below). Then s will be taken as the standard deviation of the probability distribution. The specification of the distribution is completed by choosing an appropriate form of distribution, such as a normal or Gamma distribution.

This is of course a crude procedure. The selection of a distribution is made purely by the WG without the help of outside experts, and *ignores the generally accepted principles of good practice in elicitation that are set out in this guidance* and encapsulated in the three specific protocols in chapters 6.1 to 6.3. However, provided that the more important parameters, i.e. those that most influence the final risk assessment outcomes, have been assigned carefully elicited distributions, then it should not be important to devote as much attention to the remaining parameters. The minimal assessment approach is intended to get the most important features of the uncertainty about these parameters more or less right, and this should be adequate in practice.

Important considerations

Whenever there is insufficient resource available to apply full EKE to all the relevant parameters, it is necessary to apply a more simplified approach to those deemed less important, and minimal assessment meets this need. In fact, as discussed in section 3.3.3 below, a preliminary minimal assessment of all the parameters is used to facilitate selection of the important parameters for full EKE. Therefore, minimal assessment will be an intrinsic part of most risk assessments and must be done with care and properly documented.

In assigning a mean value (m) for a parameter, the WG should review available evidence and discuss their judgements to reach a carefully considered consensus value. Similarly, the margin of error (s) must be chosen carefully to represent the degree of uncertainty that WG members have about that parameter. The description of "margin of error" in the preceding section is deliberately imprecise. The interpretation of s as the standard deviation in a normal distribution provides a more precise definition, i.e. the WG should believe that the parameter is twice as likely to lie in the range between m–s and m+s as it is to lie outside that range. However, psychological biases may tend to lead to understatement of uncertainty if this definition is applied formally. We recommend that the WG specify s such that they are *at least* twice as certain that the parameter will lie in that range rather than outside it. The WG should document these judgements, and the sensitivity analysis (see below) fully in their Technical Report.

When full EKE methods have been applied to elicit experts' probability distributions for the most important parameters in the risk model, the fact that we retain the original, crude minimal assessment distributions for the other parameters is acceptable. Only a small part of the uncertainties in the resulting risk output measures will have resulted from the parameters subjected only to minimal assessment.

It cannot be emphasised too strongly that it should *never* be acceptable to use minimal assessment in the final risk analysis for all the parameters in the model, because this would not constitute any meaningful characterisation of uncertainty. If full EKE is not done for at least *some* of the most important parameters, then the mandated risk assessment should be qualified by a statement that the conclusions may not be robust to the (unquantified) uncertainty in model parameters.

3.3.3. Sensitivity analysis

The outstanding question is how to identify the more important parameters, in the sense that these parameters influence the risk assessment outcomes most.

It is tempting to think that this can be done informally, by the Working Group simply choosing the parameters that they feel will be most important. However, this approach will generally lead to poor choices because the importance of a parameter depends on two distinct factors. The first is how strongly it features in the model, in the sense of the derivative of the risk model output with respect to that parameter. This derivative, or gradient, is called the *local sensitivity index* of that parameter. The second factor is the amount of uncertainty regarding the parameter. A parameter may have a high local sensitivity index but may nevertheless be low priority for formal elicitation if the uncertainty regarding it is small. Conversely, a parameter with low local sensitivity may be a high priority for EKE if the uncertainty regarding it is sufficiently large. In practice, attempting to judge these two factors informally is difficult, and so informally assigning importance is likely to give poor choices of priotity.

The Working Group should instead apply a formal sensitivity analysis. The first step in this process is to carry out a preliminary minimal assessment for *every* parameter. This provides at least a rough indication of the level of uncertainty regarding each parameter. The simplest form of sensitivity analysis, known as *one-way sensitivity analysis,* uses the m and s values from the minimal assessment for each parameter (but not the assigned probability distributions). It proceeds as follows for each parameter.

- Compute the output of the risk model with this parameter set to m – s but with all the other parameters set to their m values.
- Compute the output of the risk model with this parameter set to m + s but with all the other parameters set to their m values.

- The measure of importance for this parameter is the difference between these two output values (ignoring the sign if negative).

Having computed the importance value for each parameter, they are ranked in order of importance and the most important will have probability distributions elicited by full EKE (with the available resources and/or time determining how many parameters are selected). The remainder already have their minimal assessment distributions assigned.

One-way sensitivity analysis is the simplest acceptable method for this assignment and in relatively simple risk models with relatively few parameters it can be expected to give a good selection of parameters for EKE. In more complex models it is preferable to use a more sophisticated approach called *probabilistic sensitivity analysis*. Methods of sensitivity analysis generally are reviewed in Frey and Patil (2002), Oakley and O'Hagan (2004). Probabilistic sensitivity analysis (also known as variance-based sensitivity analysis) is developed in detail in Saltelli et al (2000).

3.4. Conditions and resources for the elicitation

The choice of the elicitation process is also influenced by other conditions of the risk assessment, e.g. the timeline, available resources and geographical coverage.

The hypothetical example: conditions and resources for the elicitation

The Working Group concluded:

- *The mandate and work plan allows a project duration of four months. A deadline of four months was set.*
- *The procedure should be handled in-house.*
- *The three main importing countries (e.g. AA, BB, CC) of this food item from the specific origin should be involved.*

3.5. Proposal for the Steering Group

In principle, the expert elicitation process can be handled by a separate Steering Group independently of the full risk assessment. This will have the advantage that the Working Group can later independently evaluate the results and decide on their use. For practical reasons it is recommended that there is regular communication between the Working Group and the Elicitation Group. It is the task of Steering Group to form this link. The members of the Steering Group, domain experts, experts on elicitation and administrative staff, plan and steer the elicitation process by designing the elicitation protocol. The size of the Steering Group has to reflect this task, but should focus only on the topic of the elicitation. Normally the Steering Group will be smaller than the Working Group. When the elicitation protocol is drafted and the Elicitation Group is formed, the leader (elicitor) of the Elicitation Group will also become part of the Steering Group responsible for finalising the protocol. Additional hearing experts might also be co-opted to the Steering Group to elaborate specific sub-tasks or to evaluate the feasibility of the protocol.

The hypothetical example: proposal for the Steering Group

The Steering Group should consist of:

- *One expert on the pathogen, because specific knowledge is needed about the pathogen;*
- *One EFSA staff member educated in EKE;*
- *Two staff from EFSA administration to organise contacts and invitations;*
- *One elicitor once appointed (if additional to the Steering Group).*

3.6. Deliverables

The Working Group concludes this milestone with:

- a background report including the risk assessment model and information already obtained;
- existing information (e.g. references) on the parameter(s) of interest;
- the justification and necessary conditions for the EKE;
- preliminary timeline of the project;
- proposal for the membership of the Steering Group;
- estimation of necessary resources (staff, costs) of EFSA;
- a description of the tasks and necessary resources for external contracts.

3.7. Evaluation

The project is evaluated by the chair of the panel and the head of the responsible unit to justify an EKE and the dedication of necessary resources. A positive evaluation of the initiation phase is a necessary milestone to enter the next phase.

4. Pre-elicitation

To decide on the appropriate elicitation protocol, it is necessary to establish a Steering Group. If all necessary competencies and capacities are available in the Working Group, then the Steering Group can be a sub-group of the Working Group. Normally substantial scientists will be engaged from the Working Group, while elicitation experts and further administrative support are added to The Steering Group for this specific task. For some sub-tasks, the consultation of hearing experts might be necessary, and for large elicitation exercises a full external project may also be considered.

The Steering Group specifies the concrete question (suitable for expert elicitation), defines the expert profiles and selects the experts for EKE, the elicitation method and the Elicitation Group (elicitor(s)). The elicitation protocol includes all steps of the actual elicitation and lists all resources needed for the execution.

4.1. Specification of the elicitation question

Even when the parameter is defined in the risk assessment model, the elicitation question has to be framed in such a manner that the expert is able to think about it. Regional or temporal conditions have to be specified. The wording has to be adapted to the expert's language. The quantity should be asked for in a way that it is in principle observable and, preferably, familiar to the expert (or some familiarisation needs to be provided in the EKE training). In other words, the question has to be brought to the right level and detail. The metrics, scales and units in which the parameter is usually measured have to be defined.

The hypothetical example: elicitation question (working draft)

The Steering Group concluded that practical experience on handling this specific food item is more valuable than theoretical knowledge on the pathogen (which can be provided as background information by the Working Group). This implies that the experts will be more able to think in "levels of contamination" than in "growth/survival/inactivation rates". Accordingly, the question was framed as follows:

> *"Assume that a contaminated consignment of the (specific) food item is entering Europe. The contamination with the non-toxigenic bacteria at time of crossing the border is 100 CFU/g. The consignment will then be transported and handled in Europe under usual conditions. What level of contamination (in CFU/g) would be in the consignment when it has reached the end user?"*

The elicited probability distribution will be transformed into a distribution for the final growth/survival/inactivation rate for use in the risk assessment model by dividing by 100 CFU/g. There is an implicit assumption here that the rate is independent of the actual level of contamination at the border.

It will be important to make clear to the experts that their uncertainty about the level of contamination of a single consignment at the end user is made up of two components of uncertainty. One is random variation from one consignment to another due to variability in the length of journey, external temperature and transportation conditions during the journey. The other component is scientific uncertainty about how the contamination will grow or reduce under different conditions.

In more complex situations it may be necessary to refine the risk assessment model, e.g. split the parameter of interest into several (*in the hypothetical example:* $log(R_{transport}) = log(G)T$*, where G is the average growth/survival/inactivation rate per unit time (in 1/hours) and T is the total duration (in hours)*) or introduce further stratification (*in the hypothetical example, split into northern and southern Europe*). The task is to reflect restrictions in the possible knowledge of the experts in the question.

Deciding the elicitation question is a highly interactive process in the Steering Group, and one which needs input from the substantive scientist on the problem, from a person knowledgeable about elicitation to find possible question formats, and from administrative staff to decide on resources (e.g. timeline, possible number of experts, possible number of questions, etc.). It might be also helpful to discuss the question with an (known) expert. Therefore, this step might be revised several times. The final decision will be taken after the protocol has been selected, the Elicitation Group has been appointed and the experts have been recruited.

The hypothetical example: alternative questions

It should be noted that the Steering Group rejected other framings of the questions, which were discussed during the work.

Alternative 1: *"Assume that contaminated consignments of the (specific) food item are entering Europe. The contamination with the non-toxigenic bacteria at the time of crossing the border is in each case 100 CFU/g. The consignments will then be transported and handled in Europe by a variety of end users, so the length of journey will vary from one consignment to another and the external temperature and transportation conditions will also vary. What level of contamination (in CFU/g) when it has reached the end user will be exceeded by just 5 % of consignments?" This version of the question avoids asking the experts to assess a combination of two kinds of uncertainty, and combined with elicited distributions for other percentiles (changing the 5 %) would yield more useful information for the risk assessment. However, this is a more complex question for the experts to consider.*

Alternative 2: *"It is known that the non-toxigenic bacterium will grow significantly only when the temperature of the (specific) food item exceeds 28 °C. Assume that a contaminated consignment is entering Europe and will be transported and handled in Europe under usual conditions. For how many hours between crossing the border and arrival at the end user will the temperature of parts of the consignment exceed the threshold of 28°C?" The elicited distribution will then need to be combined with a distribution for the average growth rate per hour for use in the risk assessment model.*

Alternative 3: *"It is known that the non-toxigenic bacterium will grow only when the temperature of the (specific) food item exceeds 28 °C and needs about three hours to double the number. Assume that contaminated consignments are entering Europe and will be transported and handled in Europe under usual conditions. In what percentage of the consignments will the temperature of parts of the (specific) food item exceed the threshold of 28 °C for at least six hours (summed during the whole transport from border to the end user)?" The Steering Group discussed here if the knowledge about growth rates higher than a factor of 2 will be sufficient for the risk assessment and the elicitation should focus on the percentage of these unsecure consignments.*

4.2. Definition of the expertise profiles

The Steering Group needs to construct desired expertise profiles on the basis of a thorough task analysis. Different profiles may be required for the estimation of different parameters, while complex parameters might require inputs from several sources (including related literature and databases).

A profile describes a possible expert by his or her required procedural/declarative, theoretical/practical and substantive/normative knowledge. A profile matrix might be helpful to show that all aspects are covered by the whole expert panel. The profile should be detailed enough to enable a concrete search, but not too narrow to exclude too many candidates. It will be useful to distinguish between 'desirable' and 'essential' characteristics in the profiles. For example, since there is a need for experts to express uncertainty probabilistically, normative (statistical) expertise is desirable; however, it is not essential because training can be given.

Once profiles have been determined, then indicators of satisfaction of the criteria should be devised and used to identify (or select between) candidate experts.

The hypothetical example: expertise profiles

The Working Group asked to involve experts from the three main importing countries in Europe (e.g. AA, BB, CC). Practical experience can be found at entry, in food control at the border and at the end user site, in quality control in the processing industries. Information on transport and storage is found in logistic and trading companies. Substantive knowledge is found in academic areas of microbiology of non-toxigenic bacteria and food science. Knowledge of compliance with regulations can be found in the food control authorities of the Member States and their regional offices.

It was concluded that, because of the novel nature of the problem, practical experience will be more connected to the (specific) food item than to the pathogen. The food processing industry will cover only parts of the total import, because direct sale to the consumer is another relevant pathway for this food item that needs to be covered.

We need essential expertise in the following areas:

- ***substantive knowledge of transport/handling conditions*** *of this (specific) food item in the main importing countries, e.g. the conditions that promote or inhibit multiplication of the pathogen (temperature, time, presence of air, etc.);*

- ***substantive knowledge of differences in production, trade and processing** of the (specific) food item, e.g. that promote or inhibit multiplication of the pathogen (nutrients, preservatives, etc.);*

- ***substantive knowledge of conditions of the specific food item at the border, during transport/handling and at the end user site**, especially with regard to biological contaminations, e.g. how fast biological contamination multiplies under different conditions (temperature, humidity, exposure to air, etc.).*

We also identify desirable expertise:

- ***normative knowledge on** expressing of biological contaminations in standard metrics or convertible to standard metrics;*

- *ability to quantify risk probabilistically.*

In all cases there is a need to be able to give reasons for estimates (i.e. declarative rather than procedural knowledge).

Thus, the following roles were identified:

- *importer/trader with notable volume of import/trade of this (specific) food item, including distribution to consumer;*

- *academic food scientist with knowledge of the production and processing conditions of this (specific) food item;*

- *food inspector with regular experience in control of this (specific) food item, preferably at a main point of entry;*

- *quality inspector for a main processing company of this (specific) food item/food inspection at end user level.*

Further criteria that were identified were as follows:

- *The importer/trader/processor should be able to cover the situation in at least one of the three main importing countries.*

- *Scientists and inspectors should be able to express the biological contamination in standard metrics (or in units that can be converted to a standard metric).*

- *Scientists and inspectors should have experience of comparable organisms, foodstuffs or transport conditions.*

These profiles constitute a profile matrix (Table 3).

Table 3: Profile matrix in the hypothetical example

Knowledge requirements			*Country*	*Roles*			
Expertise	***Importance***	***Specificity***		***Industry***		***Government (inspector)***	***Academia (scientist)***
				Import/ trade	***Production***		
Transport/ handling conditions	*Essential*	*Specific*	*AA*				
			BB				
			CC				
		Comparable	*AA*				
			BB				
			CC				
Production, trade and processing	*Essential*	*Specific*	*AA*				
			BB				
			CC				
		Comparable	*AA*				
			BB				
			CC				
Conditions at end user site	*Essential*	*Specific*	*AA*				
			BB				
			CC				
		Comparable	*AA*				
			BB				
			CC				
Standard contamination metrics	*Desirable*	*Specific*	*AA*				
			BB				
			CC				
		Comparable	*AA*				
			BB				
			CC				
Expressing risk and uncertainty as probability	*Desirable*	*Specific*	*AA*				
			BB				
			CC				
		Comparable	*AA*				
			BB				
			CC				

It is assumed that the microbiological facts of the pathogen can be provided as background information from the Steering Group. Further it is assumed that the quality control of processing industry will have measurements of the requested contamination, but this is not fully sufficient to describe the contamination on the end-user level (e.g. restaurants or private consumers).

In the event of there being several questions, the expertise profiles can be used to group the questions with similar profiles. The elicitation process may then be done with different experts for each group of questions.

The Steering Group needs to identify how many experts would be needed for all (separate) parts of the grid. The grid will be used for expert selection, and may also inform the selection of the method (different methods or implementations of a method, such as face-to-face vs. remote, different ideal maximum number of experts who can be involved).

The hypothetical example: indicators

The Steering Group decided to search for six experts: one importer/trader, one microbiologist, one food scientist, one food inspector at the border and two quality inspectors from two different processing sites. The inspectors should cover the three main importing countries.

It was decided that, in formulating the elicitation question, the importer/trader would have the status only of an advisor on the conditions of transport and handling, attending the discussions but not (necessarily) participating in the quantitative elicitation. This is justified by the assumption that the trader may have no experience of measurements of the pathogen.

The following indicators will be used to evaluate expertise:

Substantive

- **Practical:** CV (e.g. years of experience, type of experience), vocational qualifications, references, role, questionnaire.

- **Theoretical:** CV (e.g. publications, awards, conference presentations), academic qualifications, references, role, questionnaire.

- **Declarative:** formal training, publications, presentations, teaching experience.

Normative

- **Practical:** on-the-job experience at expressing judgements and uncertainty in the required metrics (assessed through questionnaire).

- **Theoretical:** formal training, e.g. in probability and/or risk assessment (assessed through questionnaire and CV).

The following criteria were identified to evaluate the necessary expertise:

- **Importer/trader:** At least five years' experience in importing/trading/distributing the specific food item in one of the main importing countries.

- **Academic food scientist:** Relevant academic education (e.g. agronomist, food engineer, nutritionist), at least five years' relevant experience and participation in several research projects or publications on production and processing conditions of the specific or comparable food items.

- **Academic microbiologist**: Relevant academic education (e.g. biologist, veterinarian), at least five years' relevant experience and involvement in several publications on comparable bacteria in comparable food matrices.

- **Food inspector**: At least five years' relevant experience in control of the specific or comparable food items, preferably at a main point of entry, including testing of biological contaminations.

- **Quality inspector:** At least five years' relevant experience in control of the specific or comparable food items for a main processing company or in food inspection at end user level, including testing of biological contaminations.

4.3. Selection of the expert panel

The expert selection is a stepwise process. The first step is to map all main actors or experts for the topic of interest. This step is usually undertaken by the Steering Group by reference to existing databases (e.g. EFSA's expert database, registers of members of scientific or economic associations, organisation charts of authorities/companies) or the authors of most relevant publications and using internet searches. The identification can be done on a personal (e.g. scientist) or institutional level (e.g. relevant units, institutes, companies or associations).

In the next step the "snowballing" principle can be used to create a longlist of possible candidates, i.e. to expand the list of possible experts. This is when the first approach to individual experts and institutions should be made. The main aim of this first approach is to collect information about individuals who may fit the profiles defined at the previous step. In other words, existing experts who are known to at least partially fulfil the profiles can be asked to suggest other experts who might also fit the criteria. Often it might be desirable to ask for suggestions of 'counter-experts'—experts who are known to have opposing views to the experts initially approached—at this point. To this end CVs can be requested from experts if not already held in the database. In addition, a questionnaire could be sent out to gather specific information.

The exercise continues until a pool of experts that is stable and large enough to enable selection of the expert panel has been created. It should be noted that the longlist should contain more names than needed for the elicitation. This is because some experts might subsequently be rejected once the elicitation group has been established and because not all candidates will agree to take part and others may drop out before the elicitation is complete.

> ***The hypothetical example: selection of the expert panel***
>
> *Two different approaches are used to map the main actors in the example: the institutional approach to map the possible pathways (e.g. importer, trader, processor during the lifetime of the (specific) food item) and the personal approach to map the scientific expertise.*
>
> ***Institutional approach to map the pathways***
>
> *EFSA's national contact points or other competent authorities in the main importing countries are used to identify the most important points of entry (e.g. border stations and transport means) with the local authorities responsible for food control. Local authorities are contacted to identify responsible staff according to the given profiles.*
>
> *Additional local authorities, national trade and producer associations are asked for information about main importers/distributors and processing industries of the imports of the (specific) food item, who might be willing to share their knowledge on the non-toxigenic bacteria in the (specific) food item. The companies are further contacted for "snowballing". This means to name other companies on expertise, which is not covered by them, e.g. other parts of the pathway, other regions, etc.*
>
> ***Letter to a company***
>
>> *Dear Sir or Madam,*
>>
>> *The European Food Safety Authority (EFSA) is currently preparing a scientific risk assessment of the non-toxigenic bacteria [specific pathogen] in [specific food item]. As one of the first tasks we want to map typical pathways (e.g. importer, trader, processor during the lifetime) of [specific food item] in your country. Therefore, we are collecting contact data on the main actors, e.g. companies which are importing, distributing or processing [specific food item] and which are willing to share their expertise with EFSA. If a relevant question is identified during the risk assessment EFSA will use the information to identify a panel of knowledgeable experts, who will be invited to discuss and judge the specific problem.*
>>
>> *Your company was named as a main importer/distributor/processor of [specific food item]. We would be very pleased if you would indicate your intention to help EFSA on a risk assessment of [the specific non-toxigenic bacteria] in [specific food item] by naming*

a competent contact person in your company. We are seeking for experts with the following general roles:

- *Importer/trader with notable experience in importing/trading/distributing of [specific food item], including distribution to the consumer.*

- *Quality inspector with notable experience in control of [specific food item] or comparable food items for a main processing company or in food inspection at end user level, including testing for biological contamination.*

Please use the attached questionnaire to specify the contact details and actual competence of your expert. When an answer to a concrete question is required your expert will be invited with a separate letter. In any case we will inform your expert on the outcome of EFSAs risk assessment. A public consultation of the results is planned within the next six months.

At this stage we are also interested in enlarging the network of expertise. Therefore, we would also be pleased if you can inform us of other companies that are importing/trading/distributing [specific food item], particularly in the countries [AA/BB/CC].

Individual approach for scientific expertise

For research scientists more information is available from public sources, e.g. publications, ongoing research projects (in internet) or involvement in advisory boards/scientific societies. The results from a literature search can be used to identify main authors on the specific topic. EFSAs expert database can be used as well as recommendations of the working group/panel, international or national scientific societies. Experts with a broader overview are recommended for the first step. These experts will be asked to help EFSA to identify further (more specific) scientists, different viewpoints in the scientific community and ongoing research projects. If the research topic is unfamiliar to the Steering Group the use of additional hearing experts might be recommended.

Letter to a scientist

Dear Sir or Madam,

The European Food Safety Authority (EFSA) is currently preparing a scientific risk assessment of the non-toxigenic bacteria [specific pathogen] in [specific food item]. As one of the first tasks, we want to map the current scientific research/scientists working on this topic. Therefore, we are collecting contact data for researchers who are willing to share their expertise with EFSA. If a relevant question is identified during the risk assessment, EFSA will use the information to identify a panel of knowledgeable experts, who will be invited to discuss and judge the specific problem.

We are seeking the following experts:

- *Academic food scientist with knowledge of production and processing conditions of [specific food item].*

- *Academic microbiologist with knowledge of comparable bacteria in comparable food matrices.*

You were identified by your recent publications on this topic and EFSA would be pleased if you could confirm your willingness to share your expertise with EFSA by providing us with your contact data and a short description of your expertise by completing the attached questionnaire. If a concrete question arises, you will be invited in a separate letter to participate in the discussion. In any case, we will inform you of the outcome of

EFSAs risk assessment. A public consultation of the results is planned within the next six months.

At this stage we are also interested in enlarging the network of expertise. Therefore, we would like to ask you also to name further experts on this topic in your country or at international level. Scientists who can address specific aspects that you do not have knowledge of but which are of possible value for the topic are particularly welcome. Please give further recommendations also in the case that you are unable to be involved in the EFSA project.

The hypothetical example: questionnaire

The purpose of the questionnaire (Table 4) is to gather information about the nature of individual expert's expertise and their fit to the profile. Some of the information may be gleaned from other sources e.g. CV, references, publications etc. but not all information for all experts will be available from these sources thus the questionnaire can fill in any gaps. The questionnaire also gathers information that is unlikely to be obtained from other sources such as the amount and quality of feedback regarding judgments, and normative aspects of expertise. An example generic questionnaire is given in Chapter 4, Table 4, here we give a version that is tailored to the example used in the current Chapter.

The expert panel will be selected with regard to the fulfilment of the indicators for expertise and answers in the questionnaire.

Information received from candidate experts should be matched against the desired profiles. Mismatches may be indicators of training needs. The Steering Group should also apply selection criteria at this point, not only in terms of the number of experts required but also, for example, in terms of the heterogeneity of opinions sought and the granularity of expertise.

Table 4: Expertise questionnaire for the hypothetical example

Expertise Questionnaire

This questionnaire is intended to find out about the nature of your job, and the type of judgements that you make while performing it. These answers will be used to prepare for the upcoming elicitation workshop on

<name of workshop here>

In particular, we are interested in whether or not your job requires you to make probabilistic judgements, and how you make such judgements. In addition, we are interested to find out what sort of aids you use in making judgements, whether you received any relevant training, and whether you receive feedback about the quality of your judgements.

Part A: General description of your job

1. What is the title of your job?

---

2. How would you describe your area of expertise?

---

3. How many years of experience would you say you have in your area of expertise?

---

4. Would you describe that experience to be:
   - Wholly practical and/or field-based
   - Mostly practical and/or field-based but some theoretical and/or lab-based
   - About equally practical/theoretical and field/lab-based
   - Mostly theoretical and/or lab-based
   - Wholly theoretical and/or lab-based

5. Would you describe your expertise as being mainly concerned with (please check one below):
   a. Characteristics of the pathogen, e.g. how fast it multiplies under different conditions (temperature, humidity, exposure to air, etc.)
   b. Characteristics of the food item, e.g. that promote or inhibit multiplication of the pathogen (nutrients, preservatives etc.)
   c. Characteristics of the transportation, e.g. that promote or inhibit multiplication of the pathogen (temperature, time, presence of air etc.)
   d. A combination of the above (please write down the corresponding letters a–c).

Part B: The judgements you make

6. Describe the most important judgements that you make on a regular basis in your job.

---

7. How often do you make judgements regarding the potential degree of contamination of foodstuffs? Please tick ONE box.
   - Always
   - Often
   - Sometimes
   - Rarely
   - Never

8. If applicable, in what metric(s) do you typically express contamination levels?

---

9. If applicable, what factors do you consider when assessing potential degree of contamination?

---

10. How often do you make judgements regarding the potential degree of contamination of foodstuffs by the [specific pathogen]? Please tick ONE box.
    - Always
    - Often
    - Sometimes
    - Rarely
    - Never

11. 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.)? Please tick ONE box.
    - I always use my judgement alone
    - I mostly use my own judgement
    - I use partly my own judgement, and partly other sources
    - I mostly use other sources
    - I always use other sources alone (not personal judgement)

12. If you do use other information sources, please describe them below.

---

13. Would you say that your knowledge about contamination of [specific foodstuff] and/or by [specific pathogen] is specific to local conditions or generally applicable?

---

14. If specific to local conditions could you please state which countries/geographical areas?

Part C: Data, models and feedback

15. In making your work judgements, do you receive any feedback about their accuracy? Please tick ONE box.
    - Always
    - Often
    - Sometimes
    - Rarely
    - Never

16. If you receive some feedback, what form does this take?

---

17. How soon after a judgement, on average, do you receive feedback? Please tick ONE box.
    - The same day
    - Within a week
    - Within a month
    - Within a year
    - Longer than a year
    - I do not receive feedback

18. How would you rate the ease of making good judgements in your work?

Please tick a box that best represents your opinion.

	1	2	3	4	5	6	7	
Very difficult								Very easy

19. Do you make use of a formal model for making your work judgements?

Please tick a box that best represents your opinion.

	1	2	3	4	5	6	7	
Never								Always

20. How would you rate the availability of data that you use for your work judgements?

Please tick a box that best represents your opinion.

	1	2	3	4	5	6	7	
Poor availability								Plentiful

21. How would you rate the quality of data that you use for your work judgements?

Please tick a box that best represents your opinion.

	1	2	3	4	5	6	7	
Very poor								Very good

22. Did you receive any training to make judgements? If so please describe below.

Part D: Judgements of risk and uncertainty

23. Do you ever make any of the following types of judgements at work (numerically, verbally, or by some other means)? Please tick and fill in as many as are relevant.
    - I estimate the likelihood/probability of…_________________________________
    - I estimate the chances of…____
    - I estimate confidence in…____

24. How often, on average, are you called upon to make judgements of risk or uncertainty? Please tick ONE box.
    - At least once a day
    - At least once a week
    - At least once a month
    - Less than once a month

25. When you make judgements of risk or uncertainty, what forms do they take? Please tick as many boxes as are relevant.
    - Numerical estimates (e.g. 0.5, 50 %, 1 in 2)
    - Verbal estimates (e.g. likely, infrequent)
    - Comparative (e.g. 'the risk is similar to another risk')

26. If you do make *numerical* judgements, what forms do these take? Please tick as many boxes as are relevant.
    - 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 seven-point scale of likelihood)
    - Other type of numerical judgement: please provide details below

---

27. Please give an example of the type of judgement of risk or uncertainty you typically make (if you do make such judgements).

---

28. Did you receive any training to make judgements of risk and uncertainty? If so, please describe below.

---

29. When you have to make judgements of risk and uncertainty do you rely on your judgement alone or do you also use other information sources (such as manuals of statistics, computer databases or programs, etc.)? Please tick ONE box.
    - I always use my judgement alone
    - I mostly use my own judgement
    - I use partly my own judgement, and partly other sources
    - I mostly use other sources
    - I always use other sources alone (not personal judgement)

30. If you do use other information sources, please describe them below.

Thank you for your time and effort.

4.4. Decision on the expert elicitation method

In principle every elicitation method which conforms to the detailed guidance in appendix A can be applied for the elicitation of expert knowledge about a quantitative parameter. The choice of the elicitation method will be done by the Steering Group, although in practice the Steering Group will first decide the broad approach, after which they will appoint the Elicitation Group. The details of the elicitation protocol will then be negotiated between the Steering Group and the elicitor as a member of the Steering Group. The broad approaches are exemplified by the three detailed protocols in chapter 6—the protocols of the Sheffield, Cooke's and Delphi methods.

In order to choose between these approaches, the Steering Group will find much detailed information in appendix A.4 of the guidance and in the presentations of the three protocols in chapter 6. However, some relevant considerations that are likely to influence that choice are outlined here.

4.4.1. Generic considerations

The reader may find it helpful to note some of the most important differences between the three protocols. Perhaps the most fundamental difference lies in the way that interaction between experts is permitted or handled.

- The Sheffield method employs behavioural aggregation, in which the experts meet face to face in an elicitation workshop and are allowed to interact and discuss under the management of the elicitor. There are potential problems in such interaction which may distort the final elicited distribution and lead to a poor result, but the advocates of behavioural aggregation argue that with good facilitation by the elicitor these risks are minimised and are outweighed by the potential advantages of the interaction. The principal claimed advantage is that the final elicited distribution will be better informed through the experts sharing and debating their information and judgements. Another claimed advantage is that behavioural aggregation avoids the need to pick a mathematical aggregation rule. Finally, within a face-to-face workshop it is easier to ensure that the experts understand clearly what is being asked of them.
- Cooke's method does not allow the experts to discuss their judgements; interaction is limited to initial training and briefing. Instead of behavioural aggregation, Cooke's method employs a form of mathematical aggregation. The potential problems with mathematical aggregation are that the choice of an aggregation rule is somewhat arbitrary, that every choice can be shown to have some undesirable implications and that it is not clear whose judgements the aggregated distribution represents. Nevertheless, the advantage of having an agregation rule makes the aggregation explicit, auditable and, in a sense, objective. The advocates of Cooke's method argue that their aggregation rule is founded on formal principles and the aggregated distribution represents a "rational consensus" distribution. The principal claimed advantages are that Cooke's method uniquely is able to be validated and the aggregation rule allows experts whose judgements are poorly made or relatively uninformative to be downweighted (although the efficacy of this weighting is dependent on the degree to which seed questions are comparable to the substantive elicitation questions). Advocates also believe that allowing interaction between experts may lead to poor aggregation and is not justified by the potential increase in information resulting from sharing judgements.
- The Delphi method lies between these two positions. Interaction between experts is allowed but is controlled. Judgements from each round are fed back to the experts in the subsequent round, but in an anonymised form. Although the interaction is very limited, advocates of the Delphi approach argue that it allows some benefits from the sharing of information without the risks of personal factors influencing judgements inappropriately. After all rounds of the Delphi method are completed the final distribution is obtained by a simple equal-weighting mathematical aggregation rule. Although this lacks the complexity of the Cooke rule,

advocates of the Delphi method typically doubt the value and efficacy of unequal weighting and prefer the more 'democratic' equal weighting.

The methods differ also in respect of the amount of accumulated experience and evidence regarding their performance in practice.

- Cooke's method has been in use for many years, essentially unchanged, and so there is substantial experience with its use, including some accumulating evidence that the Cooke aggregation rule is effective in producing well-validated aggregate distributions. In particular, when suitable seed questions can be obtained to allow unequal weighting then the aggregation appears to be better than equal weighting.
- Although behavioural aggregation as a generic method has also been used for many years, different practitioners have employed different detailed protocols. So these methods in general, and the Sheffield method in particular, do not have extensive practical experience as specific protocols. Partly because these methods do not employ seed questions, there is no published evidence regarding their performance.
- The Delphi protocol as described here is a new development that is yet to be employed in practice, except in a case study in EFSA (described in appendix D). However, Delphi methods have been widely used for elicitations that do not aim at quantifying uncertainty in the form of probability distributions, with numerous published results, and although the specific questions in the protocol to elicit probability distributions have not been used before in a Delphi context they are based on the foundation of the Sheffield method.

Another generic consideration is the extent to which the different methods are informed by research in psychology.

- The details of the Sheffield protocol are explicitly derived from understanding the findings of the psychology literature. This has informed the nature and sequence of the questions that are asked in the elicitation workshop, as well as the guidance to elicitors in conducting a workshop.
- Since the Delphi protocol presented in this chapter employs the same form and sequence of questions as the Sheffield protocol, it is also informed by the psychology literature. However, in a questionnaire format is not really possible to enforce the sequencing because experts are able to look ahead and to revise earlier answers after they have seen later questions. Furthermore, when using a questionnaire there is a greater risk that experts will misunderstand what is being asked for.
- Even though Cooke's method has been guided by different considerations (principles of rational consensus), potential heuristics and biases are avoided via different properties of the elicitation protocol. Motivational bias is avoided by using a proper scoring rule that should encourage the expert to state his or her true opinion; the calibration score should in principle reduce expert's overconfidence, anchoring is avoided by the format of the questions that are asked in the elicitation workshop.

4.4.2. Context-specific considerations

In addition to the above generic differences, there are further differences between the three approaches in regard to the conditions they need in practice. Therefore, the choice of approach may also depend on the context of the specific task, for instance on the characteristics of the intended expert panel. Depending on the heterogeneity in expertise, education, culture, languages and places of work some methods are more suitable than others.

The possible interaction between the experts is an important factor which may influence the selection of the method. Some limitations might be part of the setting:

- The dedicated resources do not allow EFSA to bring the experts together, or to arrange individual face-to-face meetings for elicitation. In this situation the elicitation has to be carried out via web conferences or email.
- The experts do not speak a common language (on an appropriate level), e.g. the practitioners from the production site may be familiar with only their national language.
- The expert panel is multidisciplinary to a degree, with no common understanding of the risk assessment problem or of the terminology, e.g. problems of trade (economic costs) are different from problems of quality control (optimal transportation conditions).
- There are major differences in the opinions of different groups owing to institutional relations or scientific positions.

The absence of a common language means that translation is necessary, which mostly prevents direct interactions of the experts in a common workshop or web meeting. The Delphi method and Cooke's method can be performed with written material or only individual contacts, whereas the Sheffield method might necessitate simultaneous translation during the elicitation workshop.

Having a multidisciplinary expert panel needs an organised session to generate a common view. For complex problems this needs a common workshop, where the experts can intensively exchange their viewpoints and concepts of the different disciplines. This workshop is part of the Sheffield method and possible only in a restricted (written) form in the Delphi method. Cooke‘s method will use the workshop to define a common background, but will elicit parameter judgements separately. In this case, seed variables must be provided for all disciplines. For each discipline at least ten seed variables are needed, preferably more[22]. The size of the expert panel should be restricted to the minimum needed to cover the defined expertise profiles. Nevertheless, for some elicitation questions several different viewpoints (e.g. national conditions) have to be included. The workshop required by the Sheffield method works best with about six to ten experts, whereas the Delphi and Cooke's methods are able to cope with even larger panels. But expert elicitation processes are not designed to solve hidden conflicts. In the case of known, but hidden, conflicts, which might produce strategic answers, the selection of experts should be revised.

When there are major differences in the expert judgements it is difficult to organise expert interactions in conflicting situations. The three protocols have different ways of identifying conflicts. The Sheffield method will recognise the conflicts during the common discussions and the failure to reach a final agreement of the expert panel, while in the Delphi method conflicts may be identified by the absence of convergence over multiple questionnaire rounds. For Cooke's method the differences have to be analysed by the answers to the seed questions or the specific distribution on the answers to the final question.

Other selection criteria can include those related to the resources needed to perform a method. There are differences in the resources (e.g. time, travel) required of the participants and the resources (e.g. staff, costs) required of EFSA.

Delphi is the least demanding approach for the experts, because of the written format and the adjustable timeline, whereas the Sheffield method and Cooke's method normally require attendance at a workshop or at individual face-to-face sessions. However, in Delphi elicitation there is no oversight on how much time and care the experts give their answers, and non response of some experts is common.

[22] Sometimes, it is advantageous to decompose the problem into several sub-problems corresponding to the different disciplines, elicit parameter judgements for each discipline separately (with the relevant experts), and then combine them in some appropriate way in order to answer the main question.

The EFSA resources required are broadly equivalent for each method, but differently weighted between staff (translations/Delphi) and costs (workshop/Sheffield method). Cooke's method requires the development of seed questions and is therefore more resource intensive than the others.

It is preferable to conduct a "dry run", if possible, with a small number of experts (or only one). An intensive planning of the elicitation protocol is better than repeating the whole exercise.

Once the elicitation method is selected, this will indicate the nature and perhaps number of experts needed.

The hypothetical example: decision on the method

The Steering Group concluded that three methods could be used to answer the question of interest: the Delphi method, the Sheffield method and Cooke's method. All methods will be described in chapter 6.

4.5. Proposal for the Elicitation Group (elicitor)

Depending on the method, different qualifications and capacities are necessary to conduct the elicitation process. The Steering Group will decide if the procedure can be carried out in-house or whether the additional expertise of an external contractor is needed. In the case of complex or conflictive questions, it is highly recommended that an outside elicitor with a neutral position be involved. In addition, if it is necessary to provide the experts with full confidentiality, the use of an external Elicitation Group with a corresponding contract is recommended.

The Steering Group will decide the composition of the Elicitation Group, which is responsible for conducting the elicitation process. In the case of an external contract, the Steering Group specifies the project and selection criteria.

Regardless of whether the event is held in-house or contracted out, significant administrative support will be necessary. This needs to be recognised up-front and should be dealt with by the Steering Group. Support may be needed to provide lists of appropriate experts to be approached, to contact such experts (by email, letter or phone), to provide resource materials and space for the conduct of the EKE (printing questionnaires, providing a meeting room and all necessary materials—beamers, computers, flip-charts), to arrange travel and accommodation for invited experts, and so on. Resource costs should be estimated up-front: How many and which type of EFSA staff may be necessary for how long? What are the likely budgetary implications (and from which source)?

After the nomination of the Elicitation Group, all conditions of the final elicitation are evaluated. At this stage the definition of the question for elicitation, the constitution of the expert panel and the selection of the elicitation method should be reviewed again. The Elicitation Group needs to agree that the elicitation protocol and timeline of EKE exercise are feasible.

The Steering Group and the Elicitation Group revise and agree on the elicitation protocol, that is the timeline and the schedule for the entire process, indicating at what stage certain aspects of the EKE are expected to be achieved (e.g. identification of appropriate experts; recruitment of experts; appointment of contractor; organisation of event; expected time for delivery of final report, etc.). The timeline will be provided for each protocol in chapter 6.

4.6. Deliverables

The Steering Group reports all its decisions in the elicitation protocol. This includes, for each expert panel:

- the elicitation question;
- description of the expert selection procedure, including the proposed membership of the expert panel and a number of substitutes;
- the proposed elicitation method;
- the revised timeline;
- the project plan with milestones to be achieved.

Furthermore, the Steering Group provides back-up of the Elicitation Group, being available for discussions, evaluations and further advice.

4.7. Evaluation

The elicitation protocol is evaluated by EFSA to confirm that allocated resources are sufficient to perform the EKE. An external review might be used at this stage to evaluate the use of EKE to answer the question of interest. The Working Group can evaluate the concrete question and proposal for the expert panel. A positive evaluation is a necessary criterion to enter the next step.

5. Elicitation

The actual elicitation is done by a separate Elicitation Group, typically comprising one or two elicitors with additional administrative support. These persons should be familiar and experienced with the selected elicitation protocol. All direct contacts with the experts are made by the Elicitation Group, and therefore the members should have a neutral position on the elicitation question. To enhance trust and guarantee confidentiality in ambiguous (conflictive) situations, the Elicitation Group should be independent of all parties involved.

The Elicitation Group will inform the experts on the topic, perform additional training on probabilistic judgements (if necessary) and execute the elicitation protocol.

5.1. Invitation of experts for elicitation

The Elicitation Group invites the experts to the elicitation process. Once the expert agrees to participate, all necessary background information will be provided to the expert.

- Check willingness, availability, any necessary permissions from employer etc..
- Inform experts about the problem and the reason for the elicitation and outline the conditions (e.g. workshop, attendance, workload, compensation, etc.).
- Provide detailed information on the problem including models, parameters, procedure details and the constitution of the expert group.
- Detail the methodology chosen, including information on all or some of the parameters/questions (e.g. if using Cooke's method, some questions may need to be held back for seed questions).
- At this point we can also ask the experts:
    - for additional information on the problem;
    - about any concerns they may have with the task or the procedure;
    - about conflicts of interests they may have related to the topic;

and gain their agreement with the procedure.

The invitation letter should contain the following:

- all necessary background information regarding the risk assessment, parameter(s) to be elicited, details of elicitation procedure and constitution of the expert group (although the last may be held back in certain cases, e.g. Delphi);
- a clear statement of what the expert will be expected to do and when and an estimate of the time involved;
- details of the method of reimbursement of expenses and any compensation;
- a request for confirmation of willingness and availability (and any necessary permissions).
- At this point we can also ask the experts:
    - for additional information on the problem;
    - about any concerns they may have with the task or the procedure and gain their agreement with the procedure.

The hypothetical example: invitation letter (assuming Sheffield method to be used) to an expert

Dear Professor Prugna

We have already been in communication with you regarding your possible help in conducting a non-toxigenic bacterial pathogen risk assessment. We wish now to formally invite you to take part in a knowledge elicitation exercise—further details of the nature of this exercise are given below.

The problem

A Member State reported through the rapid alert system that a non-toxigenic bacterial pathogen was detected at a regular border control in a specific food item from a third country (outside EU). Given the conditions of production in that third country it is assumed that more consignments of this food may be contaminated, while other origins seem to be safe.

The path to a solution

To determine the necessary detection level at border control a simple risk model is to be constructed. This model links the final contamination at the end user with the contamination at the border (point of entry) by a single parameter, the growth/survival/inactivation rate of this pathogen during transport from border to the end user. It is assumed that the transport conditions are so diverse and too divided into short passages for the application of a more stratified model to be feasible. Instead experienced experts will be asked to judge on the parameter taking their knowledge on transportation conditions and pathogen characteristics into account.

The parameter to be elicited

Just one parameter is to be elicited, which is the final concentration of pathogen (CFU/g) when the concentration on entry is 100 CFU/g.

The elicitation procedure

This elicitation will be conducted using the Sheffield Elicitation Framework (please see attached document).

The elicitation procedure will require attendance at a workshop of 1.5 days—this will allow sufficient time for the experts to discuss and assimilate the diverse information before making their judgements.

The workshop will consist of:

…

For further information:

The experts will receive a briefing document and an expertise questionnaire before the workshop. The briefing document will explain the purpose of the elicitation workshop and give a brief explanation of the tasks that experts would be asked to perform.

In the workshop, the elicitor will give a presentation in which the task of judging probabilities is explained in more detail. This will be followed by further training on the nature of the aggregated distribution that is the result of an elicitation. The experts will then be given an outline of the EFSA risk assessment mandate and the parameter that they will be asked to elicit.

The constitution of the expert group (*although this last may be held back in certain cases, e.g. Delphi)*

Experts (subject to their agreement)
Dr Schwarz
Professor Prugna
Mr Mosterd
Mrs Paon

Specialist advisers (subject to their agreement)
Ms Escarlata
Dr Groen

Elicitation group
A. Valkoinen (elicitor)
S. Grigio (recorder)

Dr Schwarz is a logistics scheduler with Agro-Transit, a company engaged in transporting food across Europe. He is familiar with the conditions under which food is typically transported, and the extent to which such conditions may allow the growth of pathogens.

Professor Prugna is in the public health school of the University of Transeuropia, specialising in the safety of food distribution.

Mr Mosterd is a food inspector who has experience of determining whether to admit food importation at several border entry points to the European Community.

Mrs Paon is an inspector for ReadyMealsRUs, a food processing company. She has considerable experience of measuring concentrations of bacteria in food from the company's suppliers.

Ms Escarlata (specialist adviser) is a food scientist who will provide data and expert knowledge regarding the nature of the food matrix for the specific pathogen.

Dr Groen (specialist adviser) is a microbiologist. She is familiar with the available data regarding growth/survival/inactivation rates of various bacteria under a range of experimental conditions.

The elicitor and recorder are experienced in the elicitation of expert knowledge.

Different kinds of experts might be differentiated at this point, perhaps on the basis of the screening. For instance, it may be considered that the knowledge of some experts is too specific for it to be worthwhile engaging them fully in the EKE; however, their expertise may be used to supplement the 'full' experts (this may have implications for the nature of an expert's participation, e.g. a 'partial' expert may not need to attend an elicitation workshop in person).

Once experts have agreed to participate, they need to be kept in the loop until the EKE is finalized and provided with any agreed incentives/remuneration. Mutually agreeable times and places of contact should be negotiated and necessary training should be provided. Feedback about the process should be given wherever possible.

5.2. Training on probabilistic judgements

Historically, formal expert elicitation has not regularly been used in risk assessment. Therefore, in any subject domain, one may encounter experts who have decades of subject matter experience, but have never attempted to quantify uncertainty using a probability distribution. One may also encounter experts who have had formal training in probability theory and statistics, but from a frequentist perspective in which probability distributions cannot be used to describe uncertainty about fixed quantities. Some experts (regardless of their understanding of subjective probability) may be sceptical of the use of expert judgement in the absence of hard data, and so will need persuading that the elicitation methodology is valid. Almost all experts will need training, the exception being if an expert has participated in a formal (probability distribution) elicitation session previously. The development of a stand-alone training tool is one of the recommendations in chapter 8, but it goes beyond the remit of this guidance. We suggest allowing one to two hours for training, and the recommended contents of the training are listed below. We assume that the experts already understand the broad objectives of the risk analysis.

1. Probability density functions.

   Introduce the idea of a probability density function as a device for representing subjective uncertainty. Reassure the experts that they will not have to provide density functions directly; we will construct them from simpler judgements. Choose a simple example that has nothing to do with randomness, such as the distance between two cities. Plot a plausible-looking density function, and highlight various features:

   a. values around the mode of the distribution are judged more likely than values in the tails;

   b. choose two sections of equal width, but with one section having twice the probability of the otherto contain the true value, explaining that the probability of the true value lying in a section can be visualised as the area under the curve;

   c. highlight the fact that regions well outside the tails are judged, in effect, to be impossible.

   It may be helpful to show some examples of *implausible* density functions, for example a uniform density over a very wide range, and density over a very small range that is obviously wrong. The point to get across here is that, regardless of our uncertainty about an unknown quantity, we can identify distributions that do *not* represent our beliefs, which can help to concentrate our thoughts on what distributions *do* represent our beliefs.

2. State what is needed: a probability distribution for the model inputs, to derive a probability distribution for the outputs.

   Explain that the aim of the elicitation session is to obtain a probability distribution for each uncertain risk model input. These will be used to derive a probability distribution for the risk model output. This is desirable because:

   a. The true value of the output is unknown; therefore, a single 'best estimate' of the model output will almost certainly be wrong.

   b. So EFSA should report not only a 'best estimate' but also how uncertain it is about the output, and to assess the uncertainty in the model output it is necessary to understand the uncertainties in the inputs.

   c. A probability distribution is an unambiguous way of representing uncertainty, whereas verbal descriptions mean different things to different people.

   Explain also that, once we have distributions for each input, we can investigate which inputs are the most important, which may inform strategies for reducing output uncertainty.

3. Reassure the experts that they will not be expected to claim certainty they do not have.

   The aim is to elicit distributions that represent the experts' knowledge faithfully. If the experts are genuinely very uncertain about the inputs, the elicited distributions will reflect their uncertainty through densities that spread probability across a wide range of values. However, elicited distributions can sometimes exhibit underconfidence by being wider than would be implied by their true uncertainty.

   The elicitor should emphasise that, although experts should endeavour to make their judgements as honestly and as accurately as possible, it is not feasible in practice to make precise judgements of probabilities. Reassure the experts that EFSA will recognise that the elicited judgements and fitted distribution are necessarily imprecise, and will factor that in during their risk assessment. Nevertheless, experience from practice shows that elicited distributions of this kind are indeed valuable, and that risk assessments are rarely critically sensitive to the imprecision in the expert elicitations.

4. Encourage experts to be honest!

   Experts may try to be 'helpful' by providing conservative judgements (e.g. stating a parameter to be larger than they really believe). Explain that, if appropriate, this sort of conservatism would be accommodated elsewhere in EFSA's risk assessment. If experts also give conservative judgements this may lead to EFSA's final assessment being overconservative, so it is important that the elicitation provides them with an honest assessment of the uncertainty.

5. Give the experts a practice elicitation exercise.

   Choose a quantity that is known to the elicitor, but unknown to the experts. If possible, the quantity should be chosen from a domain that is similar to the model input. In the example, a possibility would be the net growth of a different pathogen. (Using the terminology of Cooke's method, an ideal practice variable would be a "seed variable".)

   The experts will need guidance in the particular elicitation technique that the elicitor has chosen. Some suggestions when eliciting quantiles are as follows. For the purposes of this guidance, we denote the uncertain parameter by X.

a. Eliciting a median (0.5 quantile).

The median value for X has a fairly simple interpretation: X is equally likely to be greater than the median as less than the median. Suppose the expert suggests a median value of 2 for X. Consider the following two choices.

A: the expert gets a reward if $X < 2$, but receives no reward or penalty if $X > 2$.

B: the expert gets a reward if $X > 2$, but receives no reward or penalty if $X < 2$.

If the expert's median value really is 2, he or she should have no preference for option A or B. A preference for option A means that the expert's median needs to be smaller, and a preference for option B means that his or her median needs to be larger.

b. Eliciting a lower quartile (0.25 quantile).

Experts typically find this harder. Once the median has been elicited, the lower quartile can be interpreted as follows. Suppose the expert has settled on a median value of 2, and suggests a lower quartile value of 1.75. Consider the following two choices.

A: the expert gets a reward if $X < 1.75$, but receives no reward or penalty if $X > 1.75$.

B: the expert gets a reward if $1.75 < X < 2$, but receives no reward or penalty if $X < 1.75$ or $X > 2$.

If the expert's lower quartile value really is 1.75, he or she should have no preference for option A or B. A preference for option A means that the expert's lower quartile needs to be smaller, and a preference for option B means that his or her lower quartile needs to be larger.

A similar approach can be used for eliciting the upper quartile.

c. Eliciting the 0.05 quantile.

Here, one approach is to introduce a 'reference event' with a probability that is easy to specify. Suppose the expert has suggested a 0.05th quantile of 1.1. Consider the following two choices.

A: the expert gets a reward if $X < 1.1$, but receives no reward or penalty if $X > 1.1$.

B: a ball is chosen at random from a bag with 1 red ball and 19 blue balls. The expert gets a reward if the chosen ball is red, but receives no reward or penalty if the chosen ball is blue. (It may help the elicitor to have a bag of coloured balls to illustrate this.)

If the expert's 0.05th quantile really is 1.1, he or she should have no preference for option A or B. A preference for option A means that the 0.05th quantile needs to be smaller, and a preference for option B means that his/her 0.05th quantile needs to be larger. The reference event may not strictly be necessary if experts fully understand the notion of a 1 in 20 chance.

A similar approach can be used for eliciting other quantiles, by adjusting the proportion of red balls in the bag accordingly.

To conclude the practice exercise, the elicitor should reveal the true quantity of the practice variable, and comment on how 'close' the elicited distribution was to the true value.

6. Discussion of psychological biases.

   In combination with the practice exercise, the elicitor should discuss any psychological biases (discussed in Chapter 2) that may be relevant. For example, if eliciting the median first and then the quartiles, the elicitor should discuss anchoring effects, e.g. advising the experts to reflect on both the median and a lower plausible limit when choosing the lower quartile.

5.2.1. Delivery of training

It is the responsibility of the Elicitation Group to carry out the training. If the elicitation is commissioned from outside EFSA, the training should be an explicit part of the task specification. When the elicitation is to employ either the Sheffield or Cooke's method, then the training can naturally be given by the elicitor as part of the elicitation workshop. With the Delphi method, however, the delivery of training is more problematic. Simply giving written instructions will usually be inadequate because (a) experts will be unfamiliar and so require substantial training material, but (b) they will be unlikely to read such material carefully or to digest it fully when given in written form. For the Delphi method, a videoconference format is recommended for the training.

In practice, practitioners of the Cooke and Sheffield methods will often give some training in advance as well as additional face-to-face training in the workshop. In principle, it would be valuable for all methods to have access to a common, well-constructed and peer-reviewed training course. Ideally, this would be a self-tutored course so that it can be taken remotely by experts who are geographically separated (even if they will later come together for a workshop), and would include self-assessment exercises.

Unfortunately, no such course is known to the EKE Working Group or has been found in the literature searches.

The hypothetical example: training

In the hypothetical example the Elicitation Group invites the experts to attend a web-conference, during which lectures on uncertainty expressed in probability distributions and the elicitation method are given (duration about two hours).

6. Execution of the elicitation protocol

Please note that the discussion on the selection of the appropriate method is given in chapter 4.4.

6.1. Sheffield method

6.1.1. Overview of the Sheffield method

The Sheffield Elicitation Framework (SHELF) is a package of materials to assist with elicitation. It was created by Tony O'Hagan and Jeremy Oakley at the University of Sheffield, UK, and is available for free download from http://www.tonyohagan.co.uk/shelf/. The first fully defined protocol that we propose in this guidance for EFSA use, referred to here as the Sheffield method, is based on SHELF. Whereas the SHELF package includes several specific elicitation protocols, the Sheffield method employs a particular one known as the *quartile method*. Further details and alternative protocols can be found in the SHELF package.

The Sheffield method is distinguished from the other protocols proposed in this guidance by its use of behavioural aggregation. It is designed to be employed to elicit the knowledge of a group of experts in a face-to-face elicitation workshop, with the result being a distribution to represent the aggregated judgements of the experts[23]. The presence of an elicitor is essential, and it is helpful for the elicitor to be supported by a recorder.

An elicitation using the Sheffield method will be documented by the elicitor or the recorder completing two forms that are found in the SHELF package. The first of these is referred to as SHELF1 and records details about the elicitation workshop such as date, time, participants, training given, etc. The second is the SHELF2 document, specifically the quartile version of SHELF2. A SHELF2 form is completed for each elicitation conducted in the workshop. Since the method entails the use of a training elicitation to familiarise the experts with the quartile method there will in practice always be at least two SHELF2 forms, and more if the experts' knowledge is to be elicited for more than a single uncertain parameter.

The "SHELF 1 (Context)" and "SHELF2 (Distribution) Q" forms from version 2.0 of SHELF can be found later in the text, filled in for the hypothetical example.

6.1.2. Preparation for the workshop

The Steering Group will have identified the parameters to be elicited and determined that the Sheffield method is appropriate for the task of eliciting expert knowledge about these parameters. It will accordingly have appointed an Elicitation Group with expertise in applying the Sheffield method and in behavioural aggregation generally. The Steering Group will also have identified a number of relevant experts and assembled information from the initial investigations of the Working Group and from the experts themselves.

The elicitation group must now decide on how to organise the task in terms of how many elicitation workshops to run, which parameters to elicit in each workshop, the duration of each workshop and which experts to invite to each workshop.

The duration of a workshop should allow sufficient time to complete the elicitation tasks assigned for that workshop. This should be thought through in advance of sending invitations because it is generally impractical (in view of how busy experts typically are) to extend the time after the initial arrangements have been made, or to organise a second workshop when work was not completed in the first. The Sheffield method encourages the experts to discuss their judgements fully, and this can take a great deal of time (even when the elicitor is careful to curtail repetitious or peripheral discussions). To elicit knowledge about a single parameter, at least a full day will be required. As the experts become more familiar with the method, subsequent elicitations typically proceed more quickly, so that with a good group of experts it may be possible to elicit distributions for, say, four to six parameters in two days.

Two days is a sensible default workshop duration; beyond two, or at most three, days tiredness sets in and it is hard to maintain concentration. If more parameters are to be elicited than can be addressed in a single workshop, then two or more workshops will be required—in any case, in practice it is unusual for the same group of experts to be appropriate for such a large number of parameters.

> ***The hypothetical example: workshop duration***
>
> *Just one parameter is to be elicited, which is the final concentration of pathogen (CFU/g) when the concentration on entry is 100 CFU/g. A workshop of one and a half days is proposed, allowing sufficient time for the experts to discuss and assimilate the diverse information before making their judgements.*

---

[23] The method can also be used to elicit judgements from a single expert, and so could be used in a protocol in which the knowledge of several experts is elicited from each expert separately and then aggregated mathematically. However, the essence of the Sheffield method is to have a group elicitation in which the final elicited distribution is obtained by behavioural aggregation.

The number of experts participating in an elicitation workshop using the Sheffield method should ideally be no more than six to eight. With more experts than this, experience suggests that the time spent on discussion becomes excessive, without a corresponding increase in information. Furthermore, the elicitor will find it difficult to manage a large number of experts so as to ensure that all opinions are heard and discussed properly.

One way in which a larger number of experts can be accommodated is by designating some of them as having an advisory role. These are experts with specialist knowledge on one aspect of a parameter of interest but who do not have a broad view over other relevant aspects. Their advisory role is to provide the other experts with data, information or interpretations from their specialist knowledge, but they do not participate in making formal expert judgements about the parameter.

The hypothetical example: selection of experts

Based on the expertise matrix, the Steering Group identified experts in the categories of "food inspector, border", "food scientist", "microbiologist", "importer, trader" and "quality inspector, processing". The elicitation group decided to invite a food scientist and a microbiologist as advisory experts, with four other experts from the other categories invited as full elicitation experts. The reason for this was that the food inspector, importer, trader or quality inspector would have a reasonably broad view of the parameter of interest, but that it was important for them to have access during the workshop to specialist knowledge of the advisory experts.

Much of the preparation is discussed in the preceding sections of this chapter. In particular, the Sheffield method is no different from other methods in respect of the importance of defining the uncertain parameter(s) whose distribution(s) will be elicited and of identifying and recruiting experts. It is of course also necessary to recruit an elicitor with expertise in applying the Sheffield method or in behavioural aggregation generally, either by appointing an external elicitor or by training EFSA staff.

The relevant Working Group will have gathered some data at the outset to structure the problem, build a proposed risk model and identify parameters for elicitation. However, it is important to extend that exercise to ensure that as much relevant information as possible is available to the experts during the workshop. The recommended sequence of steps is as follows.

- The Working Group gathers initial information, makes an initial formulation of parameters for elicitation and appoints the Steering Group.
- The Steering Group refines the formulation of parameters for elicitation and initiates identification and recruitment of experts.
- The Steering Group solicits additional information widely, including from responsible persons in all relevant jurisdictions. This may be done by questionnaire provided that the requests are made sufficiently clearly, but a telephone call soon after sending the questionnaire is recommended to resolve any potential misunderstandings and ensure good data quality.

Having recruited experts (which will have entailed giving them at least a summary of the objectives of the elicitation workshop), a full briefing should be sent out. In addition to providing a full specification of the risk problem and the role of the elicitation, and also briefing them on the underlying concepts of elicitation (such as the nature of probability and probability judgements), it should give a digest of the information that has been obtained and its sources. Experts should be invited to tell the elicitation group of any other information that they believe is relevant.

All information gathered must be available and reviewed at the workshop, with experts' attention drawn to any items sourced since the briefing was sent to them.

The Elicitation Group should ensure that all necessary equipment is available in the workshop. This will usually include:

- *A computer, data projector and screen to complete and display the forms*. This allows the experts to see what record is being made and to inform the elicitor if there are any errors or omissions. This set-up can also be used for the elicitor to give a training presentation at the start of the workshop. A second projector/screen/PC may be useful to show fitted distributions.
- *A flip-chart or whiteboard.* It is often important for the elicitor or an expert to be able to sketch out some ideas.
- *Name cards.* It is recommended that each participating expert should be given a card to place in front of them in the workshop. In addition to the expert's own name, each card should show an anonymised name, such as "Expert A". The elicitor or recorder will use the anonymised names when recording experts' judgements, comments and opinions in the SHELF2 forms.

The elicitor may also ask for some materials to be copied and provided to the experts in the meeting.

6.1.3. Completing the SHELF1 form

The SHELF forms are designed to guide the elicitor through a well-structured protocol. The first part of the workshop should follow the sequence of steps in the SHELF1 form.

- Title, session, date and start time. The title is for the overall project and could be the title of the Working Group or the mandate. The session should identify the particular workshop.
- Attendance and roles. All people attending should be named and their roles identified. The roles of expert, elicitor and recorder are clear, but there can be others in attendance. An administrator from EFSA may be present to oversee the organisation. Other specialists may be invited in an advisory role, to provide information without participating in the expert judgements. Completing this part of the form is an opportunity at the start of the workshop to make introductions. It is important to build rapport and trust between all the participants early on, so that the workshop can run smoothly.
- Purpose of elicitation. The context of the EFSA risk assessment should be specified here. It is an opportunity to reinforce the message that the task that the experts are asked to perform is valuable and meaningful.
- This record. This field is pre-completed with the standard text, "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." There is an opportunity here to point out to the experts that the elicitation will follow a recognised standard protocol, and also to emphasise that a record will be kept of the process. The elicitor should point out that the experts can ask at any time for changes to what is recorded on the forms, if he or she feels there have been mistakes or some important points have not been recorded. The elicitor should also clarify that experts' judgements and comments will be recorded anonymously, for example as "Expert A".
- Orientation and training. This part of the form should be completed once the elicitor has conducted whatever training is felt to be needed. Training will often have been given before the elicitation workshop, in addition to the description of elicitation concepts and process that was given in the briefing document. However, it will generally be useful to provide/reiterate the training in the workshop. In all cases, the experts should engage in a practice elicitation. The uncertain quantity that is the subject of this elicitation can be chosen rather arbitrarily by the elicitor. Standard examples are so-called 'almanac questions' such as the population of a given country, the height of a mountain or the world record for some athletic event, but the experts may find the training more relevant if the quantity lies within their area of expertise and is similar to one of the uncertain quantities that are the subject of the elicitation exercise. The purpose is to take the experts through every step of an elicitation, and to discuss and clarify any difficulties of understanding that may persist even after the training. Time should

be taken to make sure that the experts fully appreciate what is being asked of them (but it is not necessary to have the kind of lengthy discussion of expert opinion that can be expected to take place when eliciting judgements about the parameters of interest). A SHELF2 form should be completed during the practice elicitation and the SHELF1 form resumed only when the training and practice elicitation are complete. The nature of the training and practice should be recorded on the SHELF1 form.

- Participants' expertise. The purpose of this step is to air what expertise each participant brings to the workshop. Often, different experts can contribute expertise in different aspects of the question. The step also allows the elicitor to understand when in the subsequent discussions an expert can be expected to be more or less authoritative, depending on the focus of the discussion at the time.
- Declaration of interests. It is important for all participants to mention openly any reason why they might benefit personally (financially or professionally) from the outcome of the elicitation. EFSA may need to conduct elicitations in which some experts are employed or paid as consultants by companies having a commercial interest in the outcome, and this should obviously be declared. Other experts may feel that they represent their country and might wish to see an outcome beneficial to their nation for reasons of loyalty. Even academics specialising in the topic of the elicitation stand to benefit professionally if the outcome supports the importance of that specialism. The elicitor should emphasise that declaring an interest does not bar a person from contributing, nor does it mean that there is any assumption that their contributions will be biased. On the contrary, everyone is assumed to be honest and contributing to the highest standards of professionalism. However, EFSA has a requirement for full disclosure of interests and openness tends to create an atmosphere of mutual trust that is beneficial for the elicitation process.
- Strengths and weaknesses. This step encourages the experts to reflect on the strengths and weaknesses of the group. Does their combined expertise cover the range of knowledge and judgement necessary to provide EFSA with a meaningful expression of aggregate knowledge? If there are areas of expertise not represented, or if there is a diversity of professional opinion and some positions are not represented, then this could be a weakness. It is important to record such potential weaknesses so that EFSA can take them into consideration when making use of the elicited distributions.
- Evidence. At this point, the elicitor or another designated participant should review the evidence that has been assembled. Psychological research shows that when making judgements experts tend to access information that they have encountered recently or that has made a strong impression, and to overlook other pieces of information. When all of the relevant information has been reviewed recently it is all equally accessible in their memories.
- Structuring. As far as possible, structuring should have been done prior to the workshop. That is, the Working Group should have identified the precise quantities to be elicited, on the basis that there was evidence and expertise available to support meaningful elicitation for those quantities, but that if these quantities were further decomposed ('structured' or 'elaborated') in terms of even more basic quantities there would not be the same availability of evidence or expertise at that level. Nevertheless, it is important to revisit those decisions with the expert group. They might strongly prefer to restructure the target quantities before embarking on the elicitation. This is another reason for making sure that there is adequate time available in the workshop.
- Definitions. Whatever parameters are to be the subject of the elicitation workshop, it is essential to define them precisely and to record those definitions (including units of measurement) on the SHELF1 form. It is necessary for all the experts to have the same understanding of what each quantity is, to avoid time-wasting misunderstandings in the elicitations.

- End time and attachments. The time at which these preliminaries are concluded should be recorded. Any supporting documents that are to form part of the documentation of this stage of the workshop should also be listed. These will usually include any training documents issued or presented during the workshop, any documents summarising the evidence and the SHELF2 form for the practice elicitation.

The hypothetical example: the SHELF1 form

The Sheffield Elicitation Framework SHELF v2.0

ELICITATION RECORD – Part 1 – Context

Elicitation title	Chapter 6: hypothetical example
Session	1 of 1
Date	14 November 2010
Part 1 start time	09:00

Attendance and roles	*Experts*: Dr Schwarz Professor Prugna Mr Mustár Mrs Pauw *Specialist advisers* Ms Scarlett Dr Verde *Elicitation group* L. Valkopesu (elicitor) S. Legris (recorder)
Purpose of elicitation	To quantify expert knowledge regarding the concentration of the specified pathogen in food on arrival at a processing facility.
This record	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.
Orientation and training	The experts received a briefing document and an expertise questionnaire before the workshop. The briefing document explained the purpose of the elicitation workshop and gave a brief explanation of the tasks that experts would be asked to perform. In the workshop, the elicitor gave a presentation in which the task of judging probabilities was explained in more detail. The experts then carried out a practice elicitation for the area of the Mediterranean Sea. This was followed by further training on the nature of the aggregated distribution that is the result of an elicitation. The experts were then given an outline of the EFSA risk assessment mandate and the parameter that they would be asked to elicit.
Participants' expertise	Dr Schwarz is a logistics scheduler with Agro-Transit, a company engaged in transporting food across Europe. He is familiar with the conditions under which food is typically transported, and the extent to which such conditions may allow the growth of pathogens. Professor Prugna is in the public health school of the University of Transeuropia, specialising in the safety of food distribution. Mr Mustár is a food inspector who has experience of determining whether to admit food importation at several border entry points to the European Community. Mrs Pauw is an inspector for ReadyMealsRUs, a food processing company. She has considerable experience of measuring concentrations of bacteria in food from the company's suppliers. Ms Scarlett (specialist adviser) is a food scientist who will provide data and expert knowledge regarding the nature of the food matrix for the specific pathogen. Dr Verde (specialist adviser) is a microbiologist. She is familiar with the available data regarding growth/survival/ inactivation rates of various bacteria under a range of experimental conditions. The elicitor and recorder are experienced in the elicitation of expert knowledge using the Sheffield method.

Elicitation Record – Part 1 – Context p1

The Sheffield Elicitation Framework SHELF v2.0

Declarations of interests	Dr Schwarz and Mrs Pauw recorded that they are employed by companies with a commercial interest in importation of the specific foods.
Strengths and weaknesses	The experts felt that their combined expertise covered the range of knowledge that was needed to evaluate the parameter of interest, although the parameter is not something about which any of them has direct knowledge.
Evidence	The elicitor reviewed the evidence that had been collected and presented separately to the experts, and invited Ms Scarlett and Dr Verde to provide initial explanation of the available evidence in their specialist fields.
Structuring	The experts accepted that although in principle the parameter of interest might be structured in terms of growth rates under various conditions, how many journeys might take place under those different conditions and the durations of such journeys, such an approach would be too complex in view of the scarcity of evidence.
Definitions	The parameter *C* is formally defined as the contamination with the bacteria at end user (in CFU/g) after a single randomly selected journey when the contamination at entry was 100 CFU/g. The experts understood that in effect they were being asked about the ratio of contamination at end user divided by contamination at entry, but felt that it was helpful for them to assess this for a specific entry concentration.

Part 1 end time	11:30
Attachments	Briefing document. Evidence summary. SHELF 2 form for the training exercise.

Elicitation Record – Part 1 – Context p2

6.1.4. The quartile method

Having completed the SHELF1 preliminaries, each parameter of interest is considered in turn, using a SHELF2 form to record the elicitation. Again, the form guides the elicitor through a sequence of steps, a key feature of which is the two-stage process beginning with elicitation of individual judgements from the experts and then a discussion leading to the group making judgements to provide an aggregate distribution.

Note that judgements, comments and opinions of the experts that are recorded in a SHELF2 form must always use the expert's anonymised names.

- Headings. The title and session name should be the same as on the SHELF1 form. The parameter name specifies the uncertain quantity whose elicitation is the subject of this SHELF2 form, and the date and time are for the commencement of this elicitation.
- Definition. This field should reiterate the definition of the parameter from the SHELF1 form, but this is an opportunity to review the definition and to make sure that it is clear and precise. The parameter will be referred to here as X.
- Evidence. Evidence relating particularly to X should be reviewed. The experts should be invited to comment on the quality of the evidence and any apparent conflicts between different pieces of information. The experts may be permitted at this stage to consider the broad implications of the data, but should not discuss likely values of X. (This is to avoid anchoring on any numeric values.)
- Plausible range. The elicitor now asks the experts to identify an upper limit U and a lower limit L for X. These 'limits' should be such that, although it may technically be possible for X to be above U or below L, the experts would be extremely surprised if it were not somewhere between the limits. First, the experts may be asked to write down their own limits without discussion. Then without revealing the experts' individual limits the elicitor initially suggests setting U at the largest of the experts' individual U values and setting L at the lowest of the individual L values. If it is generally agreed that the extreme limits were unreasonably wide, they may be changed. However, it is necessary that *all* experts agree that it is extremely unlikely that X would be higher than U or lower than L. (Limits are elicited first to avoid anchoring on a central value.)
- Median. The next two steps require experts to make individual judgements, again without discussion. The first judgement is to specify a median value M. The elicitor will explain that each expert's median is such that he or she judges it to be equally likely that X is above M or below M. Another way to put this is to say that if the expert were to be asked to choose which they would rather bet on, "X greater than M" or "X less than M", they would have no feeling that one choice would be a better bet than the other. Experts should write their M values down without revealing them to anyone at this stage.

  Note that experts who are not familiar with probability judgements may tend automatically to place M equidistant between L and U. The elicitor should invite them to consider whether values of X nearer to L are more or less likely than those nearer to U. To ensure that the experts have a good understanding of how to judge a median, the elicitor may find it helpful to show them example probability distributions like those in Figure 8 to illustrate the possibilities. However, this is best done in the training elicitation, when it does not matter that the choice of examples may unduly influence the experts' own judgements.

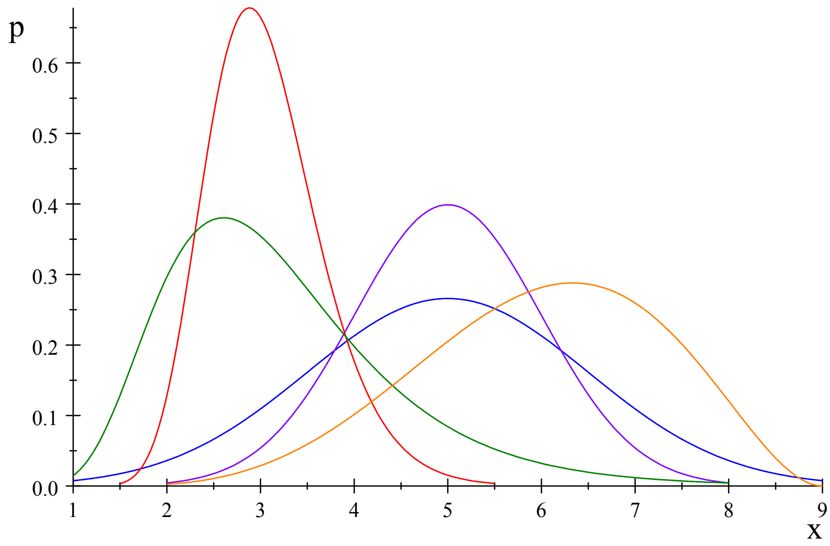


Figure 8: Various distributions with $L = 1$ and $U = 9$. Blue and purple are symmetrical with $M = 5$. Orange is skewed towards higher values of X with $M = 6.1$. Red and green are centred on lower values with $M = 3$, although red is nearly symmetrical

If X is necessarily positive and the ratio of U to L is large, then it is often reasonable to think on a log scale. For example, suppose that $U = 9L$ (U is nine times L). Then $M = 5L$ is equidistant between L and U but the experts may feel that $M = 3L$ is more sensible for the median because this says that there is a factor of 3 between M and L and the same factor between M and U. The green curve in Figure 8 illustrates this case. However, although simple diagrams and arguments like these can be helpful, particularly in the training elicitation, they are not a substitute for careful thought. The key judgement to be made by each expert is where to place M to achieve equal probability for the ranges [L, M] and [M, U].

- Upper and lower quartiles. The experts are now asked to make individual judgements of a lower quartile Q1 and an upper quartile Q3. Each expert should provide a Q1 value, between L and their median value M, such that they judge it to be equally likely that X should be between L and Q1 as between Q1 and M. Similarly, they should give a Q3 value between their median M and U such that they judge X to be equally likely to be between M and Q3 as between Q3 and U.

These are quite difficult judgements for experts to make. Again, those with little experience of the task may be tempted to place Q1 midway between L and M and Q3 midway between M and U. This would almost always be inappropriate because L and U are extreme values that in themselves will have very low probability, whereas M is a very plausible, or even typical, value for X. So values of X near to M will (unless the expert holds unusual beliefs about X) be considerably more likely than values near L or U. Consequently, Q1 and Q3 should in general both be nearer to M than to the limits. The elicitor may find it useful to point this out.

For instance, Figure 9(a) shows the part of the orange distribution in Figure 8 that lies below the median $M = 6.1$. The lower quartile is Q1 = 5.1, which is much closer to M than to L. Similarly, Figure 8(b) is the part of that distribution above the median. The probability is more evenly spread over M to U, but still it is more likely that X is close to M than to U and the upper quartile, Q3 = 7, is therefore also closer to M than to U. The message again is that it really is necessary to think carefully about suitable values for Q1 and Q3.

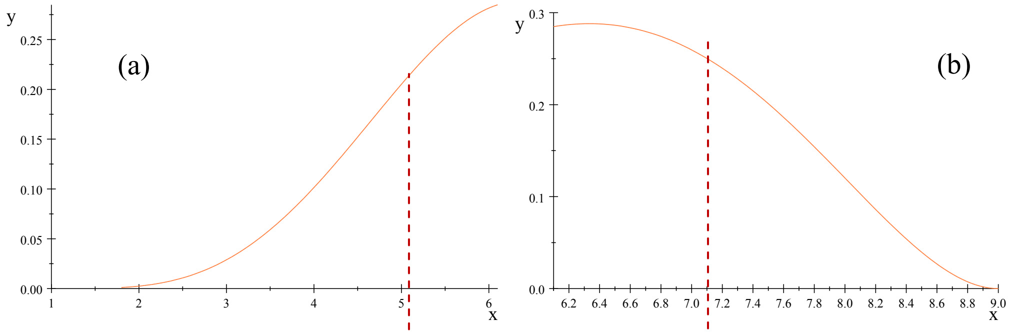


Figure 9: (a) Lower and (b) upper quartiles for the orange distribution in Figure 8

The elicitor may need to explain the quartiles in various different ways.

- As with the judgement of *M*, if the experts were asked to choose whether they would rather bet on "*X* less than Q1" or "*X* between Q1 and *M*" they should have no preference for one choice over the other. There is of course an analogous formulation for Q3. In fact, the quartiles Q1, *M* and Q3 divide the range from *L* to *U* into *four* sections into which the expert should believe *X* is equally likely to fall.

- The elicitor may find it helpful to suggest that it is revealed to the experts that *X* is certainly less than *M*, but no further information is given. Conditional on that information, judging Q1 is exactly like judging the median. Again, there is an analogous formulation for Q3.

- Another helpful device may be to point out that the expert should judge that *X* is equally likely to be between Q1 and Q3 as to be outside that range (and this could again be phrased in terms of no preference between bets).

- Fitting. Once the experts have all written down their individual judgements for *M*, Q1 and Q3, these are all revealed to the elicitor and recorded (using anonymised names) in the two relevant sections of the SHELF2 form. The next step is to fit individual probability distributions to the experts' judgements. It is usual to fit one of the standard forms of probability distributions, such as normal, beta or gamma distributions. The SHELF package includes some simple functions written for the R programming language to fit appropriate distributions and there is also a web-based implementation that can be reached from the SHELF website, but the elicitor may also choose to use other tools.

  Note that these distributions typically have two free parameters, for instance the normal distribution has a mean and variance, which define the fitted distribution. Each expert has given three numbers (and the limits *L* and *U*), and so it will not generally be possible for any of the standard distributions to fit the elicited values exactly. It is therefore necessary to optimise a measure of goodness of fit—the SHELF software uses the sum of squares of differences between the implied elicited probabilities (zero below *L*, 0.25 between *L* and Q1, 0.25 between Q1 and *M*, etc.) and the corresponding probabilities for the fitted distribution.

  The elicitor should show the experts the various fitted distributions. This is the end of the first stage of the elicitation process, the individual elicitations.

- Group elicitation. The experts now discuss the reasons for differences between their judgements and consequent fitted distributions. This is a very important part of the Sheffield method. The objective is for the experts to exchange information and interpretations of the evidence with a view to understanding each other's judgements. The elicitor should ensure that all experts have opportunities to contribute their opinions. An essential skill for the elicitor is to be aware of the group dynamics, to avoid the conversation being dominated by one or a few experts (unless it is clear that the others feel they have nothing useful to contribute on a given point). It is also important to avoid unnecessary repetition or digression; the elicitor should try to maintain a focus on understanding experts' differences of opinion. Experts may have their opinions changed in the discussion, but this should be through intellectual persuasion rather than emotional brow-beating. The elicitor/recorder should record in the SHELF2 form the principal points of this discussion. It is not necessary for the SHELF2 record to be long and detailed, but it should be adequate to explain the reasons for the main divergences of opinion.

Once the discussion has reached a point at which the elicitor judges that no more useful contributions remain to be made, he or she will ask the experts to make joint, aggregated judgements. They should be asked to give joint values for *M*, Q1 and Q3. It is important to make quite clear what such judgements mean. It is not necessary (or even usually realistic) that, despite their original differences of judgement, all experts should now agree to having the same beliefs about *X*, as represented by all having the same *M*, Q1 and Q3. Instead, they are asked to consider what an intelligent and impartial observer might now reasonably believe about *X*, having assimilated the experts' different opinions and arguments. The elicitor should make it clear that this is what EFSA requires as input to its risk assessment. Even though they may not all have the same beliefs about *X* after the discussion, they should be asked to think about reasonable impartial judgements, first for *M* and then for Q1 and Q3. These are recorded in the SHELF2 form, together with any significant new points raised in discussing these judgements[24].

The elicitor may point out that it is better for the experts themselves to make these judgements of what it would be reasonable for an impartial observer to believe, based on their careful discussion of the range of opinion, than for EFSA itself to do so[25].

- Fitting and feedback. The elicitor now fits a distribution to the experts' joint specification of *L*, Q1, *M*, Q3 and *U*. This is shown to the experts and feedback given about how well it represents their judgements. There are various forms of feedback which can be given (and some are provided by the SHELF software). First, the elicitor can show how well the fitted distribution agrees with their joint judgements (either by giving the quartiles of the fitted distribution, to be compared with those judgements, or by giving the probabilities implied by the fitted distribution for the ranges *L* to Q1, Q1 to *M*, etc.).

  The elicitor can also discuss the way in which the fitted distribution relates to the experts' original individual assessments. Where the aggregated distribution differs markedly from the judgements of an individual expert, this may be discussed—has the expert's opinion changed in the discussion, or is the expert happy to recognise that an impartial observer would give weight to the opinions of the other experts? It is important to reaffirm that all experts are comfortable with the reasonable and impartial nature of the final distribution.

  It is also important for the elicitor to be comfortable with that interpretation. In the absence of any sharing and revising of opinions, it would be expected that the final distribution would be more widely spread than any individual's distribution, such that it would spread across the full width of the individual distributions. A helpful reference for this baseline would be an equally weighted linear opinion pool of the experts' initial distributions. (The elicitor may find it helpful to compute this, but it should not be shown to the experts because it may influence

[24] It is possible that, even though the elicitor has explained the nature of the required aggregated judgments the experts cannot agree on suitable aggregate values. In this case, the elicitor will proceed to elicit aggregate judgements and fit final elicited distributions separately for each of the factions.

[25] It is one of the principal tenets of the Sheffield method that .the experts themselves are best able to understand each other's judgements and their underlying reasons, so they are best able to reach the desired conclusion of a distribution representing a rationale "consensus" view of the state of knowledge.

their discussions and push them towards that particular aggregation.) If the final distribution is appreciably narrower than this reference, the elicitor should be comfortable that it represents a genuine sharing and re-evaluation of opinions consistent with the discussion that has taken place.

If any of this feedback suggests that the fitted distribution is not a fair and impartial representation of the experts' joint state of knowledge, then the elicitor should record the reasons and ask for re-consideration of the group judgements.

- Chosen distribution. Once an acceptable aggregated distribution has been agreed, it should be recorded and drawn in this place in the SHELF2 form. This is the outcome of the elicitation for *X*.

- Discussion. Before moving to the next elicitation (or the conclusion of the workshop), the elicitor should ask the experts for comments on the process and the final conclusion. If any expert is uncomfortable with the result or with the way that the process has been conducted, this should be recorded because it is important that the SHELF2 form is an accurate record of the individual and group judgements. The elicitor may ask the experts to sign a printed version of the form to confirm this, but it is usual just to take a lack of negative discussion at the end to imply confirmation.

- End time and attachments. The end time of the elicitation for *X* is recorded and any attachments listed. Attachments may include relevant computations, tables or graphs that have not been imported into the SHELF2 form itself.

The hypothetical example: the SHELF2 form

In this hypothetrical example only limited discussions and explanations of expert opinions are given. In real elicitations it is important to record all the significant arguments made.

The Sheffield Elicitation Framework SHELF v2.0

ELICITATION RECORD – Part 2 – Distribution

Quartile Method

Elicitation title	Chapter 6: hypothetical example
Session	1 of 1
Date	14 November 2010
Quantity	*C*, the contamination at end user given that contamination at entry is 100 CFU/g.
Start time	11:45

Definition	*C* is defined as the contamination with the bacteria at end user [in CFU/g] after a single randomly selected journey when the contamination at entry was 100 CFU/g. The elicitor emphasised that the experts were required to consider a single journey, and that their uncertainty about *C* would include uncertainty about a typical or average growth rate as well as the natural variability from one journey to another. The nature of the end user was explained as being the place to which the food is delivered (without additional processing or repackaging) after entry to the EU. This may be to processing industry or to distributors, for instance.
Evidence	Evidence was reviewed earlier as described in the Part 1 form.
Plausible range	The experts spent some time discussing how the various disparate pieces of evidence might come together to indicate plausible values for *C*. It was clear that most of the evidence was of poor quality or not directly related to the specific pathogen or foods of interest. Consequently, there had to be considerable uncertainty about *C*. The four experts were asked to consider upper and lower plausible limits for *C*. Their lower limits ranged from 0 to 25 CFU/g. The elicitor pointed out that a lower bound of 0 was certainly a logical limit but asked whether it really was plausible that all contamination would disappear. The experts agreed that it would be very surprising (but not impossible) for contamination to be reduced during a journey by a factor of 20, and so a lower limit of ***L* = 5** (CFU/g) was agreed. Upper limits ranged from 1 000 to 5 000 CFU/g. The experts accepted the elicitor's proposal of ***U* = 5 000** (CFU/g) as a value above which it would be very surprising for contamination to reach on a single journey.

Elicitation Record – Part 2 – Distribution – Quartile Method p1

The Sheffield Elicitation Framework SHELF v2.0

Median

Each expert was next asked (without further discussion) to write down his or her median value (*M*) for *C*. The elicitor instructed them first to consider what a typical value might be for *C*, using the evidence regarding a typical length of journey under typical conditions and regarding the microbiology. However, they should then adjust this 'typical' value for *M* until they were comfortable that *C* was equally likely to take a value above M as below M.

Upper and lower quartiles

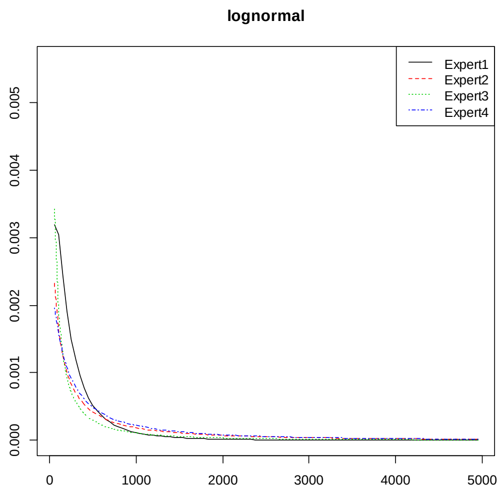
The experts were similarly asked to write down lower quartile (Q1) and upper quartile (Q3) values. The elicitor reminded them of the ways to make these judgements that had been discussed in the training and the practice elicitation.

When all experts were finished, their judgements were revealed as follows (anonymised).

Expert	**1**	**2**	**3**	4
M	200	350	100	500
Q1	100	100	20	125
Q3	400	2 000	500	1500

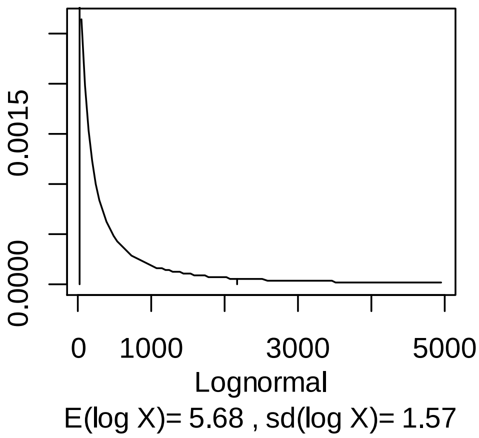
Fitting

The elicitor used SHELF software to fit a smooth lognormal distribution to each expert's judgements, as shown below.



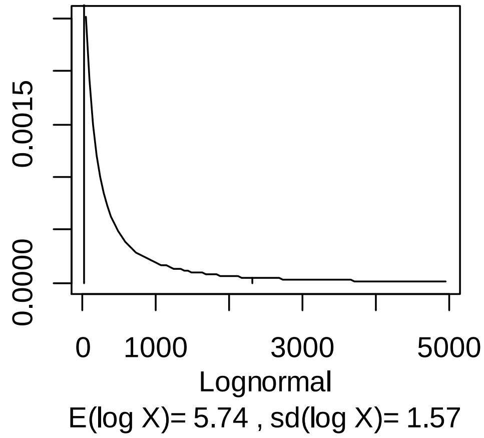
Elicitation Record – Part 2 – Distribution – Quartile Method p2

The Sheffield Elicitation Framework SHELF v2.0

Group elicitation	The elicitor invited the experts to discuss their different assessments of median and quartiles. It was noted that expert 3's (E3) median at 100 implied that E3 thought it was equally likely that the contamination level would increase or decrease in the journey, whereas the other experts all thought that it was more likely to increase. At the other extreme, E4's median of 500 implied an equal probability above and below a fivefold increase in contamination. The other two experts, E1 and E2, had Q1 values at 100, so that they believed the probability of a decrease in contamination was just 0.25. The elicitor also drew attention to similar differences in judgements of upper and lower quartiles, but the differences in median were most easily understood by the experts and led to a substantial discussion. On the basis of this discussion E1 felt that (s)he had not fully appreciated that the majority of journeys would not be made in the controlled environmental conditions (s)he had expected. Therefore, E1 thought that his/her elicited values were generally too low. On further advice from Adviser A, E2 and E4 felt that their upper quartiles (Q3) had been set too high. Although contamination might reach 1 500 or 2 000 CFU/g in some journeys, such high levels should not be expected in as many as 25 % of cases. The experts then discussed setting median and quartile values that would represent a reasonable judgement of an impartial expert based on the range of opinions expressed by E1, E2, E3 and E4. The following values were chosen: M = 275 Q1 = 100 Q3 = 950
Fitting and feedback	The elicitor showed the following fitted distribution.  She reported that, according to this distribution, the median is 290, the lower quartile is 100 and the upper quartile is 840. The experts appreciated that the fitting process would produce somewhat different values than those elicited, but they felt that on this occasion the fitted distribution would understate the risk of high levels of contamination.

Elicitation Record – Part 2 – Distribution – Quartile Method p3

The Sheffield Elicitation Framework SHELF v2.0

Fitting and feedback	They agreed after group discussion to increase the median to 330. The new fitted distribution is shown below.  According to this distribution the median is 310, the lower quartile is 110 and the upper quartile is 900. The elicitor also reported that, according to the fitted distribution, there was a 10 % probability that the contamination on arrival would be below 42 CFU/g and a 10 % probability of it being above 2 330 CFU/g.
Chosen distribution	The experts agreed to adopt the second fitted distribution shown above, which is lognormal with parameters 5.74 and 1.57.
Discussion	The experts felt that the elicitation exercise had been interesting and a good way to bring together disparate knowledge about a difficult uncertain quantity of interest to EFSA. They thought that the final distribution was a crude but reasonable representation of their combined knowledge. None of the experts expressed any concerns with the process and the outcome.

End time	14:20
Attachments	None

Elicitation Record – Part 2 – Distribution – Quartile Method p4

6.1.5. Conclusion and post-workshop reporting

The SHELF 1 and SHELF2 documents, together with any attachments, constitute a formal record of the elicitation workshop. Together with the elicitor's personal summary and commentary on the process and the results, and any other information specified in section A.6.2, these constitute the Technical Support Document which the elicitor should present to EFSA.

Under time pressure in the workshop it may not always be possible to complete the forms in as comprehensive or tidy a fashion as the elicitor/recorder would like. It is then permissible to do final work on the forms after the workshop before including them in the Technical Support Document. However, that should always be done as soon after the workshop as possible, and if any material changes have been made the revised forms must be sent to the experts for approval as an accurate record. It is good practice then to send also the forms as originally completed in the workshop (and at least implicitly approved by the experts at the time) for comparison.

It is also good practice to ask experts to complete an appraisal form, in which they are asked for their comments and evaluations about the planning, organisation and running of the workshop. This may be done in the workshop if there is time at the end of the elicitations, or sent to the experts afterwards.

6.1.6. Qualities of the elicitor

Through the sharing of knowledge and the discussion of the experts' opinions, the Sheffield method has the potential to deliver a more informative elicited distribution for the quantity of interest. However, this process of sharing and behavioural aggregation also poses challenges in implementing the method. The most important is in the management of the interactions between the experts, ensuring that all relevant data and opinions are taken account of and that the final elicited distribution truly represents a reasonable belief about the uncertain quantity based on the data, knowledge, opinions and discussion. It is essential that the elicitor is experienced in the management of expert groups and in the process of behavioural aggregation. And of courase he or she needs in particular to be familiar with the Sheffield method.

6.2. Cooke's method

Cooke's model is a performance-based linear pooling model based on statistical hypothesis testing. It aggregates individual experts' PDFs in order to obtain one combined PDF for each variable. Experts give predefined quantiles, or percentiles, of distributions, typically 5 %, 50 % and 95 %. Experts can be weighted equally or according to their (relative) expertise, as indicated by their performance on seed variables. Seed (calibration) variables are variables from the experts' field whose realisations are (or will be) known to the analysts, but unknown to the experts.

Cooke's model assumes that the (future) performance of the experts on the variables of interest (target variables) can be judged on the basis of their (past) performance on the seed variables. Therefore, the seed variables must resemble the variables of interest as much as possible.

The need for seed variables implies that extra steps are necessary in the pre-elicitation stage: the identification of seed variables and a dry-run exercise to finalise the elicitation questions.

This is followed by a plenary meeting of all experts in which the issues are discussed, the study design is explained and a short elicitation exercise is carried out. This involves a small number of seed variables, typically five. Experts are shown how the scoring and combining works. Afterwards, the experts are elicited individually.

When experts are dispersed it may be difficult and expensive to bring them together. In such cases the training is given to each expert in abbreviated form. In general, it is not advisable to configure the exercise such that the presence of all experts at one time and place is essential to the study, as this makes the study vulnerable to last minute disruptions.

Let us return to the subject of "how the weighting and combining works".

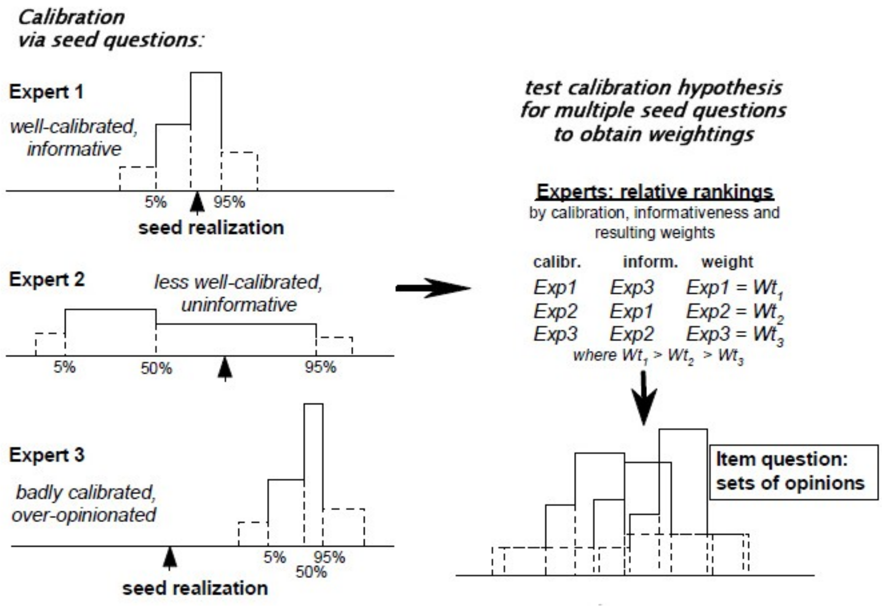
The individual experts' weights are based on two quantitative measures of performance: calibration and information. Calibration measures the statistical likelihood that the realisations of the seed variables correspond, in a statistical sense, with an expert's assessments. If this likelihood score is below a certain cut-off level, the expert is un-weighted. The cut-off could be chosen by the elicitor or determined by optimising performance of the combined virtual expert.

The calibration takes values between 0 and 1, with a high score implying that the expert's PDFs are statistically supported by the set of seed variables. Information represents the degree to which an expert's PDFs are concentrated, relative to some chosen background measure, and it is always positive. An average information score is calculated for each expert. "Good expertise" corresponds to good calibration and a large amount of information. The virtual (combined) expert resulting from the combination of experts' opinions will also have a calibration and an information score. The individual experts' performance-based weights are proportional to the product of calibration and information.

A schematic description of a hypothetical performance aggregation scheme is presented in Figure 10. This is not an accurate description of the process, but rather a vague graphical representation meant to help in acquiring some intuition behind the aggregation process. Nevertheless, it is maybe worth clarifying a few matters.

One may notice that the extreme histogram blocks are dotted. This helps pointing out that, even if the experts only give their 5 % and 95 % quantiles (and no physical bounds), a range (bounded interval) is necessary for each variable. This range will be the same for all experts (per variable), and it is called an *intrinsic range.* The intrinsic range is the smallest interval that contains all assessments for a given question/variable (and the realisation, if available), overshot by k % above and below, where k is chosen by the analyst and it is typically 10.

On the left-hand side of the figure, a typical situation for one seed variable is shown, whereas on the right hand side hypothetical weights derived from multiple seed questions are presented. Even though expert 1 seems to perform much better than the other experts on the one seed question shown, the situation does not have to stay the same when a larger number of seed variables are analysed.



from Aspinall's briefing notes

Figure 10: The schematic description of performance aggregation via seed questions

For a weighted combination of expert CDFs (Figure 11), take the weighted combination at all break points and then linearly interpolate to obtain something of the following sort:

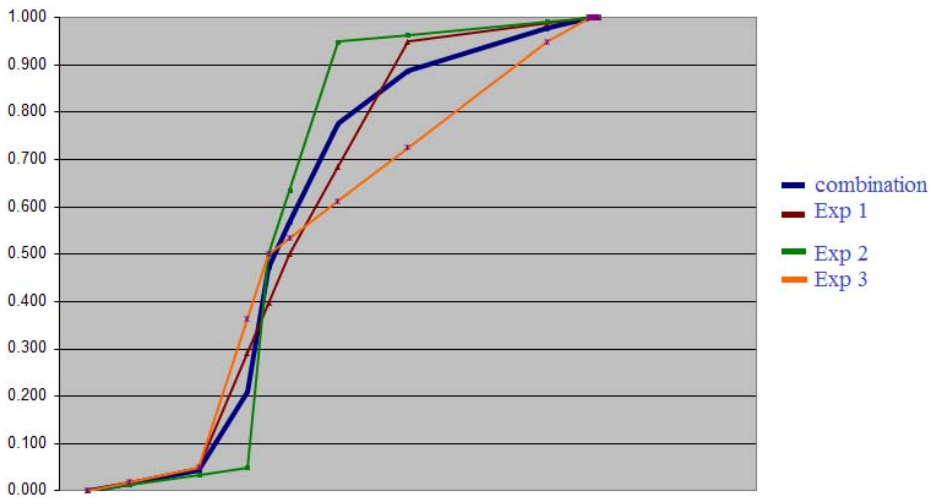


Figure 11: Cumulative distribution function (CDF) by linear interpolation of weighted results

The hypothetical example: elicitation question

Cooke's method requires seed variables that need to be as representative as possible for the variable of interest. To emphasise this need and give some guidance on how this need can be fulfilled we will make the hypothetical example a little more specific than before.

"Assume that a contaminated consignment of raw poultry is entering Europe through border control point BCP. The contamination level with Campylobacter at the time of crossing the border is 100 CFU/g. The consignment will then be transported and handled in countries A, B and C (in Europe) under usual conditions.

Consider one random journey. What level of contamination (in CFU/carcase) would be in the consignment when it reaches the end user of that journey?

From this point on, the formulation of the question could continue differently. We give here two possible endings of this question and recommend using the former:

"To express the uncertainty associated with the level of contamination, please provide the 5th, 95th, and 50th percentiles of your estimate.

Please give reasons for your opinion."

or

Please give a credible interval (your 5—95% confidence) which you judge should encompass the level of contamination, and your central estimate (median) for that value (i.e. to locate the value within your credible range).

Please give reasons for your opinion."

6.2.1. Preparation

A scientific problem is amenable to expert judgement if there is relevant scientific expertise. If a problem is an expert judgement problem, then necessarily there will exist somewhere relevant experiments, observations or measurements. The available information should be collected; the majority of it should be shared/presented in an elicitation format document. Nevertheless, a number of results should be used to derive seed variables[26]; hence, these sources and information therein should not be shared. The elicitation format document contains any general known data, the exact questions to be assessed by the experts (seed and target), necessary explanation to questions, and the format in which the assessments need to be provided.

The hypothetical example: elicitation format document

Below we present the layout/table of contents of the elicitation format document and we give an example of a possible training question in the format used in the elicitation (this example can be also found in Section "Format" in the elicitation format document).

[26] When the seed variables contain crucial and unique information about the problem at hand, they should definitely be elicited separately and released before eliciting the quantities of interest. Nevertheless that is very rarely the case, since the seed variables have to be only representative for the question of interest. That does not necessarily mean that the knowledge about these questions is essential in answering the questions of interest.

Impacts of transporting a consignment contaminated with a non-toxigenic bacterium:

Expert elicitation protocol

- *Project overview*
- *Scope*
- *Detailed definitions for the model parameters*
- *Assumptions*
- *Method*
- *Expert names*
- *Format*
- *What is a good probability assessor?*
- *Seed variables and variables of interest*
    - *Seed variables*
    - *Variable of interest*
- *Answers to practice questions:*
- *Graphs, tables, and other common reference material*

The full elicitation format document for the particular example can be found in the at the end of this section.

More details are provided about a possible training question that helps experts understand how their answers are interpreted.

> *An example of a question (in the format used in the elicitation) and the interpretation of the answer*
>
> *Commission Decision 94/360/EC prescribes the level of physical checks for certain products. What is the minimum number of consignments (in percentages) to be subjected to a physical check for poultry meat? (http://www.porthealth.eu/Import-Process-POAO.htm)?*
>
> *Please provide the 5th, 95th, and 50th percentiles of your estimate.*
>
> *5 %________________95 %___________________50 %____________________*
>
> *Presumably, this number is uncertain. If you fill in:*
>
> *5 %____10__________________95 %__70_______________ 50 %_40_________*
>
> *This means that you believe there is a 5 % chance that the actual number is below 10, a 5 % chance that the actual number is above 70, and a 50–50 chance that it is below 40.*

The true value was 50. This is not a surprising value relative to this assessment. If the value were 8 this would be surprising, as would 80. In each case, the realisation would be outside the 90 % confidence band.

An expert's probabilistic assessments are statistically accurate if 10 % of the realisations fall outside the 90 % confidence band; 50% of the realisations fall on either side of the median (50 % value).

If your assessments had been:

5 %_5__________________ 95 %__90____________ 50 %_60____________

you would have been equally unsurprised, but your assessments would be less informative.

6.2.2. List of participants/roles

The hypothetical example: experts used in the workshop

Experts were defined as persons having extensive knowledge and experience of import/transport in the main importing countries as well as the behaviour of pathogens, during different stages of transportation. They were identified as either:

- *quality managers having specialised knowledge of the behaviour of the pathogen during transportation; or*
- *food inspectors with regular experience in the control of this (specific) food item, preferable at a main point of entry; or*
- *microbiologists/food scientists with knowledge of the production, processing and transportation conditions of this (specific) food item; or*
- *border veterinarians.*

Furthermore, import/trade experts that are able to cover the situation in the three main importing countries should be available. These experts will be used for gathering/providing all available information from their field of expertise. They will participate in the workshop and present/defend all information to the experts who are elicited.

It is worth mentioning, that the workshop is essential for this particular example (and other similar situations), due to the variety of fields of expertise needed.

In total, 21 experts with varying knowledge bases and employed by a variety of organisations were identified. Three of the identified experts were to take part in a dry-run session.

The remaining 18 experts, employed by 13 different organisations, were approached to participate in the expert elicitation session with the restriction that only one person per organisation should attend the expert panel. From the potential panel of 13 experts, one person was not interested in participating and one was not available for the duration of the study, resulting in a panel of 11 experts. The experts currently applied their expertise in various fields, including government (one person), industry (four persons), science (three persons) and a combination of science and practice (three persons).

The participants can be grouped in:

- The Elicitation Group, consisting of one normative elicitor, experienced in subjective probabilities and EKE studies, and one substantive elicitor, experienced in the experts' field of interest.
- Experts used in the workshop
- Experts used in a dry run (their presence during the workshop is not required). The design of the elicitation format document is discussed with the experts selected for the dry-run exercise.

As mentioned earlier, during the elicitation, the experts are asked to provide their subjective PDFs on the variables of interest and the seed variables, through three percentiles of the distributions, the 5th, 50th, and 95th percentiles. They are also asked to provide their rationales behind their probability assessments.

Whenever substantive questions arise (ambiguities in the formulation, assumptions made in a particular question, etc), the substantive elicitor should clarify the matters. The normative elicitor's role is to clarify all questions about subjective probabilities and the EKE technique in general. Moreover the normative elicitor should pay attention and try to help the expert in avoiding

- possible biases (e.g. anchoring, overconfidence, etc.)
- giving symmetric distributions when not intended
- confusing the elicited quantiles with other quantiles

The hypothetical example: dry run

Three of the identified experts, one scientist, one working in practice and one import/trade expert, were selected for use in a dry-run session; hence, they were not used in the elicitation.

Prior to the elicitation session, a special session was held with the three dry-run experts in which they were asked to provide their comments on the elicitation format document. The documents had been given to the dry-run experts beforehand and they were explicitly asked not to give their assessments, but rather to indicate whether experts would be able to estimate the variables. After the dry-run session, the document was finalised; its final form can be found at the end of this section.

6.2.3. Equipment

There are several options when choosing the appropriate/needed equipment for an elicitation exercise. All of them involve one version or another of the software package EXCALIBUR[27] mentioned in appendix A, section A.4.3.2.1.

A computer, data projector and screen are essential during the workshop (e.g. for the training exercise, for the presentation of the context, problem, available data, etc.).

[27] Available for free download from http://risk2.ewi.tudelft.nl/oursoftware/6-excalibur

The hypothetical example: equipment for the elicitation workshop

- *a box of pencils equipped with erasers*
- *software package EXCALIBUR v. 1.5.3 light*

6.2.4. Seed variables

The seed variables are not only important in determining the weights for combining experts' assessments, but also they provide empirical evidence of the performance of the combined assessment (the combined virtual expert). Two types of seed variables (relevant to the quantity of interests) are available:

1. Domain variables: these variables fall in particular in the field of the experts. They have the same physical dimension as the variables of interest, measured from past studies at the same site or from similar circumstances.

2. Adjacent variables: these variables fall into fields which are adjacent to the field of expertise of the experts in question. They have different dimensions from the variables of interest, but represent variables for which the experts should be able to give an educated guess[28].

As a loose criterion, a seed variable should be a variable for which the expert may be expected to make an educated guess, even if it does not fall directly within the field of the study at hand. It will often happen that a given seed variable is a domain variable for one expert and an adjacent variable for another expert.

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

For each sub field at least 10 seed variables are needed, preferably more. Distinct seed variables may be drawn from the same experiment, but sufficient different studies should be also used in deriving other seed variables.

Seed variables may be, but need not be identified as such in the elicitation. If possible, the analyst should be unaware of the values of the seed variables during the elicitation. The number of seed variables for assessments of uncertain quantities with continuous ranges is typically between 10 and 20. Some more general guidance about the selection of these variables is available in Cooke (1991), Aspinall and Cooke (2013).

The abundance of applications also offers a good starting point in analysing desirable properties of seed variables. Nevertheless, further research would be helpful into what properties make a set of seed questions suitable for calibration.

The hypothetical example: seed questions

In our example, 15 domain and adjacent seed variables were defined. Five domain seed variables are presented in the elicitation format document example. The adjacent variables are similar to the first practice question in the elicitation format document.

The seed variables for this study were defined based on three different particular experiments. One example of a seed variable is given below:

Consider a flock of broiler chickens in 1995 just before it is prepared for transportation to the processing plant. The flock became colonised with Campylobacter during rearing and all birds are carrying the organism both internally and externally. A random broiler chicken is

[28] An example of a study where adjacent seed variables were used is Goossens et al. (1992).

sampled from this flock and the numbers of Campylobacter bacteria in this chicken, both in its caecum (internal carriage) and on its exterior (external contamination), are counted. Following slaughter at the farm, the carcase is sampled directly after bleeding by the carcase rinse method. The number of Campylobacter organisms is determined by a standard method (i.e. plating on charcoal cefoperazone desoxycholate agar).

Seed variable S1*.The number of Campylobacter organisms (CFU per gram) found in the caecal content of the broiler chicken just before it would have been transferred to a transport crate.*

To express the uncertainty associated with the number of Campylobacter organisms, please provide the 5th, 95th, and 50th percentiles of your estimate.

It is worth emphasizing that the seed questions chosen for this example merely serve an illustrative purpose. Much more time, effort and expertise have to be spent for the formulation of appropriate seed questions in a real application[29]*.*

When creating a case file in EXCALIBUR, one should start with adding the experts, the variables (seed and of interests), the realisations of the seed variables and the experts' assessments. Once this is done, several combinations can be investigated. Figure 12 shows a screenshot from EXCALIBUR, in which eight of the experts from our example are identified as Expert1, …, Expert8, and the full name is omitted for confidentiality reasons. Ten of the seed questions used in the example and their realisations are also shown. The first five of them are the seed variables detailed in the elicitation document.

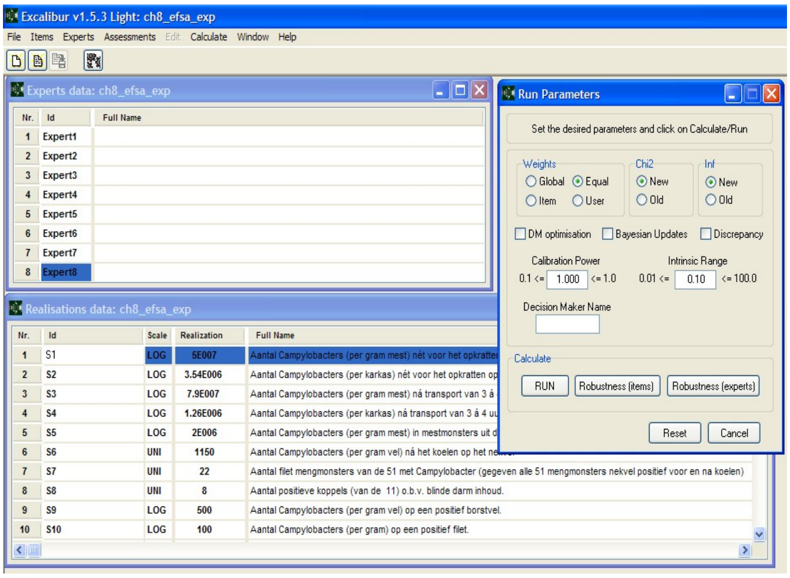


Figure 12: Screenshot of EXCALIBUR software

[29] It is desirable of avoiding excessive reliance on few sources of information.

6.2.5. Timeline of the workshop

The duration of the workshop depends on the resources: it can range from half a day to three days.

It is advisable to have a plenary meeting with all experts in which the issues are discussed, the study design is explained, and a short elicitation exercise follows. Experts are shown how the scoring and combining works. Afterwards, the experts are elicited individually.

When experts are dispersed it may be difficult and expensive to bring them together. In such cases the training is given to each expert in abbreviated form. In general, it is not advisable to configure the exercise such that the presence of all experts at one time and place is essential to the study, as this makes the study vulnerable to last minute disruptions.

Some elicitors distinguish the seed variables and elicit them before the variables of interest, allowing discussion only after the seed variables are elicited. Other elicitors elicit all variables at the same time and have the plenary meeting before the elicitation. We recommend the latter option, where a shared understanding and knowledge about the model, the context and the "language" used are achieved before the elicitation. When the resources, the time and the availability of the experts permit we recommend a three-day workshop.

The hypothetical example: workshop timeline (three days)

Day 1

Part 1: Introductory remarks

- *Introduction to the problem*
- *Introduction to EKE: aims and approach of pooled EKE, training and elicitation exercise*

Part 2: Discussions

Day 2

Part 1: Elicitation

Part 2: Discussion of preliminary results

Day 3

Presentation/discussion of results

Outstanding issues

Wrap-up, next steps

When resources are limited a shorter (one day) workshop could be organised as follows.

The hypothetical example: workshop timeline (one day)

8:30—9:00 *Arrival and coffee*

9:00 –10:00 *Introduction to the problem/model*

10:00–11:00 *Introduction to EKE: aims and approach of pooled EKE*

Break

11:15–12:45 Training session

12:45—13:30 Lunch

13:30–15:00 Discussions

Break

15:15–17:45 Elicitation

6.2.6. Elicitation

Cooke's model seeks to satisfy the demands of openness and objectivity in science, as well as demands of freedom from conflict of interest and legal liability. Expert judgement must be open to peer review; on the other hand, the experts' affiliation or professional activities may create a conflict of interests if their name is associated with the actual assessments. The proposed procedure to deal with these issues includes the following:

- Experts' names and affiliations are published in the study.
- All information is available for peer review but not available for unrestricted distribution.
- Assessments are not associated with the names but identified as "expert 1", "expert 2", etc.
- Expert rationales are available for unrestricted distribution.

Any further published use of an expert's name requires the expert's approval.

The elicitation format document is handed out to the experts for preparation of the elicitation interview.

The experts who could not attend the group meeting are sent this document by mail and given a short introduction to the above-mentioned aspects, as far as they are not familiar with them, prior to their elicitation.

Experts are interviewed individually by the elicitation team consisting of one normative elicitor and one substantive elicitor (there may be several teams).

During the elicitation, the experts are asked to provide their subjective PDFs on the variables of interest and the seed variables, through three percentiles of the distributions, the 5th, 50th, and 95th percentiles. They are also asked to provide their rationales behind their probability assessments.

6.2.7. Analysis

When the elicitation interviews are completed, the experts' assessments (with rationales) should be examined for consistency and the assessments should be returned to individual experts for confirmation.

For each variable, the individual experts' assessments are aggregated into one combined PDF (the virtual expert's distribution). Experts can be weighted equally or according to their performance, as measured on the seed variables.

The software package EXCALIBUR, which implements Cooke's model, is used to aggregate individual experts' assessments into one combined PDF per variable. Three different combination schemes are usually used, including equal weighting and performance-based weighting with or without optimisation of the performance of the virtual combined expert.

Since any combination of expert distributions yields assessments for the seed variables, any combination can be evaluated on the seed variables. In particular, we can compute the calibration and the information of any virtual combined expert (corresponding to any weighting scheme).

In case of performance-based weighting without optimisation, the cut-off level[30] for calibration is set to zero and each expert is weighted according to their un-normalised weight. With optimisation, the cut-off level is chosen such that the un-normalised weight of the virtual expert is maximal, and experts having a calibration score below this cut-off level are un-weighted. The best-performing virtual combined expert (weighting scheme) is chosen for further post-processing of the results.

In EXCALIBUR the performance weights described above are called "global weights". A variation on this scheme allows a different set of weights to be used. These weights use the information scores for each item rather than the average information score. "Item weights" are potentially more attractive as they allow experts to up- or downweight themselves for individual items according to how much they feel they know about that item. "Knowing less" is translated as choosing quantiles further apart and lowering the information score for that item. Of course, good performance of item weights requires that experts can perform this up-/downweighting successfully. Anecdotal evidence suggests that item weights improve over global weights as the experts receive more training in probabilistic assessment. In both global and item weights calibration dominates over information; information serves to modulate between more or less equally well-calibrated experts.

In addition, EXCALIBUR has a facility for importing user weights from an external source. User weights may be derived in some other way, externally, and then used in EXCALIBUR.

In the software, the virtual combined expert is called a decision maker (DM).

The hypothetical example: different weighting options

For our example we investigate three combinations: the equal-weighted combination (EWDM), the performance-weighted combination (global weights) without optimisation (GWDM_NOT), and the performance-weighted combination (global weights) with optimisation (GWDM) (Figures 13 and 14).

Expert scores: ch8_efsa_exp

Results of scoring experts
Bayesian Updates: no   Weights: global   DM Optimisation: yes
Significance Level: 0.8283   Calibration Power: 1

Nr.	Id	Calibr.	Mean relative total	Mean relative realization	Numb real	UnNormalized weight	Normaliz.weigt without DM	Normaliz.weig with DM
1	Expert1	0.002809	1.065	0.9953	10	0	0	0
2	Expert2	0.01397	1.403	1.328	10	0	0	0
3	Expert3	0.6827	0.5346	0.5661	10	0	0	0
4	Expert4	0.002809	0.8698	0.7343	10	0	0	0
5	Expert5	0.2282	0.5947	0.2557	10	0	0	0
6	Expert6	0.6827	1.064	0.978	10	0	0	0
7	Expert7	0.8283	1.675	1.632	10	1.352	1	0.5
8	Expert8	0.0002637	1.246	1.122	10	0	0	0
9	GWDM	0.8283	1.675	1.632	10	1.352		0.5
10	GWDM_NOT	0.4735	0.5697	0.5294	10	0.2507		0.09152
11	EWDM	0.2895	0.2854	0.2036	10	0.05894		0.02314

Figure 13: Different results of expert scores implemented in EXCALIBUR

[30] The cut-off level is mentioned at the beginning of this section and in appendix A.4.3.2.1.

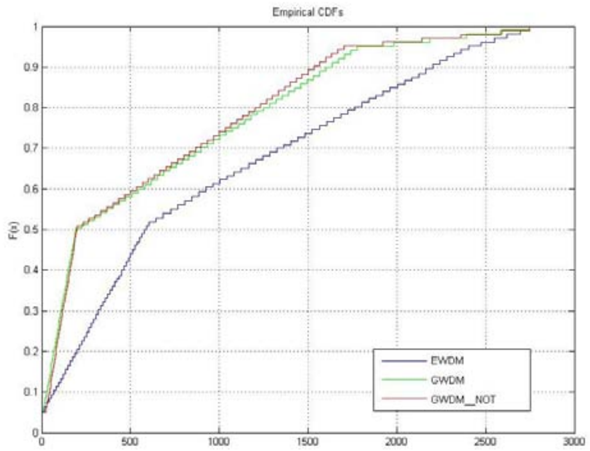


Figure 14: Different results (cumulative distribution functions) for different weighing options

The GWDM (where all experts receive a weight that is proportional to the product of their calibration and information score) obtains a higher calibration and a higher information score than the EWDM. When we optimise, the best combination coincides with the best expert, and it is even more calibrated and informative. It is worth noting at this point that this is an artificially constructed example. The performance-based combination being the best expert is not unusual but it does not always have to happen (in happens in about 30 % of the cases.)

Weight zero does not mean value zero. It simply means that the zero-weighted expert's knowledge was already contributed by other experts and adding this expert would only add a bit of noise. The value of un-weighted experts is seen in the robustness of the answers against the loss of experts (see further in this section).

There is no mathematical theorem that either item weights or global weights outperform equal weighting or outperform the best expert. It is not difficult to construct artificial examples where this is not the case. Performance of these weighting schemes is a matter of experience. In practice, global weights are used unless item weights perform markedly better. Of course, there may be other ways of defining weights that perform better, and indeed there might be better performance measures.

6.2.8. Robustness

Robustness analysis includes the removal of one expert or one seed variable from the dataset at a time and recalculation of the resulting combined virtual expert distribution, to assess the relative information loss of the new combined distribution w.r.t the original one. The larger the relative information, the more the expert/seed variable contributes to the combination, and the results may not be replicated if different seed variables or different experts were to be chosen.

The hypothetical example: robustness

For our example (Figures 15 and 16), we chose the GWDM with no optimisation to perform the robustness analysis. One other measure (except the relative information loss) that we could investigate is the change in calibration scores (fifth column) when one expert/item is removed at a time. It is worth mentioning that the calibration scores are very robust with respect to the choice of experts, and even when expert 7 (the best expert) is removed from the analysis the performance based weighted combination of the other experts achieves a similar calibration score. Experts' 6 assessments or lack thereof will influence the entire analysis in ways that are not obvious by just looking at the displayed results. Details are not necessary here; this is just to point out that the influence of each expert is more subtle than just investigated through their weight.

Robustness for Experts: ch8_efsa_exp

Robustness analysis on Experts
Bayesian Updates: no Weights: global DM Optimisation: no
Significance Level: 0 Calibration Power: 1

Nr.	Id excl.exp	Rel.info/bgr. total	Rel.info/bgr. realization	Calibr.	Rel.info/or.DM total	Rel.info/or.DM realization
1	Expert1	0.57	0.5297	0.4735	0.0006169	0.0006106
2	Expert2	0.5692	0.5291	0.4735	0.002559	0.002405
3	Expert3	0.6251	0.6618	0.4735	0.1831	0.1218
4	Expert4	0.5712	0.5311	0.4735	0.001364	0.001447
5	Expert5	0.4795	0.427	0.4735	0.09569	0.1031
6	Expert6	0.6041	0.5731	0.2895	0.1081	0.104
7	Expert7	0.3791	0.345	0.4925	0.316	0.313
8	Expert8	0.5687	0.5283	0.4735	0.0004015	0.0004133
9	None	0.5697	0.5294	0.4735	0	0

Figure 15: Robustness analysis (exclude experts) implemented in EXCALIBUR

In this artificially constructed example, the selection of seed questions seems to be very appropriate since the results do not significantly fluctuate with the removal of any of the seeds. Of course, this is not a rule, and robustness per items could give insights into badly formulated or unrepresentative seed questions that alter the analysis.

Robustness analysis / items

Robustness analysis on seed Items
Bayesian Updates: no Weights: global DM Optimisation: no
Significance Level: 0.0000 Calibration Power: 1.0000

Nr.	Id of excl. item	Rel.info/bgr. total	Rel.info/bgr. realization	Calibr.	Rel.info/or.DM total	Rel.info/or.DM realization
1	S1	0.4775	0.4233	0.571	0.09442	0.1038
2	S2	0.5432	0.493	0.571	0.07217	0.07824
3	S3	0.4745	0.4198	0.571	0.09878	0.1085
4	S4	0.5366	0.4845	0.4048	0.06915	0.07406
5	S5	0.4777	0.4242	0.571	0.1016	0.1111
6	S6	0.5469	0.4975	0.571	0.07054	0.07672
7	S7	0.5007	0.4456	0.4048	0.05734	0.0613
8	S8	0.6258	0.5871	0.4048	0.01058	0.01081
9	S9	0.5725	0.533	0.4048	0.02273	0.01701
10	S10	0.5388	0.4953	0.4048	0.01902	0.01338
11	None	0.5697	0.5294	0.4735		

Figure 16: Robustness analysis (exclude seed questions) implemented in EXCALIBUR

6.2.9. Feedback and documentation

Experts must have access to their assessments, their calibration and information scores, their weighting factors, any conclusions about over- or underconfidence, and conclusions about their tendency to over- or underestimate. This information should be made available upon request.

Names of the experts should only be used with the experts' permission.

All relevant information and data is recorded in a formal report to be presented to the problem owner and the experts. The results of individual experts are treated anonymously. This is referred to as a result report at the end of this section.

The hypothetical example: the Elicitation Protocol

Impacts of transporting contaminated consignment with a non-toxigenic bacteria: expert elicitation protocol

Project overview

A Member State reported through the rapid alert system that a non-toxigenic bacterial pathogen was detected at a regular border control in a specific food item from a third country (outside the EU). Regarding the conditions of production in that third country it is assumed that more consignments of this food may be contaminated, while other origins seem to be safe.

To determine the necessary detection level at border control a simple risk model is constructed. This model links the final contamination at the end user with the contamination at the border (point of entry) by a single parameter, the growth/survival/inactivation rate of this pathogen during transport from border to the end user. It is assumed that the transport conditions are so diverse and too divided into short passages for the application of a more stratified model to be feasible. Instead experienced experts should judge on the parameter taking their knowledge on transportation conditions and pathogen characteristics into account.

Scope

The Working Group concludes that the conditions during transportation could be especially favourable for growth of the pathogen, but no specific information on conditions during transport that would enable the modelto be refined or the global growing rate $R_{transport}$, to be estimated is known. The Working Group therefore recommends the use of EKE to estimate this parameter.

Detailed definitions for the model parameters

$$C_{\text{end user}} = C_{\text{entry}} \times R_{\text{transport}}$$

where

$C_{end\ user}$ *is the contamination with the bacteria at end user (in CFU/g)*

C_{entry} *is the contamination with the bacteria at border (in CFU/g] and*

$R_{transport}$ *is the growing/survival/inactivation rate during transport and storage (dimensionless)*

- *food item—poultry*
- *non-toxigenic bacteria—Campylobacter*
- *end users*

Assumptions and background information

- *All available information on model parameters or influencing factors*
  - *growth, survival and inactivation parameters of the pathogen under experimental conditions;*
  - *characterisation of the food matrix, e.g. pH, water activity, constituents, etc.;*
  - *conditions of usual transport and storage, e.g.size of consignments, transportation means, storage conditions: ambient temperatures / chilled / frozen, life period / shelf period / transport distances;*
  - *intracommunity trade regarding this food (no stratification by origin);*
  - *import of this food into each EU country from the specific origin.*

Levels of the pathogen detected at border control

Method

The model quantification will be based on available data, whenever possible. However, incomplete and/or unreliable data should be complemented by data provided by experts via a structured expert judgement elicitation. Structured expert judgement has been widely applied in risk analysis for many years but (understandably) still generates scepticism among researchers, stakeholders and the general public. The use of structured expert judgement typically involves greater uncertainty. For these reasons, it is imperative to fully document the process, and to validate the uncertainty assessments to the maximum extent possible. Validation requires eliciting uncertainty on variables whose true values will be known within the time frame of the study.

Expert names
Expert names and affiliations are part of the published documentation, as are the individual assessments. The association of names and assessments is preserved in the ubpublished records of the research group and is accessible for review. However, the association of names with individual assessments is never included in open publications.

Format
All of the questions will have a similar format. You will be given the description of an uncertain quantity taking values in a continuous range. You are asked to quantify your uncertainty by giving the 5th, 50th and 95th percentiles of your uncertainty distribution. Examples of quantities of model variables for published models from previous years, related variables and quantities are given below, in Tables 1,2,….

For example:

Commission Decision 94/360/EC prescribes the level of physical checks for certain products. What is the minimum number of consignments (in percentages) to be subjected to a physical check for poultry meat? (http://www.porthealth.eu/Import-Process-POAO.htm)?

Please provide the 5th, 95th, and 50th percentiles of your estimate.

5 ____________________ 95 %____________________ 50 %____________________

Presumably, this number is uncertain. If you fill in:

5 %____10________________ 95 %__70_____________ 50 %_40_________

This means that you believe that there is a 5 % chance that the actual number is below 10, a 5 % chance that the actual number is above 70, and a 50–50 chance that it is below 40.

The true value was 50. This is not a suprising value relative to this assessment. If the value were 8 this would be surprising, as would 80. In each case, the realisation would be outside the 90 % confidence band.

An expert's probabilistic assessments are statistically accurate if 10 % of the realisations fall outside the 90 % confidence band; 50 % of the realisations fall on either side of the median (50 % value).

If your assessments had been:

5 %_5__________________ 95 %__90___________ 50 %_60___________

you would have been equally unsurprised, but your assessments would be less informative.

To get a feeling for this format, please complete the following assessments:

A
B
C

What is a good probability assessor?
A good probability assessor is one whose assessments, taken together, show good statistical accuracy, and are informative. Of these two, statistical accuracy is more important; informativeness is important to discriminate between statistically accurate assessments. "Little knowledge" should translate into wide uncertainty bands, and that in itself is valuable information which must be propagated through the model.

It is essential for the credibility of the results that the combined expert judgements display good statistical accuracy and high informativeness. For this reason, we will ask you to assess items whose true values will become known within the time frame of the study (seed variables).

Seed variables and variables of interest
Please reason your opinion for all questions/variables.

Seed variables
Consider a flock of broiler chickens in 1995 just before it is prepared for transportation to the processing plant. The flock became colonised with *Campylobacter* during rearing and all birds are carrying the organism both internally and externally. A random broiler chicken is sampled from this flock and the numbers of *Campylobacter* organisms in this chicken, both in its caecum (internal carriage) and on its exterior (external contamination), is enumerated. Following slaughter at the farm, the carcase is sampled directly after bleeding by the carcase rinse method. The number of *Campylobacter* organisms is determined by a standard method (i.e. plating on charcoal cefoperazone desoxycholate agar).

> **Seed variable S1**. The number of *Campylobacter* organisms (CFU/g) that is found in the caecal content of the broiler chicken just before it would have been transferred to a transport crate.
>
> To express the uncertainty associated with the number of Campylobacter organisms, please provide the 5th, 95th, and 50th percentiles of your estimate.
>
> **Seed variable S2**. The number of *Campylobacter* organisms (CFU per carcase) that is found on the exterior of the broiler chicken just before it would have been transferred to a transport crate.
>
> Please give a credible interval (your 5—95 % confidence) which you judge should encompass the number of *Campylobacter* organisms, and your central estimate (median) for that value (i.e. to locate the value within your credible range).

The flock is transported (14 broiler chickens per crate) for three to four hours. On arrival at the processing plant, the levels of internal and external carriage are determined again by the above-mentioned method.

Seed variable S3. The number of *Campylobacter* organisms (CFU/g) that are found in the caecal content of the broiler chicken after transport for three to four hours.

To express the uncertainty associated with the number of Campylobacter organisms, please provide the 5th, 95th, and 50th percentiles of your estimate.

Seed variable S4.The number of *Campylobacter* organisms (CFU per carcase) that are found on the exterior of the broiler chicken after transport for three to four hours.

To express the uncertainty associated with the number of Campylobacter organisms, please provide the 5th, 95th, and 50th percentiles of your estimate.

On arrival, faecal samples are obtained from the crates used to transport the broiler chickens and analysed for *Campylobacter*. The number of *Campylobacter* organisms is determined by the standard method.

Seed variable S5.The number of *Campylobacter* organisms (CFU/g) that is found in faecal samples from the crates after transporting the broiler chickens for three to four hours.

To express the uncertainty associated with the number of Campylobacter organisms, please provide the 5th, 95th, and 50th percentiles of your estimate.

Variable of interest

VI. Assume that a contaminated consignment of raw poultry is entering Europe through border control point BCP. The level of contamination with *Campylobacter* at the time of crossing the border is 100 CFU/g. The consignment will subsequently be transported and handled in countries A, B and C (in Europe) under usual conditions.

Consider one random journey. What level of contamination (in CFU/carcase) would be in the consignment when it reaches the end user of that journey?

To express the uncertainty associated with the level of contamination, please provide the 5th, 95th, and 50th percentiles of your estimate.

Answers to practice questions:
…

Graphs, tables, and other common reference material

6.3. Delphi method

Survey methods are ideal when a quick response is needed, opting for a one-off questionnaire when little but crucial information needs to be obtained. The Delphi method is essentially an iterative survey that has the advantage of providing feedback from the involved experts over successive rounds, enabling a degree of highly restricted *expert interaction*, and providing an opportunity for consensus to emerge as experts review their opinions in the face of novel information from their peers. Anonymity of experts is a specific feature of Delphi, as the technique is intended to reduce the social and political pressures to accept judgements that can arise in interacting groups; by removing identifying information from feedback, it is supposed that experts can/will concentrate on the merits of the feedback information itself without being influenced by potentially irrelevant cues. When using a survey, the group output is obtained by taking an equal weighting of the different judgments. When using Delphi, the (usually equal-weighted) judgements are aggregated from the responses of experts on the 'final' round. Evidence suggests that the *accuracy* of judgements (on judgement and short-term forecasting tasks—albeit often from studies using students rather than experts as subjects) does tend to increase over rounds (Rowe and Wright, 1999), and that this tends to occur because 'less expert' participants tend to change their judgements to an greater extent than more expert subjects, which has a corresponding tendency to shift the averaged judgement towards the 'true' answer on successive rounds. For further discussion of the Delphi method, see appendix A, section 4.4.

Both survey methods require the same initial steps in execution, but differ in content and length of process steps (the iterative Delphi requires more steps).

This section will include examples of how to execute a Delphi survey as it is assumed that this type of survey will be carried out more often. Steps for conducting a Delphi are shown in table 5. Suggestions for preparing a one-off questionnaire are compiled at the end of this section.

Table 5: Steps for executing a Delphi survey

	Step	
Preparation	1	Choose survey media
	2	Develop the survey
	2a	Write an introduction to the survey
	2b	List all questions that need to be answered
	2c	Write a closure to your survey
	3	Pilot survey
Timeline	4	Estimated timeline for expert involvement
Execution	5a	Training on probabilistic judgements
	5b	Send out survey
Analysis	6	Collect results and analysis
Subsequent Delphi round	Repeat steps	Subsequent Delphi rounds: repeat steps
	2a	Develop the survey (including collate answers and design feedback)
	3	Pilot survey
	5	Send out survey
	6	Collect results and analysis
		Final data collection and analysis

Participants

Within this Delphi protocol the participant list is already developed. In case there are experts who have a conflict of interest (or vested interests) in the survey topic, it is suggested that this information is provided at the onset of the Delphi to ensure transparency. This can be done by stating that due to the nature of the topic a variety of experts is included, including those with possible vested interest in the topic beginning. If clear vested interests are shown in the judgements, and it is of relevance to the Delphi, the elicitor may choose to label those as such (note, these responses do need to be shown!).

Within the Delphi study itself experts are kept anonymous. It should also be made clear at the onset of the Delphi study to what extent the experts will be named at the end of the study (i.e. reports). Shall experts' names and institutions be provided, or solely the latter? Experts need to provide consent in using the data you wish to include, this can be included in the Delphi survey itself.

Basis of probabilistic judgements as adopted within this Delphi protocol

The Delphi method combines behavioural aggregation with mathematical aggregation. This protocol adopts the *quartile method* in collecting the judgements, as shown in the Sheffield protocol. If desired this Delphi protocol can be adapted towards the 5^{th} and 95^{th} percentiles methods as used in the Cooke protocol.

6.3.1. Step 1. Choose survey media

Before you create the survey, you need to decide how the survey will be implemented. The most popular options nowadays are:

- using a web-based survey tool (e.g. Survey Monkey, Delphi decision aid: http://armstrong.wharton.upenn.edu/delphi2/ (freeware)) or
- via email (document attached or in the body of the email).

Web-based survey tools automatically collect your responses, which makes it easier for you to start drawing results. Keep in mind that programming the questionnaire may take some time—though how long will vary depending on the survey, the survey tool and the amount of programming experience you have.

Executing the questionnaire using *email with a document attachment* is preferred if respondents have limited access to the internet, as this allows the survey to be filled in off-line.

6.3.2. Step 2. Develop the survey

Survey methods require a questionnaire, which basically contains three parts:

- introduction on the survey process (time, topic, training reminder);
- elicitation of parameters (survey questions);
- closure: thank expert for participation and provide details on next steps.

It is suggested that online individual *training sessions on probability distributions* are run in advance of the Delphi survey. For details how to do this, please refer to section 5.2. For the hypothetical example, it is assumed that this training session is completed and only a reminder about this training is needed.

Delphi makes use of multiple (sequential) questionnaires (called 'rounds', usually two or three rounds). When listing the questions, it may become apparent that other details are needed before some parameters can be estimated; clearly these details need to be elicited first, which means that the estimation of this parameter needs to be shifted to a later round. However, try to execute your Delphi in about three rounds, as otherwise the process may become too lengthy and you may lose expert participant commitment.

Step 2a. Write an introduction to the survey

Include the following items in your introduction:

Delphi
Round 1
• Topic of Delphi; explain why their expertise is needed (in brief, as this information is also included in invitation letter) • Explain what will be done in this first Delphi round • Provide the amount of time needed to fill in the questionnaire* • Explain how to fill in the questionnaire (for example, if digitally, how to scroll through the pages; or, if on paper, use black or blue pen)
• If training is needed you can include a reminder of the previous training session in the first section of your Delphi
Round 2 (or any subsequent round)
• Short summary of what was done in last round • What will be done in this Delphi round. • Provide the amount of time needed to fill in the questionnaire* • Explain how to fill in the questionnaire

*Obtain timing in step 3.

Make sure that you remind participants to provide responses in their own capacity. This is especially of importance in case a Working Group member is part of the Delphi, the expert should provide judgements reflecting their own beliefs, which may not necessarily be the belief of the WG.

Below you find an example introduction text for a Delphi survey and a short text to remind the expert of the earlier training session.

The hypothetical example: introductory text for Delphi survey

Dear participant

Welcome to the online Delphi questionnaire on a non-toxigenic bacterial pathogen risk assessment! Many thanks for your willingness to participate in our research. The purpose of this survey is to elicit your knowledge to determine the necessary detection level of the non-toxicogenic bacterial pathogen at border control. Before answering the questions, you will receive some more background information about this specific situation, as well as a summary of the training session to refresh your memory on the next page.

[explanation: filling in survey]

Please note that, when we provide feedback in subsequent surveys, individual comments will not be identifiable. All results will be anonymised.

As already mentioned in the request letter, we estimate it will take approximately [XX] minutes to [XX] minutes to answer the questions.

If you have any difficulties with this questionnaire, or if you would like to have some extra information about the survey, please contact the survey team at the following email address: [XXX]

Kind regards [on behalf of the Elicitation Group]

[The elicitor]

[contact details]

The hypothetical example: reminder on training session

Recall that, in the training session that you took part in, the concept of a probability distribution was introduced. [provide main details of training session]

In this survey we wish to collect judgments on probability distributions, and the questions will generally ask: 'state 10 % ...; 90 % ...'. Below is a worked example showing what we would like you to consider. [short example]

Now, before going to the questionnaire, please fill in this example question. [something like a seed question, for example ask to write the quantities expected]

Thank you, now progress onto the questionnaire by [insert how to go there].

Step 2b. Elicitation of parameters (survey questions).

In this step you turn the list of all your parameters and other issues you need to elicit (which was created earlier) into questions. In our example we need to elicit the following (see section 3.2 for full details):

The hypothetical example: elicitation question

Given that the contamination at entry is 100 CFU/g, what level of contamination (in CFU/g) would be in the consignment, when it has reached the end user?

These items are listed in the table below and, in the case of the Delphi method, you can see how the items are divided over the two rounds.

The hypothetical example: elicitation overview

Delphi
Round 1
• *Total number of bacteria at end user (in CFU/g) for a single randomly selected journey for which the contamination at entry was 100 CFU/g*
Round 2
• *Feedback on round 1 outcome (e.g. present results/use results as basis for new questions)**   • *Invite participants to provide feedback on the outcome of round 1*

*Type of feedback possible: averages, estimates, rationales; presented in, for example, tables, or graphically. Rationales may be presented as received from experts or combined per topic (where relevant).

Collecting rationales. It is important to collect rationales (i.e. ask the participant to provide a reason for his or her response) so that the opinions and outcomes of the Delphi are understood, as well as a means to provide transparency of the process. Ideally, this should be done all the time, but it may not always be logistically feasible. With relatively small expert samples and / or short questionnaires all questions should be accompanied with the option to provide a rationale. For large expert samples or in case there are many questions, this may be (logistically) too much to handle. In such cases it is recommended to collect rationales for a set of questions. Always include at the end of the questionnaire an option to provide other comments (that are of importance to the Delphi, but were not included in the questionnaire), to allow for feedback and identify gaps in the study.

The hypothetical example: Delphi questions, round 1

Before we can start the estimation, some definitions need to be presented in order to ensure that all participants start at the same point.

This elicitation aims to define C, the contamination at end user given that contamination at entry is 100 CFU/g.

The following items need to be taken into account:

- *C is defined as the contamination with the bacteria at end user (in CFU/g) after a single randomly selected journey when the contamination level at entry was 100 CFU/g.*
- *The end user is defined as the place to which the food is delivered (without additional processing or repackaging) after entry to the EU. This may be to the processing industry or to distributors, for instance.*
- *A single journey is considered.*
- *Your uncertainty about C includes uncertainty about a typical or average net growth as well as the natural variability from one journey to another.*

Question 1: additional information

Currently the model that is used within this risk assessment is defined as follows:

$$C_{end\ user} = C_{entry} \times R_{transport}$$

where

$C_{end\ user}$ *is the contamination with the bacteria at end user (in CFU/g)*

C_{entry} *is the contamination with the bacteria at border (in CFU/g) and*

$R_{transport}$ *is the growing/survival/inactivation rate during transport and storage (dimensionless).*

The Working Group has concluded that conditions during transportation could be especially favourable for the growth of the pathogen, but no specific information on conditions during transport are known to refine the model or estimate the global growing rate $R_{transport}$.

Do you have any additional information on any of these parameters? [open response]

Question 2: define upper and lower limits for C.

Question 2a. Please suggest a lower limit for C (in CFU/g). In addition, please provide a rationale for your suggestion.

Question 2b. Please suggest a upper limit for C (in CFU/g). In addition, please provide a rationale for your suggestion.

Question 3: define median for C.

Question 3a. Please provide below a value what you consider might be a typical value for C (in CFU/g). In addition, please provide a rationale for your suggestion.

Question 3b. Please now provide the median value (M, in CFU/g) by adjusting your 'typical' value until you are comfortable that C is equally likely to take a value above M as below M. In addition, please provide a rationale for your suggestion.

Question 4: define upper and lower quartiles for C.

Question 4a. Please suggest a lower quartile (Q1) for C (in CFU/g). In addition, please provide a rationale for your suggestion.

Question 4b. Please suggest a upper quartile (Q3) for C (in CFU/g). In addition, please provide a rationale for your suggestion.

Step 2c. Write a closure to your survey.

It is very important to close your survey by thanking the experts for participation and provide details on (possible) next steps.

Delphi
Round 1 (or subsequent round)
• Thank the expert for participating • Say when to expect next Delphi round • Explain how to submit the Delphi survey
Round 2 (final round)
• Thank the expert for participating • Do you need any more input from expert?* • Explain how to submit the Delphi survey

*Note that if this was the final Delphi round, results will be calculated. Then it needs to be said what will be expected from expert (e.g. report to experts for feedback).

The hypothetical example: closure text, Delphi round 1 (or other non-final round)

Thank you for your participation!

Please note, by pressing "Finish" you finalise the survey, which is then sent to us automatically.

All responses will be aggregated and analysed, at which time a new questionnaire will be developed. You will receive an invitation to participate in about [XX] weeks.

6.3.3. Step 3. Pilot survey

By testing the survey (i.e. to pilot) with several people (about three to five), you can check that the survey is understandable and does not contain any mistakes and how long it takes to respond to the survey. To ensure an independent pilot, you need to ask people who were not involved in the development of the survey to pilot the survey (for example, a colleague or work group members). As this is a pilot, it is not important whether or not the pilot participants know the exact answer to the questions (but they should understand the topic).

To pilot your survey:

- send your pilot participants the letter requesting that they take part in the survey, including the (link to) survey;
- in addition, ask your pilot participants to time their effort, and to look for typing errors (etc.).

The survey will be amended slightly by adding some items. Some suggestions are provided below.

The hypothetical example: example pilot items

****INSERT at the beginning of the survey:***

Dear pilot participant

Please keep in mind that you will need to fill in at the end of the questionnaire how long it takes you to complete the survey; this includes reading the introduction to the survey.

Thank you very much for your participation.

INSERT the following items at the end of the survey (and provide space for answering):

Regarding the timing of the questionnaire:

1a. How long did you take to complete your survey?

1b. Were you able to complete the questionnaire in one session? [Yes/No]

1c. If not, can you explain why?

2. Regarding the questionnaire, are the questions easy to understand?

3. Regarding the use of jargon. Whenever there was jargon used within the text, was this sufficiently explained?

This expert elicitation aims to [provide survey topic].

4. Do you think we missed out on a topic within this survey?

5. Do you think the way to elicit probability assessment [XXX] is appropriately asked?

6. Do you have any other comments regarding this survey?

Responses to pilot. Once you have received your pilot responses you can check these on the following issues:

- comments on survey, invitation letter;
- comments on pilot questions;
- do test analysis with responses (if possible and desirable).

Adapt survey. Depending on the feedback received, amend the survey and invitation letter as appropriate. The pilot has provided as well the timing for the survey, which can be inserted on the first page of the Delphi questionnaire as well as included in the request letter.

The survey is now ready to be sent out!

6.3.4. Step 4. Estimated timeline for expert involvement

The timeline for expert participation (Table 6) using a survey method is detailed below. In general, it takes about five to ten weeks to complete a single Delphi round (Table 7): two to three weeks (full time) to prepare a Delphi survey (up to sending it out), two to four weeks for the experts to complete and return the survey and about one week to analyse the (final) results.

Table 6: The estimated timeline focused on expert participation (per Delphi round)

Step	Estimation of time needed
Expert training on probabilistic judgements	1 day
Survey out with expert participants	2–4 weeks
Send out participant reminder for survey	1 day
Closure of survey and data collation	1 day–2 weeks

With the Delphi method you need to remember that it is a repetitive process. The amount of repetition of steps that need to be executed depends on the number of survey rounds, which provides you with the following timeline. For the Delphi multiple (short) surveys are needed, for which you repeat the items in step 2. With our example these have been divided into Round 1 and Round 2. See full timeline for a Delphi study below, a timeline for executing a questionnaire is placed at the end of this section.

Please note: Especially the preparation of subsequent Delphi rounds may go faster than suggested here, however this all depends on the amount of information needed to elicit and the amount of resources available to prepare the next survey.

Table 7: The estimated timeline

Step	Estimation of time needed
Set estimated timeline	1 day
Delphi round 1	**About 5–10 weeks**
Survey development	1–2 weeks
Pilot of survey	1 week
Expert training on probabilistic judgements	1 day
Send out survey	1 day
Survey out with expert participants	2–4 weeks
Send out participant reminder for survey	1 day
Closure of survey and data collation	1 day–2 weeks
Data analysis	1 week
Delphi subsequent round	**About 5–10 weeks**
Survey development (including feedback)	1–2 weeks
Pilot of survey	1 day–1 week
send out survey	1 day
Survey out with expert participants	2–4 weeks
Send out participant reminder for survey	1 day
Closure of survey and data collation	1 day–2 weeks
Data analysis	1 week
After final Delphi round	
Combine data analysis of all rounds	**1 week**

6.3.5. Step 5. Execute Delphi survey

Step 5a: Expert training on probabilistic judgements

In advance of the Delphi survey, experts need to be trained on giving probabilistic judgements. See section 5.2 for details.

Step 5b: Send out survey

In order to send out the survey you need the following items, which are described in more detail below:

- participant list
- request letter
- Delphi survey

Survey with experts. Usually the survey is out with expert participants for about two to four weeks. How long you allow for response may vary depend on the survey (and how quickly you need the results), but it is suggested that you give your experts at least two weeks to respond.

Participant list. Previously you have collated a list of participants, it is now time to use that list and send out the survey to all expert participants.

The hypothetical example: participant list

Name	***Expertise***	***Institution***	***Contact details***
Dr Schwarz	*Food transport conditions*	*Logistics scheduler, Agro-Transit*	*Schwarz@agrotransit.eu*
Professor Prugna	*Safety of food distribution*	*Public health school, University of Transeuropia*	*prugna@universitytransitopia.eu*
Mr Mustár	*Determine food importation admittance at border entry points to the EC*	*Food inspector, border control,European Community*	*mustar@bordercontrol.eu*
Mrs Pauw	*Experience of measuring concentrations of bacteria in food*	*Inspector, ReadyMealsRUs*	*pauw@readymealsrus.eu*

In case there is a large variation in participating experts (due to large variation in input needed), one may provide an option to only respond to those they feel expertise on. If not all questions are answered, it is strongly suggested to check the recruitment again, as apparently not the right experts were included in the survey (expertise gap). Alternatively, if topics vary too much different survey versions may be developed specific to expertise.

Request letter. Although you have contacted your expert participant already, you need an request letter to accompany your survey. In this request letter you need to detail in brief:

- the "why" of the survey: what is this about and why is it important;
- what you expect from the expert
    - respond to survey (questionnaire, Delphi round 1/2/X)

  - how shall this survey be executed (online, using email, etc.)
  - before what date the response needs to be in (e.g. three weeks from now being [DATE XX])
  - how long it takes to respond (e.g. [15] minutes)
- include the survey (e.g. weblink to survey or attached as document);
- thank the expert for participating (provide incentive if relevant);
- explain how to contact you in case of queries and provide contact details.

Sending out the survey: make sure you include the survey or a link to the survey!

Reminder. If responses are low you may wish to remind the experts to respond to the survey, for example one week before the deadline (possibly at a later stage again). Make sure that you remind only those who did not respond. If you have a small group of experts, phoning them will probably be a better option to increase your response rate. Make sure you have the following items prepared, so that you do not forget to mention something. If you need to send an email after your conversation (e.g. to re-send the survey), it will be useful to include these items as well. Alternatively, if you choose to send only a reminder email, make sure it is short but complete! The following items are suggested:

- introduce yourself;
- explain this is a reminder for survey [NAME];
- repeat deadline [DATE];
- provide link to survey;
- thank the expert for participating (provide incentive if relevant);
- note how to contact you in case of queries.

6.3.6. Step 6. Collection of results and analysis

Data collection

After closing your survey, it is time to collect all data into one dataset and analyse the results. If you use an automated survey system, data collation will be easy. If you need to manually collect all data and insert into one datasheet, this may take some time (depending on the length of the survey and the number of participants). Keep this in mind when estimating your timetable.

Rationales. These responses can be used to feed back in the next round (ask question again preceded with feedback to be considered before answering) or can be used to develop a new question (e.g. respondents have provided the following options, do you agree/disagree?). In this hypothetical example the reasoning is limited, but in real elicitations the experts should be encouraged to give full explanations, which should be reported. When short responses are provided by the experts, make sure that in a subsequent round experts are reminded to provide full explanations. This can be done within the questionnaire itself, as more personally within the request letter (or both).

The hypothetical example: collated raw data

The data received from all participants are summarised in the table below.

Expert	***Question***					
	2a	***2b***	***3a***	***3b***	***4a***	***4b***
				M	***Q1***	***Q3***
1	*10*	*5 000*	*400*	*200*	*100*	*400*
2	*25*	*3 000*	*500*	*350*	*100*	*2 000*
3	*0*	*1 000*	*95*	*100*	*20*	*500*
4	*0*	*5000*	*650*	*500*	*125*	*1 500*

Responses to open questions. [remember: Delphi participants are anonymous!]

Question 1: additional information

Do you have any additional information on any of these parameters? [open response]

Expert 1: No data available.

Expert 2: No additional information on transport data, all other info has been provided to your working group by my colleague earlier.

Expert 3: None.

Expert 4: See my email on transport data.

Question 2: define upper and lower limits for C

Question 2a. Please suggest a lower limit for C (in CFU/g). In addition, please provide a rationale for your suggestion.

Expert 1: Only a few are expected.

Expert 2: I always detect small amounts.

Expert 3: None.

Expert 4: Often no bacteria are detected.

Question 2b. Please suggest a upper limit for C (in CFU/g). In addition, please provide a rationale for your suggestion.

Expert 1: I have checked our logs and found this extreme value.

Expert 2: Commonly found

Expert 3: --

Expert 4: Often high values are observed.

Question 3: define median for C

Question 3a: Please provide below a value what you consider might be a typical value for C (in CFU/g). In addition, please provide a rationale for your suggestion.

Expert 1: Many samples have this value.

Expert 2: Frequently detected.

Expert 3: –

Expert 4: I think this is the average.

Question 3b: Please provide now the median value (M, in CFU/g) by adjusting your 'typical' value until you are comfortable that C is equally likely to take a value above M as below M. In addition, please provide a rationale for your suggestion.

Expert 1: I think it is lower than the above figure I gave.

Expert 2: I guess.

Expert 3: –

Expert 4: As above.

Question 4: define upper and lower quartiles for C.

Question 4a. Please suggest a lower quartile (Q1) for C (in CFU/g). In addition, please provide a rationale for your suggestion.

Expert 1: Looking at my previous answers this should be right.

Expert 2: This is what my dataset provides me with.

Expert 3: –

Expert 4: This is the average lower amount I come across.

Question 4b. Please suggest a upper quartile (Q3) for C (in CFU/g). In addition, please provide a rationale for your suggestion.

Expert 1: Idem.

Expert 2: From dataset

Expert 3: –

Expert 4: This is the average high amount I come across

Data analysis

Once you have all data in one spreadsheet, data analysis can be started. The time needed to conduct the analysis may vary on the length of your survey and the number of responses received, as well as the difficulty of the variables needed to investigate. It is suggested to equally-weight the Delphi responses.

Rationales. These may be provided just as you received them (in a list, random order) or collated per topic (listed per topic or tabulated).

The hypothetical example: data analysis

Expert	*Question*		
	M	***Q1***	***Q3***
1	*200*	*100*	*400*
2	*350*	*100*	*2 000*
3	*100*	*20*	*500*
4	*500*	*125*	*1500*
Range	*100–500*	*20–125*	*400–2 000*

6.3.7. Step 7. Subsequent Delphi rounds: repeat steps

The results from one round will be used to develop a subsequent questionnaire (possibly in combination with questions you already allotted to this Delphi round). In order to develop the new survey, repeat steps 2, 3, 5,and 6 until the Delphi process is completed. It is assumed that you have already dealt with step 4 (estimated timeline for expert involvement).

Step 7.1: Choose survey (not applicable here)

Step 7.2: Development of the survey

The aim of second/subsequent Delphi survey round is to ‘respond’ to the previous round and take it further. In general you can say that such a Delphi round contains the following items:

- feedback on the results of the previous round;
- questions.

Feedback. Examples include averages, estimates, rationales, which may be presented, for example, in tables, or graphically. Rationales (responses to open questions) may be presented as received from experts or combined per topic (where relevant). Your choice will depend on the specific situation, i.e. many similar responses or mostly different responses, few or large group of expert participants/responses.

Questions. The questions in subsequent Delphi rounds continue where the last round ended. Most often, the question is asked again with feedback/responses to consider provided in advance (the process is repeated until no changes in responses are observed).

The order in which the feedback and the questions are provided can be as follows:

- all feedback, all questions;
- feedback question 1, question 1, feedback question 2, question 2, etc.

Both options have their merits and drawbacks; in the case of larger questionnaires, the second option is recommended.

Look again at step 2 for details on other issues related to development of the survey; the introduction and closure of the survey are not repeated here.

The hypothetical example: Delphi questions, round 2

[NOTE in this example only feedback is provided to questions that will be continued in this second round.]

In the first round you have made some estimations to estimate C (the contamination at end user given that contamination at entry is 100 CFU/g; for more details please see the textbox below). In the table below you can find the responses of all experts plus the response ranges.

Expert	***Parameter M***	***Rationales***
1	*200*	*I think it is lower than the above figure I gave.*
2	*350*	*I guess*
3	*100*	*I think this is the average*
4	*500*	
Range	*100–500*	

Expert	***Parameter Q1***	***Rationales***
1	*100*	*Looking at my previous answers this should be right.*
2	*100*	*This is what my dataset provides me with*
3	*20*	*This is the average lower amount I come across*
4	*125*	
Range	*20–125*	

Expert	***Parameter Q3***	***Rationales***
1	*400*	*Looking at my previous answers this should be right*
2	*2 000*	*From dataset*
3	*500*	*This is the average high amount I come across*
4	*1 500*	
Range	*400–2 000*	

Please consider again the parameter details as provided in the previous survey round.

This elicitation aims to define C, the contamination at end user given that contamination at entry is 100 CFU/g. The following items need to be taken into account:

- *C is defined as the contamination with the bacteria at end user (in CFU/g) after a single randomly selected journey when the contamination at entry was 100 CFU/g.*
- *The nature of the end user is defined as the place to which the food is delivered (without additional processing or repackaging) after entry to the EU. This may be to processing industry or to distributors, for instance.*
- *A single journey is considered.*
- *Your uncertainty about C includes uncertainty about a typical or average net growth as well as the natural variability from one journey to another.*

Within this second round we would like you to respond to some of the first round questions again now that you have considered the feedback from all other expert participants. Please answer the following questions.

Question 1: define median for C

Please provide the median value (M) (in CFU/g), for which you are comfortable that C is equally likely to take a value above M as below M.

Question 2: define upper and lower quartiles for C

Question 2a. Please suggest a lower quartile (Q1) for C (in CFU/g).

Question 2b. Please suggest a upper quartile (Q3) for C (in CFU/g).

Step 7.3: Pilot survey

Please note the following thoughts on repeating the pilot survey. Most likely it is unnecessary to repeat the full pilot, assuming that your questionnaire has not changed much between rounds (only added respondent feedback to the questions). In such cases a shortened pilot is advised: let somebody who did not develop the questionnaire look at it (check for typing errors, feedback correctly inserted, etc.) and make corrections where needed. If the Delphi questionnaire did change dramatically, there is a clear advantage in re-piloting the subsequent questionnaire.

Step 7.4: Estimated timeline for expert involvement (not applicable here)

Step 7.5: Send out survey

This should be executed as described in step 5 above.

It is good practice to ask experts to complete an appraisal form at the end of the expert elicitation, in which they are asked for their comments and evaluations about the planning, organisation and running of the workshop. This may be sent to the expert once the final Delphi round is completed. For details see section 7.1.

Step 7.6. Collect results and analysis

This should be executed as described in step 6 above.

The hypothetical example: collated raw data Delphi round 2

The data received from all participants are summarised in the table below.

Expert	***Question***		
	1	***2a***	***3b***
	M	***Q1***	***Q3***
1	*200*	*100*	*400*
2	*350*	*100*	*2 000*
3	*100*	*20*	*500*
4	*500*	*125*	*1 500*

6.3.7.1. Final data collection and analysis

If no more Delphi rounds are needed, as all information is gathered, a final analysis is needed. In this final analysis, the data from the last round are analysed and combined (where relevant) with previous Delphi rounds.

If it is thought that the probability distributions for the parameters (quantiles) over the various Delphi rounds will be redefined, with the assumption that expert responses move to the median. After the final Delphi round the data will be aggregated and an equal weighting of distributions will be calculated. If it happens that two groups of responses emerge, it should be concluded that follow-up actions are needed. Either another Delphi round should be executed or the outcomes should be fed back to the Working Group.

When preparing the final report, a small amount of editing of expert responses is allowed. Though these responses should be fed back to the expert to check whether they agree with the made changes.

6.3.8. Post-elicitation reporting

To create a formal record of the Delphi elicitation, the following items need to be included in the Technical support Document (included either in report itself or in an appendix):

- pilot questionnaire round 1
- pilot participant list
- outcomes pilot questionnaire round 1 ('raw data')
- written version of "Expert training on probabilistic judgements"
- participant list
- request letter
- round 1 questionnaire
- round 1 responses ('raw data')
- round 1 analysis
- round 2 questionnaire
- round 2 responses ('raw data')
- round 2 analysis
- subsequent round questionnaire(s) (if applicable)
- subsequent round responses ('raw data') (if applicable)
- subsequent round analysis (if applicable)
- final data analysis
- expert appraisal form
- outcomes expert appraisal.

For further details on the Technical Support Document, see part II, section 7.2.

6.4. Deliverables

In addition to the Technical Documentation of the elicitation process the Elicitation Group has to summarise the elicitation results in two different reports.

1. The result report is targeted to the risk assessors to answer the question of interest. It should be a stand-alone documentation for publication in parallel (e.g. as an annex) with the risk assessment.

2. The expert feedback which will be provided to the individual experts on request to fulfil the obligation of transparency against the expert.

3. Additional to the expert feedback, an evaluation of the elicitation process should be requested.

The result report and documentation of the elicitation phase will be reviewed by the Steering Group, which should confirm that the elicitation was completed as planned and that the results are valid for further use.

7. Post-elicitation

The post-elicitation phase is the critical appraisal of the whole elicitation process and the final decision on the use of the results.

While the final documentation of the elicitation process will be done by the Steering Group, normally by summarising the results of the different milestones, the Working Group reviews the whole process regarding to the needs of the risk assessment procedure. A description of the final use will be added to the final documentation and sent to the expert panel.

7.1. Technical documentation

The technical documentation consists of the results and evaluations of the three milestones:

- Problem definition of the initiation phase
    - constitution of the Working Group (responsible authors)
    - background report including the risk assessment model
    - existing information on the parameter of interest
    - justification and necessary conditions for EKE
    - evaluation by the corresponding panel and EFSA administration
- Elicitation protocol of the pre-elicitation phase
    - constitution of the Steering Group (responsible authors)
    - the final elicitation question
    - description of the expert selection procedure
    - the decision on the elicitation method including the selection of the Elicitation Group
    - the final project plan for elicitation
    - external review, if applicable
- Result report of the elicitation phase
    - constitution of the Elicitation Group (responsible authors)

- documentation of the background material and training sessions given to the expert panel
- documentation of the elicitation process, including the time line, the questions, the expert panel, and the methods to gather and analyse the answers
- documentation on the data analysis
- anonymised version of expert rationales for their judgements
- results for use in risk assessment
- discussion of assumptions, qualitative uncertainties and constraints of the result
- any complaints regarding the result, if declaimed by a participant
- evaluation of the process and the results by the Steering and Working Group

7.2. Summary on the use of the elicitation results

The final step of the elicitation process is the description of the use of the elicitation results in the risk assessment.

The corresponding panel or EFSA administration thanks all participants involved in the elicitation process for their input.

CONCLUSIONS AND RECOMMENDATIONS

8. Implementation of EKE in EFSA

According to the mandate one task of the Guidance was to give practical advice on how to conduct an expert knowledge elicitation in the context of EFSA's risk assessments. The working group considered this by incorporating three concrete protocols into the Guidance.

Nevertheless the working group has to notice that a written Guidance alone is not sufficient to put a new methodology into the practice of an institution. Additionally the protocols should be seen as possible solutions for most problems, and should be used as starting point, but can be adapted case-by-case to fit specific situations better.

Therefore the working group offers some recommendations on how expert knowledge elicitation could be implemented into EFSA's daily work.

8.1. Training and enabling

The first and most important step to establish expert knowledge elicitation within EFSA is to provide an appropriate training, material, and tools, for those who will be involved in the process, to guarantee a consistent understanding and implementation of the key concepts and processes. This includes a large number of experts, who will be elicited, but also all EFSA staff and panel members and their working groups.

Different types of training are needed for the different roles in the process.

Training on probabilistic judgements

All people involved in the process need training on making probability judgements. Furthermore all new EFSA staff and experts in panels and working groups should receive this training at the beginning of their term in EFSA, as well as existing staff in due course. This should guarantee that all quantitative judgements are done consistently in accordance with the Guidance.

Because this is a large and recurrent need, our recommendation is to commission the development of an interactive, self-paced online training module. Ideally this will be accompanied by written documentation and supporting material. The course should cover understanding concepts of probability as a judgement, basic rules of probability and probability distributions; the kinds of judgement needed in elicitations, common sources of biases and judgement errors, and should include practical exercises covering the whole range of EFSA topics. Finally the trainee's understanding should be tested, and successful completion of the course should be required for participation in the elicitations. Translations to languages other than English might be necessary.

Training on steering an expert knowledge elicitation

A second form of training for EFSA staff, members of panels and working groups is in the use and steering of expert knowledge elicitation exercises. This includes how to identify the need for EKE, prioritizing parameters for EKE, role of the Steering Group, e.g. framing questions, identifying and selecting experts, choosing the method, appointment of elicitors, and documentation of the process. The course will be based mainly on the Guidance document, and be enriched by practical group exercises to consolidate the ideas. A duration of 1.5-2 days is recommended.

The priority of attendance on the first runs of the course should be given to the network of EKE advocates (see below) and to staff and experts using quantitative models in risk assessments, but in due course all staff and panel members and working group chairs should take this training.

Elicitation tools

Ultimately there are two approaches EFSA can take to carry out EKE. The first is to commission external contractors with appropriate expertise in the chosen EKE methods; the second is to develop sufficient in-house expertise to conduct one or more of the EKE methods itself. We anticipate that the first will be favoured in the short term, however over time EFSA may prefer to establish in-house capability. In the latter case, the access to, and further training on, specific software tools will be needed. Although public domain software is already available for the Sheffield (SHELF) and Cooke (EXCALIBUR) protocols, there is a present lack of software to perform the Delphi protocol. We recommend that EFSA commission the adaptation of existing survey software to this specific purpose.

External elicitors

In case external elicitors should perform the protocols, EFSA should define clear requirements to select appropriate contractors. A checklist for each protocol might help to formulate the concrete tasks of the contracts and to verify their correct execution. The signing of a framework contract will help to shorten the time for starting external procurements.

8.2. Organisational structure

Best practice

This working group is firmly convinced that the formal procedures as set out in this Guidance will constitute best practice for EFSA. In order to facilitate the implementation of EKE, EFSA management can ensure that sufficient resources are allocated routinely for all mandates. This includes financial resources, staff allocation and appropriate timescales.

Contact point

Widespread and timely adoption of EKE will require a dedicated full time EKE specialist within EFSA, providing support to a network of contacts covering all the units. This expert should have technical knowledge and practical experience in EKE. A background in social science and statistics will be advantageous. Whilst this person may initially be skilled in just one protocol, it is important for them to understand the range of protocols and their contingent utilities, and being proficient in all three methods.

Their duties will include providing advice on EKE, both in general terms and practical hands-on support, for instance by membership of the Steering Group on controversial or sensitive topics; coordinating and performing training; supporting documentation and evaluating finalized EKE, generally acting as an ambassador for EKE, and promoting further developments to enlarge the applicability of EKE within EFSA.

In short term a network of EKE advocates from each unit)coordinated by the AMU unit) should be achivable as starting point and providing support to the EKE specialist, once appointed..

8.3. Expert identification and management

Expert identification

A fundamental challenge in EKE is the identification and recruitment of appropriate experts. EFSA does currently possess information on a variety of experts in its expert database. However this is of limited utility for EKE purposes, because the experts have not been systematically and actively identified, and the information contained in the database is insufficient for appropriately classifying the nature of their expertise.

The working group recommends that a more systematic approach is taken to expert identification, particularly in order to enlarge the pool of experts for EKE beyond academia. One approach would be to map what expertise resides in which institution on main topics in the remit of EFSA. This is in order to establish initial institutional contacts for forthcoming expert identifications. These institutional contacts can be used as starting point for snowballing, as described in the Guidance. Other approaches might also be considered, such as social network analysis.

Expert database and expert retention

In order to aid the identification of experts for EKE, a better, more comprehensive database is required. This can either be archived by amending the present database, or develop a new specific database for EKE purposes. The design of such a database should be a priority, incorporating questions related to substantive and normative issues identified in this Guidance.

This database should also include information on experts' past engagements with EFSA, contact information from institutions, and expert information from selection procedure. EFSA can use this information to retain experts for future EKE excercises.

8.4. Extensions of the EKE process

This Guidance does not address all the aspects of expert elicitation that are potentially of value to EFSA. Some aspects (e.g. dependencies and imprecision, see below) have been excluded by the mandate because they are more advanced topics that should not be employed until EFSA has acquired sufficient experience with the basic protocols. Others are not addressed here because they are areas of current research in the field. EFSA should continue to expand its EKE capability with a view to creating a transparent, harmonized and consistent set of EKE methods that extend best practice to all the situations arising in risk assessment.

The Guidance does not cover the important role that expert elicitation can play in qualitative judgements such as the development of an appropriate risk model. EFSA should identify best practice in the use of experts to provide fully-informed, transparent and documented judgements in these cases.

Very limited discussion is given here on the elicitation of expert knowledge regarding two or more uncertain quantities. When these are not judged to be independent, the risk assessment can be substantially enhanced by eliciting a suitable joint distribution that reflects expert beliefs about correlation. Some methods for multivariate elicitation currently exist and new techniques can be expected to emerge in the near future. EFSA should seek to expand its capability in this important but challenging area of elicitation.

The Guidance has identified three specific protocols in the expectation that EFSA will initially rely on these when EKE is performed. However, the extensive discussion in Appendix A can be used to explore variations on or alternatives to these basic protocols. EFSA should periodically review its EKE practices with a view to tailoring the protocols to the specific requirements of EFSA risk assessments.

Responsibility for promoting and managing these enhancements should rest in the first instance with the full-time EKE specialist identified in section 8.2.

REFERENCES

ACERA (Australian Centre of Excellence for Risk Analysis), 2009. Eliciting Expert Judgments, report #1: Literature Review. ACERA Project 0611, Melbourne, Australia,71 pp. Available online: http://www.acera.unimelb.edu.au/materials/endorsed/0611.pdf

ACERA (Australian Centre of Excellence for Risk Analysis), 2010. Process manual – Elicitation Tool. ACERA, Melbourne, Australia, 40 pp. Available online: http://www.acera.unimelb.edu.au/materials/endorsed/0611-process-manual.pdf

Algers B, Anil H, Blokhuis H, Fuchs K, Hultgren J, Lambooij B, Nunes T, Paulsen P, and Smulders F, 2009. Project to develop animal aelfare risk assessment guidelines on stunning and killing. Technical report submitted to EFSA, CFP/EFSA/AHAW/2007/01, 88 pp.

Anderson JR, 1982. Acquisition of cognitive skill. Psychological Review, 89, 369–406.

Asante-Duah K, 2002. Public health risk assessment for human exposure to chemicals. Environmental Pollution Vol. 6, Kluwer Academic Publishers, Dordrecht, the Netherlands.

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

Aspinall WP, and Cooke RM, 2013. Quantifying scientific uncertainty from expert judgement elicitation. In: Risk and Uncertainty Assessment for Natural Hazards. Eds. J Rougier, S Sparks, and HJ Hill, University Press, Cambridge, 2013.

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

Bastian M, Heymann S and Jacomy M, 2009. Gephi: An Open Source software for exploring and manipulating networks. Proceedings of the Third International ICWSM Conference.

Berry DC, Raynor DK, Knapp P and Bersellini E, 2003. Patients' understanding of risk associated with medication use – impact of European Commission guidelines and other risk scales. Drug Safety, 26(1), 1–11.

Bolger F and Wright G, 1994. Assessing the quality of expert judgment: Issues and analysis. Decision Support Systems, 11, 1–24.

Bolger F and Wright G, 2011. Improving the Delphi process: lessons from social psychological research. Technological Forecasting and Social Change, 78, 1500–1513.

Booker JM, and Meyer MA, 1988. Sources and effects of interexpert correlation: an empirical study. IEEE Transactions on Systems, Man & Cybernetics, 18, 135–142.

Borgatti SP, Everett MG and Freeman LC, 2002. Ucinet for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies.

Brenner LA, 2003. A random support model of the calibration of subjective probabilities. Organizational Behavior and Human Decision Processes, 90, 87–110.

Brown J, Ehrhardt J, Goossens LHJ, Cooke RM, Fischer F, Hasemann I, Jones JA, Kraan BCP, and Smith JG,, 2001. Probabilistic accident consequence uncertainty assessment using COSYMA: uncertainty from the food chain module. EC project report EUR-18823, 126 pp. Available online: ftp://ftp.cordis.europa.eu/pub/fp5-euratom/docs/eur18823_en.pdf.

Brown AJ and Aspinall WP, 2004. Use of expert opinion elicitation to quantify the internal erosion process in dams. Proceedings of the 13th Biennial British Dams Society Conference, University of Kent, Canterbury, UK, 22–26 June 2004, 16 pp.

Burgman MA, McBride M, Ashton R, Speirs-Bridge A, Flander L, Wintle B, Fidler F, Rumpff L, and Twardy C, 2011. Expert status and performance. PLoS ONE, 6(7), e22998. doi: 10.1371/journal.pone.0022998.

CAC (Codex Alimentarius Commission), 2011. Procedural Manual, 20th edn.. FAO, Rome, Italy, and WHO, Geneva, Switzerland.

Cheeseman P, 1985. In defense of probability. Proceedings of the Ninth International Joint Conference on Artificial Intelligence, 1002–1009.

Chen C, 2006. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for Information Science and Technology, 57(3), 359-377.

Clemen RT, 2008. Comment on Cooke's classical method. Reliability Engineering & System Safety, 93, 760–765.

Clemen RT and Winkler RL, 1999. Combining probability distributions from experts in risk analysis. Risk Analysis, 19, 187–203.

Cooke RM, 1991. Experts in uncertainty. Oxford University Press, Oxford, UK.

Cooke RM, 2008. Response to discussants. Reliability Engineering & System Safety, 93, 775–777.

Cooke RM, 2013. Validating expert judgment with the classical model. In: Experts and consensus in social science – critical perspectives from economics, sociology, politics, and philosophy. Eds Martini C and Boumans M. Springer, New York, USA, in preparation.

Cooke RM, 2013b. Deep and shallow uncertainty in messaging climate change. Proceedings of the 22nd European Safety and Reliability Conference "Safety, Reliability and Risk Analysis: Beyond the Horizon", Amsterdam, The Netherlands.

Cooke RM, ElSaadanay S, and Huang X, 2008. On the performance of social network and likelihood-based expert weighting schemes. Reliability Engineering & System Safety, 93, 745–756.

Cooke RM and Goossens LJH, 1999. Procedures guide for structured expert judgment. Project report EUR 18820EN. Luxembourg, Euratom, 50 pp.

Cooke RM and Goossens LJH, 2000. Procedures guide for structured expert judgment in accident consequence modelling. Radiation Protection Dosimetry 90, 303-309.

Cooke RM, and Goossens LJH, 2008. TU Delft Expert Judgment Data Base. Reliability Engineering & System Safety, 93, 657–674.

Cooke RM and Jager E, 1998 Failure frequency of underground gas pipelines: methods for assessment with structured expertjudgment. Risk Analysis, 18, 511–527.

Dalgleish LI, 1988. Decision making in child abuse cases: applications of social judgment theory and signal detection theory. In: Human judgment: the SJT view. Eds Brehmer B and Joyce CB. North Holland: Elsevier, Amsterdam, the Netherlands.

Dalkey NC, 1975. Toward a theory of group estimation. In: The Delphi method: techniques and applications. Eds Linstone HA and Turoff M. Addison-Wesley, Reading, MA, USA.

Daneshkhah A, and Oakley JE, 2010. Eliciting multivariate probability distributions. In: Rethinking risk measurement and reporting: Volume I. Ed Böcker K. Risk Books, London, UK.

DeGroot MH, 1970. Optimal statistical decisions. Reissued in Wiley Classics Library, Hoboken, USA.

de Nooy W, Mrvar A and & Batagelj V, 2005. Exploratory Social Network Analysis with Pajek. Cambridge University Press, Cambridge, UK.

Develee JLA and Knol AB, 2011. Software to support elicitation – An explonatory study of existing software packages. RIVM Letter Report 630003001/2011. National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands, 100 pp.

Edwards W, Miles RF Jr. and von Winterfeldt D, 2007. Advances in decision analysis: from foundations to applications. Cambridge University Press, Cambridge, UK, 638pp.

EFSA (European Food Safety Authority), 2010a. Application of systematic review methodology to food and feed safety assessments to support decision making. EFSA Journal 2010;8(6):1637, 90 pp. doi:10.2903/j.efsa.2010.1637

EFSA (European Food Safety Authority), 2010b. Report on data collection: future directions. EFSA Journal 2010; 8(5):1533, 35 pp. doi:10.2903/j.efsa.2010.1533

EFSA (European Food Safety Authority), 2012. Science strategy 2012-2016, 40 pp. doi:10.2805/96760

EFSA AHAW Panel (EFSA Panel on Animal Health and Welfare), 2009. Guidance on good practice in conducting scientific assessments in animal health using modelling. EFSA Journal 2009;7(12):1419, 38 pp.

EFSA AHAW Panel (EFSA Panel on Animal Health and Welfare), 2012. Guidance on risk assessment for animal welfare. EFSA Journal 2012;10(1):2513, 30 pp. doi:10.2903/j.efsa.2012.2513

EFSA ED (European Food Safety Authority, Executive Director), 2013. Decision of the Executive Director concerning the selection of members of the Scientific Committee, scientific panels and external experts to assist EFSA with its scientific work. Parma, 27th Feb. 2013.

EFSA MB (European Food Safety Authority—Management Board), 2005. Decision concerning implementing measures of transparency and confidentiality requirements. EFSA Document MB 10.03.2005–10.

EFSA PLH Panel (EFSA Panel on Plant Health), 2010a. Guidance on a harmonised framework for pest risk assessment and the identification and evaluation of pest risk management options by EFSA. EFSA Journal 2010;8(2):1495, 66 pp.

EFSA PLH Panel (EFSA Panel on Plant Health), 2010b. Scientific opinion on a quantitative pathway analysis of the likelihood of Tilletia indica M. introduction into EU with importation of US wheat. EFSA Journal 2010;8(6):1621, 88 pp.

EFSA PLH Panel (EFSA Panel on Plant Health), 2011. Guidance on the environmental risk assessment of plant pests. EFSA Journal 2011;9(12):2460, 121 pp.

EFSA SC (EFSA Scientific Committee), 2006a. Transparency in risk assessment carried out by EFSA: Guidance document on procedural aspects. EFSA Journal 2006, 353, 1-16.

EFSA SC (EFSA Scientific Committee), 2006b. Guidance of the Scientific Committee on a request from EFSA related to Uncertainties in Dietary Exposure Assessment. The EFSA Journal 2006, 438, 1-54.

EFSA SC (EFSA Scientific Committee), 2009. Guidance of the Scientific Committee on transparency in the scientific aspects of risk assessment carried out by EFSA. Part 2: general principles. The EFSA Journal 2009, 1051, 1-22.

EFSA SC (EFSA Scientific Committee), 2012. Scientific Opinion on risk assessment terminology. EFSA Journal 10(5):2664, 43 pp. doi:10.2903/j.efsa.2012.2664

Eggstaff JW, Mazzuchi TA and Sarkani S, 2014. The effect of the number of seed variables on the performance of Cooke′s classical model. Reliability Engineering & System Safety, 121, 72–82.

Ferrell WR, 1985. Combining individual judgments. In: Behavioral decision making. Ed Wright G. Plenum Press, New York, USA.

Flandoli F, Giorgi E, Aspinall WP and Neri A, 2011. Comparison of a new expert elicitation model with the Classical Model, equal weights and single experts, using a cross-validation technique. Reliability Engineering & System Safety, 96, 1292–1310.

Fraidin SN, 2004. When is one head better than two? Interdependent information in group decision making. Organizational Behavior and Human Decision Processes, 93, 102–113.

French S, 1985. Group consensus probability distributions: a critical survey In: Bayesian Statistics. Eds Bernardo JM, De Groot MH, Lindley DV and Smith AFM. Elsevier North Holland, 182-201.

French, S, 2011. Aggregating expert judgements. Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A, Matematicas, 105, 181-206.

French S and Rios Insua D, 2000. Statistical decision theory. Arnold, London, UK.

Frey HC and Patil SR, 2002. Identification and review of sensitivity analysis methods. Risk Analysis 22, 553-578.

Garthwaite PH, Kadane JB and O'Hagan A, 2005. Statistical methods for eliciting probability distributions. Journal of the American Statistical Association, 100, 680–701.

Greitemeyer T and Schulz-Hardt S, 2003. Preference-consistent evaluation of information in the hidden profile paradigm: beyond group-level explanations for the dominance of shared information in group decisions. Journal of Personality and Social Psychology, 84, 322–339.

Genest C and Zidek JV, 1986. Combining probability distributions: a critique and an annotated bibliography. Statistical Science 1, 114-118.

Genre V, Kenny G, Meyler A and Timmermann A, 2013. Combining expert forecasts: can anything beat the simple average? International Journal of Forecasting, 108–121.

Gerard HB, Wilhelmy RA and Conolley ES, 1968. Conformity and groupsize. Journal of Personality and Social Psychology, 8, 79–82.

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.

Gigerenzer G, 1991. How to make cognitive illusions disappear: beyond "heuristics and biases". European Review of Social Psychology 2, 83–115.

Gigerenzer G and Hoffrage U, 1994. How to improve Bayesian reasoning without instruction: Frequency formats. Psychological Review, 102, 684-704.

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

Goossens LHJ and Kelly GN, 2000. Special: issue: expert judgement and accident consequence uncertainty analysis (COSYMA). Radiation Protection Dosimetry 90, 293–294.

Gore SM, 1987. Biostatistics and the Medical Research Council. Medical Research Council News.

Hardman D, 2009. Judgment and decision making: psychological perspectives. BPS textbooks in psychology, Blackwell, Chichester, UK, 2009.

Hastie R, 1986. Review essay: Experimental evidence on group accuracy. In: Information Pooling and Group Decision Making: Proceedings of the Second University of California, Irvine, Conference on Political Economy. Eds. Grofman B and Owen G. JAI Press, Greenwich, CT.

Hilbert M, 2012. Toward a synthesis of cognitive biases: how noisy information processing can bias human decision making. Psychological Bulletin, 138, 211–237.

Henry N and Fekete JD, 2006. MatrixExplorer: A dual-representation system to explore social networks. IEEE Transactions on Visualization and Computer Graphics, 12(5), 677–684.

Hoffmann S, Fischbeck P, Krupnick A and McWilliams M, 2006. Eliciting information on uncertainty from heterogeneous expert panels attributing U.S. foodborne pathogen illness to food consumption. Resources for the Future Discussion Paper DP 06-17-Rev. Resources for the Future, Washington, DC, USA.

INL (Idaho National Laboratory), 2005. Simplified expert elicitation guideline for risk assessment of operating events. Available online: http://www.inl.gov/technicalpublications/documents/3310952.pdf

IPCC (Intergovernmental Panel on Climate Change), 2007. Summary for policymakers: contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Available online: http://www.ipcc.ch/pdf/assessment-report/ar4/wg3/ar4-wg3-spm.pdf

IPCS (International Programme on Chemical Safety), 2004. IPCS risk assessment terminology. Harmonization Project document No.1. FAO/WHO, Geneva, Switzerland.

IPCS (International Programme on Chemical Safety), 2009. Principles and methods for the risk assessment of chemicals in food. Environmental Health Criteria 240. FAO/WHO, Geneva, Switzerland.

IPPC (International Plant Protection Convention), 2004. Pest risk analysis for quarantine pests, including analysis of environmental risks and living modified organisms. International Standards for Phytosanitary Measures (ISPM) No 11. FAO (Food and agricultural Organization of the United Nations), Rome, Italy, 30 pp.

IRGC (International Risk Governance Council), 2005. White Paper 1 on risk governance: towards an integrative approach. IRGC, Geneva, Switzerland. Available online: www.irgc.org/IMG/pdf/IRGC_WP_No_1_Risk_Governance__reprinted_version_.pdf

Jacobs RA, 1995. Methods for combining experts' probability assessments. Neural Computing, 7, 867–888.

Janis IL, 1982. Groupthink: psychological studies of policy decisions and fiascoes, 2nd edn. Houghton Mifflin, Boston, MA, USA.

Johnson SR, Tomlinson GA, Hawker GA, Granton JT, Grosbein, HA and Feldman B, 2010. A valid and reliable belief elicitation method for Bayesian priors. Journal of Clinical Epidemiology, 63, 370–383.

Jones JA, Ehrhardt J, Goossens LHJ, Brown J, Cooke RM, Fischer F, Hasemann I, and Kraan BCP, 2001. Probabilistic accident consequence uncertainty assessment using COSYMA: Methodology and processing techniques. EC project report, EUR 18827, 77 pp. Available online: ftp://ftp.cordis.europa.eu/pub/fp5-euratom/docs/eur18827_en.pdf

Kahneman D and Tversky A, 1979. Prospect theory: an analysis of decision under risk. Econometrica, 47, 263–291.

Kahneman D, Slovic P and Tversky A (Eds), 1982. Judgment under uncertainty: heuristics and biases. Cambridge University Press, Cambridge, UK.

Kaplan S and Garrick BJ, 1981. On the quantitative definition of risk. Risk Analysis, 1, 11–27.

Keith DW, 1996. When is it appropriate to combine expert judgements? Climatic Change, 33, 139–143.

Keren G, 1991. Calibration and probability judgments: conceptual and methodological issues. Acta Psychologica, 77, 217–273.

Kerr NL, MacCoun RJ and Kramer GP, 1996. Bias in judgement: comparing individuals and groups. Psychological Review, 103, 687–719.

Kharroubi SA, Brazier JE, Roberts J and O'Hagan A, 2007. Modelling SF-6D health state preference data using a nonparametric Bayesian method. Journal of Health Economics, 26, 597–612.

Knol AB, Slottje P, van der Sluijs JP, and Lebret E, 2010. The use of expert elicitation in environmental health impact assessment: a seven step procedure. Environmental Health, 9, 1-16.

Koriat A, Lichtenstein S and Fischhoff B, 1980. Reasons for confidence. Journal of Experimental Psychology: Human Learning and Memory, 6, 107-118.

Kurowicka D, and Cooke RM, 2006. Uncertainty analysis with high dimensional dependence modelling. John Wiley and Sons, Chichester, UK.

Kynn M, 2008. The 'heuristics and biases' bias in expert elicitation. Journal of the Royal Statistical Society, Series A 171, 239–264.

Larson JR Jr, Christensen C, Franz TM and Abbott AS, 1998. Diagnosing groups: the pooling, management and impact of shared and unshared case information in team-based medical decision making. Journal of Personality and Social Psychology, 75, 93–108.

Lave J and Wenger E, 1991. Situated learning: legitimate peripheral participation. Cambridge: University Press, Cambridge, UK.

Lawrence MJ, Edmundson RH and O'Connor MJ, 1986. The accuracy of combining judgmental and statistical forecasts. Management Science, 32, 1521–1532.

Lin S and Cheng CH, 2009. The reliability of aggregated probability judgments obtained through Cooke's classical model. Journal of Modeling in Management, 4, 149–161.

Lindley DV, 1982. Scoring rules and the inevitability of probability. International Statistical Review, 50, 1–11.

Lindley DV, 1987. The probability approach to the treatment of uncertainty in artificial intelligence and expert systems. Statistical Science 2, 17–24.

Lock A, 1987. Integrating group judgments in subjective forecasts. In: Judgmental forecasting. Eds Wright G and Ayton P. John Wiley and Sons, Chichester, UK.

MacIntyre A, 2007. After virtue: a study in moral theory, 3rd edn. Duckworth, London, UK.

Malkiel BG, 2011. A random walk down Wall Street: the time-tested strategy for successful investing. W. Norton, New York, USA.

Meyer MA and Booker JM, 1987. Sources of Correlation between Experts: Empirical Results from two Extremes. Los Alamos National Laboratory Report LA-10918-LA: Los Alamos, New Mexico and Nuclear Regulatory Commission Report NUREG/CR-4814, Washington, DC, USA.

Meyer MA and Booker JM, 2001. Eliciting and analyzing expert judgment: a practical guide. American Statistical Association and the Society for Industrial and Applied Mathematics, Alexandria, Virginia, USA.

Moore G and Beadle R, 2006. In search of organizational virtue in business: agents, goods, practices, institutions and environments. Organization Studies, 27, 369–389.

Moore PG and Thomas H, 1988. The anatomy of decisions, 2nd edn. Penguin, Harmondsworth, UK.

Murphy AH and Brown BG, 1985. A comparative evaluation of objective and subjective weather forecasts in the United States. In: Behavioral decision making. Ed. Wright G. Plenum Press, New York, USA.

NRC (National Research Council, Committee on the institutional means for assessment of risks to public health), 1983. Risk assessment in the federal government: managing the process. National Academy Press, Washington, DC, USA.

Oakley JE and O'Hagan A, 2004. Probabilistic sensitivity analysis of complex models: a Bayesian approach. Journal of the Royal Statistical Society, B 66, 751–769.

Oakley JE and O'Hagan A, 2007. Uncertainty in prior elicitations: a nonparametric approach. Biometrika, 94, 427–441.

O'Hagan A, 1998. Eliciting expert beliefs in substantial practical applications. The Statistician, 47, 21–35 (with discussion, pp. 55–68).

O'Hagan A, 2012. Probabilistic uncertainty specification: overview, elaboration techniques and their application to a mechanistic model of carbon flux. Environmental Modelling and Software, 36, 35–48.

O'Hagan A and Wells FS, 1993. Use of prior information to estimate costs in a sewerage operation. In: Case studies in Bayesian statistics. Eds Gatsonis C, Hodges JS, Kass RE and Singpurwalla ND. Springer-Verlag, New York, USA, 118–163.

O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson D, Oakley JE and Rakow T, 2006. Uncertain judgements: eliciting experts' probabilities. John Wiley and Sons, Chichester, UK.

OIE(Office International des Épizooties - World Organisation for Animal Health), 2012. Terrestrial Animal Health Code, 21st edn. OIE, Paris, France.

Oskamp S, 1962. The relationship of clinical experience and training methods to several criteria of clinical prediction. Psychological Monographs, 176 (28, Whole No. 547).

Oskamp S, 1965. Overconfidence in case study judgments. Journal of Consulting Psychology 1965; 29 (3): 261-265.

Parenté FJ and Anderson-Parenté JK, 1987. Delphi inquiry systems. In: Judgmental forecasting. Eds Wright G and Ayton P. John Wiley and Sons, Chichester, UK.

Phillips LD, 1987. On the adequacy of judgmental forecasts. In: Judgmental forecasting. Eds Wright G and Ayton P. John Wiley and Sons, Chichester, UK.

Plous S, 1993. The psychology of judgement and decision making. Mc Graw-Hill Series in social psychology, New York, USA.

RIVM (Rijksinstituut voor Volksgezondheid en Milieu), 2008. Expert Elicitation: Methodological suggestions for its use in environmental health impact assessments. RIVM Letter report 630004001/2008. RIVM, Bilthoven, the Netherlands.

Roosen J and Hennessy DA, 2001. Capturing experts' uncertainty in welfare analysis: an application to Organophosphate use regulation in U.S. apple production. American Journal of Agricultural Economics 83, 166–182.

Rowe G, 1992. Perspectives on expertise in the aggregation of judgments. In: Expertise and decision support. Eds Wright G and Bolger F. Plenum, Press, London, UK, 155–180.

Rowe G, 2010. Assessment of the COT uncertainty framework from a social science perspective: a theoretical evaluation. Food Standards Agency, London, UK, 58 pp..

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

Rowe G and Wright G, 2001. Differences in expert and lay judgments of risk: myth or reality? Risk Analysis, 21 (2), 341–356.

Saltelli A, Chan K and Scott EM, 2000. Sensitivity Analysis. John Wiley and Sons.

Schweiger DM, Sandberg WR and Ragan JW, 1986. Group approaches for improving strategic decision-making: a comparative analysis of dialectical inquiry, devils advocacy, and consensus. Academy of Management Journal, 29, 51–71.

Schweiger DM, Sandberg WR and Rechner PL, 1989. Experiential effects of dialectical inquiry, devils advocacy, and consensus approaches to strategic decision making. Academy of Management Journal, 32, 745-772.

Seaver DA, 1978. Assessing probability with multiple individuals: Group interaction versus mathematical aggregation (Report No. 78-3). Social Science Research Institute, University of Southern California, Los Angeles CA, USA.

Shafer G, 1976. A mathematical theory of evidence. Princeton University Press, Princeton, NJ, USA.

Smits T, and Hoorens V, 2005. How probable is probable? It depends on whom you are talking about. Journal of Behavioral Decision Making, 18, 83–96.

Sniezek JA, 1992. Groups under uncertainty: an examination of confidence in group decision making. Organizational Behavior and Human Decision Processes, 52, 124–155.

Soll JB and Klayman J, 2004. Overconfidence in interval estimates. Journal of Experimental Psychology: Learning, Memory and Cognition, 30, 299–314.

Stanovich KE and West RF, 2007. Natural myside bias is independent of cognitive ability. Thinking and Reasoning, 13, 225–247.

Stasser G, Stewart DD, and Wittenbaum GM, 1995. Expert roles and information exchange during discussion: the importance of knowing who knows what. Journal of Experimental and Social Psychology, 31, 244–265.

Stasser G, and Titus W, 1985. Pooling of unshared information in group decision making: biased information processing during discussion. Journal of Personality and Social Psychology, 48, 1467–1478.

Swanson RA, and Holton III EF, 2001. Foundations of human resource development. Berrett-Koehler Publishers, Inc., San Francisco, CA, USA.

Szwed PS, and van Dorp JR, 2002. A Bayesian model for rare event risk assessment using expert judgement about paired scenario comparisons. ASEM National Conference.

Torrance GW, 1986. Measurement of health state utilities for economic appraisal: a review. Journal of Health Economics 5, 1–30.

US-EPA, 2011 (United States – Environmental Protection Agency) Expert Elicitation Task Force White Paper. US_EPA, Washington, DC, USA (Draft version, 2009). Available online: http://www.epa.gov/stpc/pdfs/ee-white-paper-final.pdf, http://www.epa.gov/osa/pdfs/elicitation/Expert_Elicitation_White_Paper-January_06_2009.pdf

Van de Ven AH, and Delbecq AL, 1971. Nominal versus interacting group processes for committee decision making effectiveness. Academic Management Journal, 14, 203–213.

Walley P, 1991. Statistical reasoning with imprecise probabilities. Chapman & Hall, London, UK.

Wenger E, 1998. Communities of practice: learning, meaning and identity. Cambridge University Press, Cambridge, UK.

Wiper MW and French S, 1995. Combining experts' opinions using a Normal-Wishart model. Journal of Forecasting 14, 25–34.

Wright G, Bolger F and Rowe G, 2002. An empirical test of the relative validity of expert and lay judgments of risk. Risk Analysis, 22, 1107–1122.

Wright G, Rowe G, Bolger F and Gammack J, 1994. Coherence, calibration and expertise in judgmental probability forecasting. Organizational Behavior and Human Decision Processes, 57, 1–25.

Yunusov AR, Cooke RM and Krymsky VG, 1999. Rexcalibr-integrated system for processing expert judgement: In: Proceedings 9th Annual Conference Risk Analysis: Facing the New Millenium Rotterdam, The Netherlands, Oktober 10-13. Eds. Goossens LHJ. University Press, Delft, The Netherlands.

Zimmermann HJ, 2001. Fuzzy set theory—and its applications, 4th edn. Springer, New York, USA.

APPENDICES

Contents:

A. Principles and Practical Recommendations 131

B. Extensive Web search on Guidance Documents on expert knowledge elicitation 191
B.1. Objectives and scope of the literature search 191
B.2. The literature search and selection processes 191
B.2.1. Criteria for assessing report relevance 191
B.2.2. Information sources searched 192
B.2.3. Search terms 192
B.2.4. Exclusion of web pages 193
B.2.5. Identification of relevant publications 193
B.3. Results of the literature search and selection processes 194
B.4. Discussion 199
B.5. References 199

C. Case Study in Animal Health and Welfare 200
C.1. Risk of introduction of Rift Valley fever into the Southern Mediterranean area through movement of infected animals 200
C.2. Lessons learnt 200

D. Case Study in Plant Health 202

E. Case Study in Biological Hazards 257
E.1. Global Burden of Foodborne diseases: Expert elicitation on sources of disease 257

A. PRINCIPLES AND PRACTICAL RECOMMENDATIONS

A.1. PRINCIPLES AND PRACTICE OF PROBLEM, PROCESS AND PROTOCOL DEFINITION

A.1.1. Search for evidence (identification of information deficit)

Once an issue has emerged as of interest to EFSA, it is clearly necessary to first ascertain what is known about that issue, before attempting to use EKE to fill any remaining information gaps. The usual approach will be to search relevant databases and publications. EFSA already has advice on how to conduct systematic literature reviews or use fit-for-purpose databases (see EFSA, 2010a, b). However, it is possible that there may be relevant information that is not known to EFSA and not uncovered through its search strategies. Experts may be aware of such information, and using members of the Working Group in this identification process would be good practice. Thus, the first stage of any knowledge elicitation might involve consulting experts and then reconsidering whether or not further EKE is required (i.e. sufficient information is now available). Experts may also have information on which *other* experts might be able to fill information deficits (a snowballing approach to expert identification). Which experts to approach, and the way to approach these, are elucidated in a subsequent section.

At this point it is important to note that the expert utilisation process of EFSA essentially has two phases, with different tasks allocated to the experts identified in each phase. The first phase generally involves the identification and selection of a Working Group to consider a problem. EFSA has particular rules and procedures for the identification of experts for such groups (EFSA ED, 2013). These groups are then required to follow a predefined mandate that details their roles and the nature of the problem to be considered. Identifying information gaps, and expert or information sources that might fill these gaps, would seem a task suited to *this* group. For example, 'hearing experts' might be identified to attend group meetings to provide additional evidence on the problem at hand. Ultimately, such groups might be used by EFSA to help define the problem to be answered (within the constraints of the group's mandate), such as by identifying an appropriate risk model; to select experts to approach; and to propose a process to elicit the information needed to provide a suitable risk assessment. Other experts might then be approached in a second phase to take part in formal knowledge elicitation with regards estimating the parameters of the identified risk model.

> **Recommendation:** Conduct a consultation with relevant experts at the start of a process to identify relevant information sources of which EFSA may be unaware, and to identify a preliminary list of other experts who might be able to address deficits later in a formal EKE approach. This process may take place through EFSA working groups or the use of hearing experts.

A.1.2. Defining the question for the experts

EFSA's primary role is to conduct risk assessments on potential hazards that might impact human, plant, and animal health and welfare. A risk analysis tries to answer three questions: (a) What can happen?/what is the hazard? (b) How likely is it to happen? (c) Given that it occurs, what are the consequences? A way to formalise these questions is to think in terms of scenarios, each with a certain probability and a set of consequences (Kaplan and Garrick 1981). An important part of the identification and quantification of scenarios, probabilities and consequences is the data gathering process. Whenever empirical data is unreliable, sparse, or simply not available, only expert judgement can fill the gaps. How this is done can be critical to a risk assessment. When performed rigorously and within a formalised approach, structured elicitation and pooling of opinions from a spectrum of experts is a powerful means for quantifying a risk model.

A distinction can be made between models that are tautological and those that are science based. A tautological model simply expresses the output quantities in terms of input quantities, with the input quantities defined and combined in ways that logically produce the desired output quantities. For example, if a desired output is the weight of infected meat entering the food chain, then this can be logically expressed as the number of infected animals times the average weight of infected meat per animal. But models often also contain relationships that are empirical or based on theory. Such science-based relationships are not necessarily tautologically true.

The choice of science-based elements in the model is clearly a judgement that should be made carefully and with the use of all available expertise—it is a form of expert elicitation. Ideally, uncertainty about the validity and accuracy of a science-based model should be elicited, although this is difficult. When eliciting expert knowledge about quantities in a science-based model, it is important for the experts to accept and understand the model. Otherwise they may decline to participate or may provide judgements that are not correctly aligned to the meaning of those quantities within the model.

Even a purely tautological model involves judgement, because choices must be made regarding the level of detail to employ. For example, if the weight of infected meat entering the food chain is not a final output of the model, then we could choose to regard it as a primary input or as an intermediate quantity which is the product of the two primary inputs (number of animals and weight per animal). This choice depends on available evidence and expertise. There might, for instance, be evidence and expertise from slaughterhouses about the number of infected animals, while there is separate veterinary expertise regarding what parts of the infected animal might lead to infected meat products. That would clearly support the more detailed model, with separate elicitation for the two inputs. On the other hand, if that kind of specialised expertise is not available, or if, for instance, the available evidence is from food inspection, then it may be better to use the weight of infected meat as a primary input, to be elicited directly.

Some tautological elements may be introduced during the elicitation. Thus, we may seek to elicit from experts their knowledge about the weight of infected meat, but the experts find it helpful to separate the task into thinking about the two component quantities individually. In general, there are choices to make over how to define and structure the quantities of interest in order to facilitate elicitation, and these choices can be made at the stage of modelling or during elicitation.

Different experts may be suitable for answering different questions. It is important to ensure the appropriateness of one's experts (see section A.2.2) to address the questions at hand. As such, elicitation may need to follow a staged procedure, with some experts used to address questions on, for example, the suitability of a particular risk model, and others used to assess the parameters within the model. Different processes may be apt for eliciting the different types of information needed: for example, a qualitative approach may be necessary to help identify suitable risk models and choose between them, while a quantitative approach is strongly recommended in eliciting uncertain parameters and is essential when it comes to combining these through the risk model.

In general, the more 'qualitative' tasks—such as risk model specification—seem suited to experts participating in Working Groups (see section A.1.1). The focus of this report is on the more quantitative task of eliciting expert knowledge about parameters. We will therefore say little about techniques for eliciting qualitative information (such as using card sorts, or processes from scenario planning to identify key variables), which however could be used by members in working groups (facilitated by EFSA staff or external professional elicitors).

One important issue is to consider *how* experts are typically asked to comment on the subjects on which they are expert, that is to consider the *metrics* they might use to give responses. Questions need to be framed in a sensible and familiar manner to experts to make best use of their expertise; inappropriate question framing may even negate expertise. For example, if particular experts are used to giving judgements using certain metrics (probabilities), it makes little sense to ask them to provide similar judgements using different metrics (e.g. verbal descriptors from a given scale). (See section A.2.2, but also, for example, Rowe and Wright, 2001.) However, when it comes to the elicitation of parameters, the probabilistic representation of uncertainty is recommended, and this may require use of unfamiliar metrics. Because experts might not handle probabilities with great aplomb, there is ample room for improvement and strong indication that training in *reasoning with uncertainty* is necessary, such as via dry-runs, elicitation exercises and workshops. A section on training follows in a later part of this report (chapter A.3). In short, it is important to characterise how one's chosen experts make judgements in their everyday tasks, either to inform how one frames a question or to establish whether it is necessary to train experts to answer questions in a preferred format.

Beyond this, there are various rules for good practice in developing questions, which can be found in any good textbook on social science methodology, as follows.

Recommendation

- Avoid the use of leading questions (e.g. framing questions so that they imply a certain answer, such as the riskiness of a particular hazard).
- Use neutral language (avoiding emotive expressions).
- Avoid compounding questions (e.g. that ask about two or more things, such as 'health AND safety').
- Use language appropriate to the audience (e.g. the precise experts chosen).
- Provide succinct definitions of any technical/scientific terms that might be open to variable interpretation (e.g. by people from different scientific disciplines).
- Try to avoid the use of specific numbers in a question (such as in giving an example), as these can act as anchors for subsequent judgements.

Good questionnaire design in the specific context of eliciting probabilities is considered later, and is detailed in one or two specific elicitation processes. The key rule here is that the problem needs to be consistently framed in the same manner for all experts. Other information that should be given includes:

- the purpose of the elicitation task;
- the assumptions that are being made concerning the problem (with justifications for those assumptions);

the way in which the derived estimates and output from the task will be used.

If there is any relevant scientific information on the problem being addressed, this may also be sent to the experts. When the elicitation method involves bringing the experts together as a group, then the requirement to give all experts *the same* information is easily achieved, but when experts provide their judgements separately (for instance via individual interviews or a questionnaire), they should all be sent the same written briefing material. When the elicitation involves an element of unstructured interaction between expert(s) and elicitor, the interaction can be used to clarify and elaborate on any initial form of words. However, the elicitor must take care throughout the dialogue to avoid making contributions that may inappropriately lead or mislead the experts.

Recommendation: Once a problem is identified, consider the nature of the questions that need to be answered, what type of information is required (identifying models; choosing between models or parameters; quantifying uncertain parameters) and what experts are appropriate to answer the questions.

Recommendation: Ensure one's questions ask experts to express their expertise using familiar language and metrics if possible. If you wish to use a specific metric (e.g. probabilities) that is unfamiliar to experts, then training will be required.

Recommendation: When experts are to provide their judgements individually, or when the elicitation is to be conducted using a questionnaire un-moderated by interaction with the elicitor, perform a dry-run exercise that aims at finding out whether the material (e.g. case structure document and the elicitation format document) are unambiguously outlined and whether they capture all relevant information and questions.

A.1.3. Standard scales

Within EFSA, the most popular elicitation procedures for uncertain parameters of risk models are semi-quantitative, i.e. qualitative expressions are given alongside numeric estimates. Conversion from verbal scales to numerical assessments and vice versa is a process fraught with difficulty and prone to hidden traps. Even though we strongly recommend a purely quantitative representation of uncertainty, we shall further touch upon a number of issues that can arise when a semi-quantitative approach is used.

We have noticed that different EFSA units in different cases have used a variety of scales to elicit and express probabilistic information, particularly in using verbal scales that have varied in the labels used to describe ranges of probability, as well as in the number of scale items (e.g. 5, 7, 10). It may be that the particular scales used in any one case are used in order to reflect the ways in which the chosen experts naturally make probability estimates in their everyday jobs—which would be apt, given the recommendation above about matching expert with question format. However, we suspect that this is generally not the case, and this needs to be confirmed. Whether or not the scale is familiar to experts, the issue of whether verbal scales ought to be used at all (rather than numerical values) is controversial.

The first point to note here is that experts sometimes baulk at using numbers to express uncertainty. There are several reasons for this. First, some experts find it difficult to express uncertainties about questions in this way (perhaps because they commonly do not do this in their everyday lives); second, some experts fear that using numbers can imply excessive (and unjustified) precision; and, third, because experts may be concerned that numerical estimates can be misinterpreted by decision-makers, stakeholders and the public (etc.). Indeed, in response to this concern, several organisations have attempted to produce formal verbal scales—for example, to describe levels of uncertainty about side effects related to taking medicines (e.g. see scales devised by the EC Pharmaceutical Committee and the German Federal Health Agency). Perhaps the best known of such scales is that used by the Intergovernmental Panel on Climate Change (IPCC), which advocates a scale that it has used in its various reports. As an example, the guidelines for the lead authors of the IPCC fourth report (IPCC 2007) recommended using a scale consisting of seven verbal terms (Table 2).

Table 8: Verbal description of probabilities used by IPCC (2007)

Verbal description	Interpretation as probability
Virtually certain	> 0.99
Very likely	> 0.9
Likely	> 0.66
About as likely as not	0.33–0.66
Unlikely	< 0.33
Very unlikely	< 0.1
Exceptionally unlikely	< 0.01

It is important to realise the limitations of using verbal terms rather than providing numerical values. The main problem is that people have different 'linguistic probability lexicons', which is to say they possess different understandings of different verbal terms, such as 'likely'. The interpretation of verbal expressions has been found to differ between people from different countries (cultural differences), between people of different ages, between people with different numeracy skills (and education level) and even *between experts working within domains that frequently deal with uncertainty* (such as accountants, auditors, clinical pathologists and veterinarians). The most robust finding is that between-individual variability is higher than within-individual variability when judging the same terms—although individual variability can also be significant (and influenced by the context in which one asks for evaluations, as demonstrated in the 'heuristics and biases' literature). In short, most people (including experts) perceive the meaning of verbal probabilities *fairly* consistently and reliably, but differently from each other, i.e. research has shown that individuals may have different rank orderings of probability terms, though they do tend to have a *relatively* stable rank ordering of the phrases over time (see Rowe, 2010, for references).

In several contexts, verbal scales have been proposed for *communicating the outputs* of a risk analysis, but some of these have been empirically examined and revealed to be problematic. For example, the EC Pharmaceutical Committee's guidelines in 1998 advocated a system for conveying the level of risk of side effects to people taking a medicine, using the labels "very common", "common", "uncommon", "rare" and "very rare" (e.g. "common" is 1–10 % frequency and "rare" is 0.01–0.1 %). Berry et al. (2003) summarised a number of studies involving members of the general public, patients and hospital doctors that evaluated the utility of the EC guideline descriptors. In all of these studies it was found that people significantly overestimated the likelihood of adverse effects occurring given specific verbal descriptors, which resulted in significantly higher ratings of their perceived risks to health and significantly lower ratings of their likelihood of taking the medicine. The IPCC system has also been studied, and similar problems in interpretation have been identified, including order reversal and beliefs (e.g. about global warming) impacting on the interpretation of phrases. (See Rowe (2010) for references and full discussion of this issue.) Given that numerical terms are less open to variable interpretation, it would seem that these should be sought when possible (EFSA SC 2009).

The use of verbal scales for *describing uncertainty in the inputs* to a risk analysis is even more problematic. Different understandings of the meanings of verbal terms between the expert(s) and the elicitor/analyst may result in bad judgements. Furthermore, as discussed in chapter A.4.6, algorithms used to combine qualitative expressions of uncertainty are arbitrary and can lead to misleading conclusions regarding overall risk. It is therefore important to elicit quantitative probability judgements from experts if possible. However, it may be that experts refuse to use numbers despite the elicitor's requests and explanations. In anticipation of such problems it is appropriate to make a strategic recommendation, which is that a standard scale be constructed for general use by EFSA whenever experts strongly resist quantitative elicitation. Use of a standard scale in all EKE exercises is likely to enhance the utility of such a scale by encouraging the learning of a single 'language of probability' across EFSA units, experts, and so on.

> **Recommendation:** In eliciting uncertainty, elicit numerical probabilities instead of verbal expressions of probability if you can.
>
> **Recommendation:** If you cannot elicit numerical values for some reason, and need to elicit verbal descriptors instead, it is best for EFSA to settle for a single scale and encourage its generic use (where appropriate) in order to help develop a common linguistic probability lexicon.

A.1.4. How to select the most suitable method for this problem

The most appropriate method will depend upon the nature of the question being asked (identifying models, identifying parameters, risk factors and the relationship between these; choosing amongst a number of alternatives; eliciting uncertain parameters) and the nature and availability of appropriate experts. In any one case, there may be several stages of EKE, marked by asking different experts different types of questions. There are many different methods for eliciting information from experts, ranging from the simple individual interview (e.g. for model selection) to more complex processes for eliciting probability distributions for uncertain parameters of models. However, as noted, the focus here is on eliciting uncertain parameters.

The most important dimensions for distinguishing between methods are largely based upon (a) whether the method is intended to be used with individuals or with groups and (b) whether one's experts are able to interact face to face or only remotely. When a panel of experts is used, the methods also differ in the type of consensus which they aim for, and the manner in which this consensus is reached.

In choosing methods, the general rules are that (a) it is better to use multiple experts rather than one and (b) face-to-face interactions (using structured/facilitated approaches – *not* unstructured ones) are preferable to 'remote' methods (*all else being equal …* although in many cases, face-to-face approaches may not be practically viable). The rationale for the former is that more experts possess more potentially relevant knowledge on the topic than any one expert. (A subsidiary benefit is that using experts from different groups or stakeholders may also help militate against claims of bias that might arise from the choice of a single expert whose views might be seen – justly or otherwise – to somehow be influenced by the orientation of the expert's organisation). The rationale for the latter is that interaction between the experts *can in theory* lead to 'process gain', in which learning takes place and misconceptions are revealed, such that the group judgement is better than that of any individual within it. Naturally, there are ways in which the likelihood of such process gain can be increased, such as through the use of a skilled elicitor and the use of a dry run (training) before the elicitation. In contrast, unfacilitated groups rarely show 'process gain' and instead tend to show 'process loss', when extraneous factors hinder good judgement, for example the group's judgement is drawn towards the position of the arguments of the most dogmatic individual (which are not necessarily the best). Furthermore, agreement among interacting experts may not be possible.

Assuming that multiple experts are used in the elicitation, their opinions need to be aggregated. Aggregation approaches can be classified into behavioural and mathematical aggregation approaches (Clemen and Winkler, 1999), or 'mixed' approaches that combine aspects of behavioural and mathematical approaches (Ferrell, 1985; Rowe, 1992). Behavioural aggregation essentially involves experts aggregating judgments themselves, as when a group, following discussion, comes to an agreement about a particular judgement value. Mathematical aggregation involves collecting judgements from the individual experts, which are then aggregated by an external person, such as the elicitor, using an aggregation rule. Mixed approaches allow a degree of interaction, but then use an aggregation rule rather than allowing the experts to combine the judgements (hoping to avoid difficulties and biases than can be associated with this stage of the judgement process).

If mathematical aggregation is chosen, interaction between experts is usually discouraged as it introduces undesired dependence. Moreover, when scientists disagree, attempts to impose agreement may promote confusion between consensus and certainty. Several approaches for mathematical aggregation are proposed and reviewed in, for example, Cooke (1991), Jacobs (1995), Clemen and Winkler (1999), French and Rios Insua (2000), Goossens and Kelly (2000), Garthwaite et al. (2005) and O'Hagan et al. (2006). Amongst others, they include Bayesian approaches and linear pooling approaches (discussed later).

The simplest way to combine experts' opinions mathematically is by using equal weights. Equal weighting is often undertaken because the decision maker cannot find a reasonable basis on which to differentially weight the experts, or for convenience (i.e. this is a quick and unproblematic approach). Having said this, an aggregate judgement can be *better* than the judgement of the 'best' expert, depending upon the spread of judgements of the experts and where these lie with respect to the 'true' answer to a problem (thus justifying an aggregation approach). That is, when all of the experts' judgements lie to one side of the 'true' value (e.g. they all overestimate a quantity), then the 'best' expert will outperform a group aggregation, though when the experts' judgements straddle the 'true' value (e.g. some overestimate and some underestimate) then an averaging approach *can* (though not always *will*) lead to judgement that is better than the 'best' expert.

However, it has been argued that an equal weighting approach implicitly assumes that all experts are equally informed, equally proficient and free of bias. In some methods, aggregation is then done using weights that reflect the perceived quality of the experts' judgements, as indicated—for example—by their performance in expressing their knowledge about 'seed' quantities that are designed to be similar to the quantities of interest but whose values are known to the elicitor (Cooke, 1991). A performance-based weighted combination distribution is seldom identical to the distribution of any one expert but does represent a rational consensus of the information provided by the experts as a group, differentiated by their performance.

In general, a *structured* procedure and the *presence of the elicitor* during the elicitation are important, with structured approaches having been shown to be better than unstructured ones. For example, the Delphi method (a structured 'remote' 'mixed' method) has been shown to outperform unstructured groups as well as statistical groups (comprising the equal-weighted combination of judgements from a number of non-interacting individuals), albeit in studies that have generally used non-expert subjects (e.g. students) in almanac estimation or short-term forecasting tasks (i.e. not using probability distributions) (Rowe and Wright, 1999). Elsewhere, in a large study (involving over 67 000 experts' subjective probability distributions), structured performance-based aggregation of non-interacting experts opinions has been shown to outperform both the equal weighted combination and the best expert, in most cases (Cooke and Goossens, 2008) (that is, better at predicting the seed quantities, not the substantive quantities of interest).

It is possible that it will be beneficial to use a two-step procedure, in which individual experts are used first, with a second, group stage only taking place if there is significant disagreement between the experts (e.g. beyond a particular magnitude or pre-set criterion) (see the approach in INL, 2005). Use of the Delphi technique also intrinsically includes this possibility, with experts being approached by questionnaire on a number of rounds, a process which could be limited to a single round if there is close agreement between expert responses.

Beyond choice of a broad 'type' of method for EKE, it is the actual choices of specific processes used within each method, and the competent facilitation and enactment of these methods, that will lead to the best and most efficient EKE. Principles regarding the enactment of the various stages of such methods are provided later.

> **Recommendation:** Use multiple-expert methods rather than single-expert methods when possible, as multiple experts possess more potentially relevant information than any single best expert.
>
> **Recommendation:** Use *structured* face-to-face approaches rather than 'remote' approaches when possible.

A.2. PRINCIPLES AND PRACTICE OF SELECTING AND MOTIVATING EXPERTS

A.2.1. Rationale for chapter

Clearly for expert elicitation to occur you need at least one expert—for most applications you will need several. However, there are a number of theoretical and practical considerations involved in finding experts for knowledge elicitation and, once they have been recruited, ensuring their involvement until the exercise is completed. For example, theoretically speaking, how many experts is optimal? What kind of expertise should the experts have? And is it better to have experts with homogeneous or heterogeneous knowledge? And, from a practical point of view, how do you find experts? How do you persuade them to join a project, and stick with it? And how do you evaluate their expertise? In this chapter we aim to give some answers to these questions and others, and also provide some guidance and tools for those embarking on a knowledge elicitation exercise.

Before continuing to our discussion we wish to remind the reader that this guidance addresses *the elicitation of probability distributions for numeric estimates of quantities*. Other types of knowledge can be elicited from experts, in which case the theoretical and practical considerations might be different. For example, if one wished to elicit decision-making strategies to incorporate into an expert decision-making system, then one would probably spend a long time with a few experts delving deep into their knowledge structures. However, for our purpose we will usually prefer more experts, with varied opinions (and ability to express these opinions probabilistically), over fewer, but will need each expert to devote comparatively less time to the elicitation process than in the expert system case. In both cases, though, selection of experts who are able and willing to commit time and effort, and who will respond positively to the elicitation method, will be important considerations.

Chapter goals

We wish to:

- identify the principles for selecting, recruiting, motivating and retaining high calibre experts
- make practical suggestions for realising these principles.

Chapter structure

- First define expertise and consider variations in its extent and quality.
- Second, on the basis of this analysis, select high-quality experts with expertise appropriate to a project.
- Third, identify the important considerations when selecting multiple experts.
- Finally, address the practicalities of recruiting and retaining experts (and getting them to give their best).

A.2.2. Defining expertise

A.2.2.1. Common definitions

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, 2001, p. 85)

However, as we shall see, 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—as opposed to so-called ‘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 —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.

Further, expertise is often ascribed on the basis of role (and symbols of that role such as the scientist’s or doctor’s white coat), while 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 as the correlation between social rank and skill or knowledge is weak (Burgman et. al, 2011) because of the number of ways to gain rank in many fields other than by knowing a lot (e.g. ‘old boy’ networks, personality traits such as self-confidence and charisma, being a ‘squeaky wheel’, appearing to work hard, coming from a wealthy family, being in ‘the right place at the right time’ etc.).

Two basic views of expertise underlie the definitions above:

- expertise 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);
- expertise 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.

Germain (2006), drawing on earlier work by Swanson and Holton (2001), developed the Generalized Expertise Measure (GEM), which mainly captures aspects of the knowledge-based views of expertise. The GEM’s 16-item scale contains both objective (e.g. education, training and qualifications) and more subjective expertise items (e.g. self-assurance, potential for self-improvement and intuition). The potential of the GEM for identifying expertise will be discussed in section A.2.3.

Although seemingly perhaps 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 presented to them) as requiring knowledge or skills, or both, 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 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).

Usually 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, arcaneness) might be used by outsiders—neither might be particularly well correlated with knowledge or skill-based expertise.

A.2.2.2. 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. Thus, if the aim is to model decision processes then an intermediate or even a novice might be more useful.

For the purpose of eliciting probability distributions we would normally wish to recruit experts with as much relevant experience as possible; however, there are a couple of exceptions:

- 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 many fewer years of experience in total, but more of these have been acquired after the change point.
- There is reason to believe that greater experience leads to entrenched thinking or biases 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.

A.2.2.3. Granularity and scope of expertise

> *the individual should not be considered expert unless he or she is knowledgeable at the level of detail being elicited*
>
> (Meyer and Booker, 2001, 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, an even better choice might be an expert who has studied a variety of threats to that specific crop—including the insect in question—particularly if they have local contextual knowledge.

A.2.2.4. Types of expertise

A.2.2.4.1 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. In the context of eliciting probability distributions we are generally more interested in declarative knowledge—quantitative judgements and risk assessments—than the manner in which these assessments are arrived at, so perhaps the problem of automaticity, and consequent lack of access, is not a problem as such (i.e. the knowledge we want can, in principle, if not in practice, be stated by the expert).

The distinction between procedural and declarative knowledge might be blurred, though. For instance, as was noted in part I, chapter 2, 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 then be arguments for using less experienced experts for whom reasoning processes have not become implicit. However, given that the nature of expert probabilistic reasoning remains an empirical question, it would be premature to make any particular recommendations in this respect.

A.2.2.4.2 Theoretical versus practical

This distinction is often related to both the previous and the following distinctions but the distinctions are 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). In the context of eliciting probability distributions we will mostly be concerned with practical expertise (i.e. the application of expertise in risk assessment to the particular case in hand), but possibly interested in theoretical knowledge too (e.g. to formally document the elicitation process or if reasons for a probability judgement are elicited in a Delphi procedure).

A.2.2.4.3 Substantive versus normative

Again this is related, but not identical, to the previous distinctions. Substantive knowledge concerns particular domains whereas normative knowledge is about formal, abstract methods for expressing domain knowledge (e.g. knowledge regarding factors affecting likelihood of precipitation vs. knowledge regarding probability theory and expressing forecasts on a probability scale). While substantive expertise is what we are chiefly after, normative expertise will usually assist in the elicitation (e.g. of well-calibrated and coherent probabilities).

Lack of normative expertise may be a reason why expert judgements are sometimes little or no better than lay judgements. Koriat, Lichtenstein and Fischhoff (1980) propose that probability judgment 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. To be specific, 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. Noisy and/or incoherent judgements may result in frequently observed biases such as sub-additivity and overconfidence being actually or apparently manifest (see, for example, Brenner, 2003; Hilbert, 2012). As argued elsewhere (Wright et al., 1994), reliable or coherent probability judgements may be a prerequisite for valid (i.e. well-calibrated) judgements. Further, 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) while it is often observed in practice that experts are reluctant to put any numbers on their uncertainty, preferring instead to make verbal statements such as "I am fairly sure" or "it is pretty unlikely". For such reasons we (and others, e.g. Phillips, 1987) propose that experts are trained in expressing uncertainty as probabilities.

A.2.2.5. Principles and practice of defining expertise

Principles

Greater degrees of expertise are not always to be preferred to lesser—look out for evidence of structural change in the domain of interest, or of institutional biases.

Furthermore, expertise must also be at the appropriate level of granularity and of the appropriate type—for elicitation of risk estimates we will primarily want experts with declarative, practical and substantive knowledge, but normative and theoretical expertise is also desirable (although it may be possible to train up experts in these latter skills and/or design the elicitation process to help experts overcome any deficiencies in normative and theoretical knowledge).

Practice

The type of expertise needed for the elicitation needs to be defined, and variations in its extent and quality need to be considered. There are three potential solutions to this problem, (in decreasing order of desirability):

- Select experts with both appropriate substantive and normative expertise—these may be evaluated using our questionnaire (Table 11). We will discuss the questionnaire in detail below in section A.2.3.2, 'Limits of expertise'.

- Select experts with appropriate substantive expertise and train them in the response mode (it is likely that in many domains the two forms of expertise are related; thus, those with both normative and substantive expertise will be better substantively than a substantive-only expert who has been thoroughly schooled in the normative for the purpose of EKE). Question 15 of our questionnaire (Table 11) aims to determine whether experts have subsequently received training in quantitative expression of uncertainty.

- Select experts with appropriate substantive expertise and use a response mode with which the expert is familiar (e.g. elicit verbal probability terms then map them on to a numeric scale; however, research shows that this is difficult to do reliably (e.g. Moore and Thomas, 1988; Smits and Hoorens, 2005).

A.2.2.6. Identifying experts

The first step in identification of experts is to determine appropriate expertise profiles and expert roles. This requires the Steering Group to decide what type of expertise is needed to answer the elicitation question, as well as the likely expert roles and other criteria that are needed to compile the list of potential experts. Identification roles/criteria could be for example: disciplines (e.g. animal welfare; communication; production technology), institutions (e.g. governmental body; feed industry; research), or geographical domains (e.g. West Africa; Eastern Europe; global).

An example of potential expert profiles is as follows (it is useful to divide required expertise into 'essential' and 'desirable' categories to avoid creating criteria that are so specific that they cannot be satisfied—desirable characteristics may sometimes be instilled through training or those lacking in them might be engaged in an advisory, non-elicitation role (see part II, chapter3 to 7 for the detailed example):

Essential expertise

- substantive knowledge of transport/handling conditions of a specific food item;
- substantive knowledge of differences in production, trade and processing of a specific food item;
- substantive knowledge of conditions of a specific food item at the border, during transport/handling and at the end user site.

Desirable expertise

- normative knowledge of expressing biological contaminations in standard metrics;
- normative ability to quantify risk probabilistically.

The following roles were consequently identified:

- importer/trader with notable volume of import/trade of the specific food item, including distribution to consumer;
- academic food scientist with knowledge of the production and processing conditions of the specific food item;
- food inspector with regular experience in control of the specific food item, preferably at a main point of entry;
- quality inspector for a main processing company of the specific food item/food inspection at end user level.

Further criteria identified were:

- The importer/trader/processor should be able to cover the situation in at least one of the main importing countries.
- Scientists and inspectors should preferably be able to express the biological contamination in standard metrics (or in a metric that can be converted to a standard metric).
- Scientists and inspectors with experience with comparable organisms, foodstuffs or transport conditions.

The expertise profiles should be combined with the roles and other identification criteria to produce a profile matrix. The empty cells in the matrix indicate the 'space' for searching relevant experts. However, we would not expect to find a different expert for each cell—for example, we might not expect to find scientists with country-specific knowledge, or importers/traders with experience with the standard contamination metrics or expressing risk and uncertainty as probabilities. An example profile matrix is given in Table 9.

Once the general profile(s) of the type of experts required for the elicitation has been determined, the next step is to identify specific experts who satisfy the profile(s). It is therefore now time to match expert names to each compartment of the matrix. This list of expert names requires a second matrix which allows you to fill in all needed expert information, such as name, contact details or area of expertise. See below for an example (Table 10). To generate the list of candidate experts (by Steering Group members), searches can be performed using EFSA expert database (see section A.4.4.1 below), institutional contacts and the internet (e.g. professional and social networking sites, specialized search engines and databases e.g. patent databases).—'Snowballing' may also be used (see sections A.2.4.1 and A.2.4.2), but only in addition to the other methods because it usually operates within a single network, thereby risking missing alternative scientific and professional networks.

Table 9: Example of a profile matrix

Knowledge requirements			**Country**	**Roles**			
Expertise	**Importance**	**Specificity**		**Industry**		**Government (inspector)**	**Academia (scientist)**
				Import/ trade	**Production**		
Transport/ handling conditions	Essential	Specific	AA				
			BB				
			CC				
		Comparable	AA				
			BB				
			CC				
Production, trade and processing	Essential	Specific	AA				
			BB				
			CC				
		Comparable	AA				
			BB				
			CC				
Conditions at end user site	Essential	Specific	AA				
			BB				
			CC				
		Comparable	AA				
			BB				
			CC				
Standard contamination metrics	Desirable	Specific	AA				
			BB				
			CC				
		Comparable	AA				
			BB				
			CC				
Expressing risk and uncertainty as probability	Desirable	Specific	AA				
			BB				
			CC				
		Comparable	AA				
			BB				
			CC				

Table 10: Example of contact details of experts

Name	Address (organisation)	Country	Email address	Telephone no	Type of institution	Discipline	Commitment (yes/perhaps)

Where possible, collect a CV from prospective experts. It is also recommended that a questionnaire is sent to the expert and/or to colleagues and peers or a prospective expert in order to determine the nature and potential quality of that person's expertise. Some things to look out for are described below. A template for a questionnaire is given in Table 11.

'Good' indicators of expertise

- Formal qualifications and training
- On-the-job experience
  - amount, e.g. years, of experience
  - type of experience (in terms of types of section A.2.2.4)
  - quality of experience, e.g. practical vs. academic. Not all experience is equal, in particular it is important that the sort of expert knowledge to be elicited can actually be learned; this issue is discussed below in section A.2.3.2)
- Outputs (e.g. papers, reports, talks)
  - evidence of ability to communicate
  - opportunity for peer review (quality of outputs might be assessed by, for example, examining impact ratings, comparing with published rankings or asking for peer opinion)
  - note that experts outside academia will be less likely to publish so other types of output may be better indicators of expertise (e.g. patents) also when experts from industry do publish it is much less likely they will do so in peer-reviewed journals, thus it will be necessary to look at trade publications, technical reports, conference proceedings etc. (but due to confidentiality issues much of the output of such experts may not be in the public domain at all)
- Awards and other 'esteem indicators'
- References (e.g. from other 'experts')—this is to establish that a potential expert has credibility within his or her field (but should not to be used as a basis for giving differential weighting to experts because research suggests that the ability of experts to evaluate the quality of the expertise of themselves or others is poor relative to more objective criteria (Oskamp, 1962), such as experts' calibration (correspondence between objective and subjective probabilities, e.g. Cooke et al., 2008). Accuracy of past judgement is another objective criterion that has been proposed (e.g. Genre et al., 2013). However, a problem with objective weighting criteria is that they can be difficult to measure reliably or obtain at all (see, for example, Lock, 1987).

'Bad' indicators of expertise

- Job title/social role
- Confidence
- Being verbose/prolific
- Media presence
- The reputation of the organisation where the expert is stationed

Although these indicators are 'bad' in the sense that they are unreliable, they may still be associated with expertise. In addition, it should be noted that sometimes a big name might be a useful asset on a project, for instance lending it credibility and 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 elicitiation process. Thus, 'experts' may sometimes be selected for reasons other than the quality of their knowledge.

Information regarding formal qualifications and training can be gathered either by asking experts for a CV, by obtaining a reference from an employer or by asking the opinion of peers—outputs, awards and esteem indicators are probably best obtained from CVs. To obtain information about on-the-job experience, we suggest using a questionnaire. Table 11 presents a questionnaire that was developed and piloted with experts in the case study on the Rift Valley fever elicitation (see Annex B). It is designed to identify aspects of substantive and normative expertise—specifically, parts A to C are designed to collect information from experts regarding their substantive expertise whereas part D is more concerned with their normative expertise. For example, with regard to substantive expertise there are questions concerning the opportunities to learn in the role (e.g. number and quality of data, availability and regularity of useful feedback) and with respect to normative expertise there are questions about the degree of statistical training and support (e.g. availability of formal models, experience and training at expressing uncertainty in terms of probabilities).

Table 11: Template for an expertise questionnaire

Expertise Questionnaire

This questionnaire is intended to find out about the nature of your job, and the type of judgements that you make while performing it. These answers will be used to prepare for the upcoming elicitation workshop on

<name of workshop here>

In particular, we are interested in whether or not your job requires you to make probabilistic judgements, and how you make such judgements. In addition, we are interested to find out what sort of aids you use in making judgements, whether you received any relevant training, and whether you receive feedback about the quality of your judgements.

Part A: General description of your job

1. What is the title of your job?

---

2. How would you describe your area of expertise?

---

3. How many years of experience would you say you had in your area of expertise?

---

4. Would you describe that experience to be:
   Please tick ONE box.

   Wholly practical and/or field-based
   Mostly practical and/or field-based but some theoretical and/or lab-based
   About equally practical/theoretical and field/lab-based
   Mostly theoretical and/or lab-based
   Wholly theoretical and/or lab-based

Part B: The judgements you make

5. Describe the most important judgements that you make on a regular basis in your job.

---

6. 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.)?
   Please tick ONE box.

   I always use my judgement alone
   I mostly use my own judgement
   I use partly my own judgement, and partly other sources
   I mostly use other sources
   I always use other sources alone (not personal judgement)

7. If you do use other information sources, please describe them below.

Part C: Data, models and feedback

8. In making your work judgements, do you receive any feedback about their accuracy?
Please tick ONE box.

   - Always
   - Often
   - Sometimes
   - Rarely
   - Never

9. If you receive some feedback, what form does this take?

---

10. How soon after a judgement, on average, do you receive feedback?
Please tick ONE box.

   - The same day
   - Within a week
   - Within a month
   - Within a year
   - Longer than a year
   - I do not receive feedback

11. How would you rate the ease of making good judgements in your work?
Please tick a box that best represents your opinion.

	1	2	3	4	5	6	7	
Very difficult								Very easy

12. Do you make use of a formal model for making your work judgements?
Please tick a box that best represents your opinion.

	1	2	3	4	5	6	7	
Never								Always

13. How would you rate the availability of data that you use for your work judgements?
Please tick a box that best represents your opinion.

	1	2	3	4	5	6	7	
Poor availability								Plentiful

14. How would you rate the quality of data that you use for your work judgements?
Please tick a box that best represents your opinion.

	1	2	3	4	5	6	7	
Very poor								Very good

15. Did you receive any training to make judgements? If so please describe below.

Part D: Judgments of risk and uncertainty

16. Do you ever make any of the following types of judgements at work (numerically, verbally, or by some other means)? Please tick and fill in as many as are relevant.

I estimate the likelihood/probability of… ________________________________
I estimate the chances of… ____
I estimate confidence in… ____

17. How often, on average, are you called upon to make judgements of risk or uncertainty? Please tick ONE box.

At least once a day
At least once a week
At least once a month
Less than once a month

18. When you make judgments of risk or uncertainty, what forms do they take? Please tick as many boxes as are relevant.

Numerical estimates (e.g. 0.5, 50 %, 1 in 2)
Verbal estimates (e.g. likely, infrequent)
Comparative (e.g. 'the risk is similar to another risk')

19. If you do make numerical judgements, what forms do these take? Please tick as many boxes as are relevant.

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 below

____

20. Please give an example of the type of judgement of risk or uncertainty you typically make (if you do make such judgements).

____

21. Did you receive any training to make judgements of risk and uncertainty? If so please describe below.

____

22. When you have to make judgements of risk and uncertainty do you rely on your judgement alone or do you also use other information sources (such as manuals of statistics, computer databases or programs, etc.)? Please tick ONE box.

I always use my judgement alone
I mostly use my own judgement
I use partly my own judgement, and partly other sources
I mostly use other sources
I always use other sources alone (not personal judgement)

23. If you do use other information sources, please describe them below.

____

Thank you for your time and effort.

Table 12 shows the Generalized Expertise Measure (GEM). The items are designed to identify whether experts have the following characteristics relative to a specific field:

- specific education, training, and knowledge;
- required qualifications;
- ability to assess importance in work-related situations;
- capability to improve themselves;
- intuition;
- self-assurance and confidence in their knowledge.

Since this instrument largely eschews social expertise, it may be a useful instrument for identifying experts; however, as we have already mentioned, the last two dimensions above—self-assurance and confidence—may not be good indicators of expertise (although still useful to know, especially if you plan to use interacting groups). In our view, the GEM would be best given to colleagues or peers of an expert rather than the expert him- or herself because it might be difficult for an expert to answer the questions in an impartial manner—in this respect, the GEM differs from, and is somewhat complementary to, our questionnaire, which is designed to be completed by the experts themselves. Overall, we think that the GEM is most useful if used to assess the potential of candidates suggested by other experts during snowballing (the GEM would be sent to the proposing experts). Having said this we wish to give a couple of words of caution. First it is unlikely that GEM respondents can ever be truly impartial about their assessments of colleagues: most aspects of the questionnaire are highly subjective. Second, it is of course of paramount importance that the responses to GEM be kept confidential and, in fact, it might be advisable to destroy any information derived from it after selection of the expert panel.

Table 12: Generalized Expertise Measure (GEM) (Germain, 2006)

[Please rate the proposed candidate on the characteristics below using the scale

	1	*2*	*3*	*4*	*5*	
Not true at all						*Definitely true]*

This person:

1. has knowledge specific to a field of work.
2. shows they have the education necessary to be an expert in the field.
3. has the qualifications required to be an expert in the field.
4. has been trained in their area of expertise.
5. is ambitious about their work in the company.
6. can assess whether a work-related situation is important or not.
7. is capable of improving themselves.
8. is charismatic.
9. can deduce things from work-related situations easily.
10. is intuitive in the job.
11. is able to judge what things are important in their job.
12. has the drive to become what they are capable of becoming in their field.
13. is self-assured.
14. has self-confidence.
15. is outgoing.

Our suggestions for use of the scale are given in square brackets.

[The higher the better score for answers to 1–7, 9, 11–12 — for the remaining scores, caution should be adopted if the expert is used in a freely interacting group as they could potentially dominate at the expense of possibly more knowledgeable members].

A.2.2.7. Screening experts

If there is doubt about the suitability of an expert, it may be desirable to use a screening process. Clearly, screening could be problematic if framed in a negative way (e.g. 'we are assessing you to see if you are expert enough') but may be acceptable if framed in a more positive manner (e.g. 'familiarisation with the elicitation process'). Screening could take the form of some 'test' questions designed to establish whether an expert meets certain criteria—these criteria would normally be determined by the Working Group responsible. Given that the criteria are likely to be specific to a particular elicitation problem we do not believe that it is meaningful to try to develop some generic screening instrument here. We also recommend that screening is used carefully (and sparingly). As discussed in section A2.5, 'Constructing groups of experts', it is usually desirable to have a range of expert opinions expressed—overly strict criteria or overzealous application of screening may run counter to the goals of creating large and heterogeneous groups. Further to this, as far as possible, a balanced selection of experts is to be desired. In other words, avoid any obvious bias in selection, e.g. regarding education or affiliation (although there is some evidence that such things may not be significant influences on the outcomes of expert elicitation, see for example, Meyer and Booker, 1987; Booker and Meyer, 1988).

A.2.2.8. Limits of expertise (what we can reasonably expect)

Learnability and predictability of the domain

- Bolger and Wright (1994) analysed 20 studies of expert judgement—several 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. The authors argue that, for a judgement task to be 'learnable', it is necessary to receive regular, rapid and reliable outcome feedback—this is particularly true if one is expected to calibrate one's probabilistic judgements. For example, weather forecasters making short-term forecasts of precipitation 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 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 for 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 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, we have included questions about the nature, availability and speed of feedback in our questionnaire in Table 11, along with a request for 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 (see also paragraph after next).

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. 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, we pose several questions of potential experts in our questionnaire in Table 11 regarding the nature, calibre and number of data perceived in the task domain. However, it should be noted that if large numbers 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), permit experimentation with parameters through simulation, and are readily available if future 'judgments' 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.

A.2.2.9. Principles and practice of identifying experts

Principles

- Give greater weight to evidence of training and experience than indicators of 'social' expertise such as rank or media presence (but with due consideration of the next principle).
- Be clear about what you require from an expert, e.g. do you want a person who actually knows a lot or a person who is perceived as knowing a lot (who may attract other experts to join the project)?
- Have reasonable expectations about what can be expected of experts based on a thorough analysis of what they do.
- Exercise sensitivity in selection and screening (you want experts to feel that they are being taken seriously but not being evaluated or, worse, rejected).

Practice

Once the expert list is compiled, experts should be contacted by email, letter or telephone stating that EFSA has an important problem, and that the recipient has been identified as expert in the area; and provided with brief information regarding the problem.

If the expert agrees in principle to co-operate, he or she should be sent the expertise questionnaire in Table 11.

An example of a contact letter can be found in part II, section 5.1. The expertise questionnaire allows EFSA to make a more informed decision whether or not the right experts are included in the elicitation exercise. The questionnaire not only polls the experts' expertise but also asks for suggestions of other (alternative or complementary) experts. In this way a larger sample of experts is obtained as it uses snowballing (Working Group members first fill in the grid, then the identified experts are subsequently invited to provide additional names of experts).

A.2.3. Expert recruitment and retention

A.2.3.1. Where to find experts (e.g. databases)

To have access to the knowledge of experts in the field of food and feed safety, EFSA has established, in cooperation with the Member States, an expert database. Scientists with relevant expertise in the remit of EFSA and who are willing to work with EFSA can apply to be included in the database, which also contains details of all experts who are currently working with EFSA or are still connected with it. Similar databases can be obtained from other appropriate organisations, especially the membership lists of scientific or industrial associations. The search for experts should be taken as an opportunity to enlarge the EFSA database: even if experts are not considered suitable for the current EKE they may be a useful resource in future.

Nevertheless, to answer a concrete question, a description of the expertise required will seldom be sufficient to perform a selection relying only on these databases. In the case of the academic sector, searching for expertise by extensive search of relevant publications is well established. The main authors and their affiliations / corporate authors of these publications will be a starting point to find additional experts. However, in many cases where expert knowledge is needed to provide knowledge for use in EFSA risk assessment, the best available expert group is likely to include industry personnel with practical experience of production/distribution/use etc. Relevant industry associations will play a key role in identifying appropriate experts of this kind. In addition, to introduce a balanced approach, authors of key publications may be used to provide input on the research/publication but should not usually be the sole source of experts as important perspectives might be omitted (e.g. researchers who failed to replicate and consequently could not get published).

For the non-academic sector, the defined expert profiles and roles can be used to obtain further contact points. The structure of international organisations, scientific or industrial associations or national administrations can be used to select relevant units. Associations of relevant producers can be used to select companies with specific knowledge, as can databases of patents. Bibliometric methods could be usefully integrated into this process (commonly used databases include Google Scholar[31], Publish or Perish[32], Scopus[33] & Web of Science[34]) not only through generation of lists of experts (as stated in 4.4.1.), but also generation of lists of corporate experts (public bodies, private companies, NGO, etc).

The EU project FAHRE[35] mapped the key players in food and health research (funding) in each of the 30 countries of the European Economic Union (EEC).

Finally, professional social media (e.g. LinkedIn, ResearchGate) might be a further source to retrieve specialised experts, although we do not recommend this as the sole means for searching experts (as the cliquey nature of such networks inhibits the goal of creating heterogeneity of opinion). It is also worth mentioning recent work in mapping social networks: software has been developed to aid the visualization process (e.g. Gephi (Bastian et al., 2009); Ucinet (Borgatti et al., 2002); CiteSpace II (Chen 2006); Matrix Explorer (Henry & Fekete, 2006); and Pajek (de Nooy et al, (2005)). Because of the ongoing development of these tools further experience is needed to conclude on their use.

[31] Google Scholar: http://scholar.google.co.uk/
[32] Publish or Perish: http://www.harzing.com/pop.htm
[33] Scopus: http://www.elsevier.com/online-tools/scopus
[34] Web of Science: http://thomsonreuters.com/web-of-science/
[35] Food and Health Research in Europe; http://www2.spi.pt/fahre/

All these searches will be used to establish first contact points and start a long-list with possible experts. In the next step the contacts will be used to populate the long-list further and gather additional experts by snowballing.

A.2.3.2. How to approach experts (importance of identity of messenger)

Experts will typically be busy people and, as such, may receive many requests for their time each day. Unsolicited approaches from strangers by email in particular might therefore be ignored (or filtered out as spam). Experts could, then, be initially approached more successfully by regular mail or by telephone than be email, although 'cold' phone calls will be construed as an intrusion by some. Further, approaches from those known to the experts should stand a greater chance of being considered than those from strangers, which is a major advantage of snowballing (although a disadvantage is that it may conflict with the principle of diversity expressed below, because experts will recommend other experts with similar opinions to their own, who are also likely to come from same background and have the same vested interests). Whether or not snowballing is used then it is important that the credentials of the person approaching the expert are established as quickly as possible as otherwise the message may not be attended to.

A.2.3.3. Motivating experts

The initial encounter is very important so it is important to make a convincing case. The initial part of the initial encounter is particularly important: it is necessary to make the reason for the approach clear and stress why the project is important and why the expert is important to the project. Some suggestions for achieving this are:

- Approaches by someone known to the expert can be particularly effective (hence usefulness of snowballing, although this can work against the principle of creating diversity).
- Face-to-face solicitation may be more effective than telephone, telephone more effective than letter, letter more effective than email.
- Flatter the expert. For example, make it clear that the reason for the approach is that he or she plays a leading role in the field and be as specific as possible about the indicators of the expert's standing. If the expert has been recommended by someone prominent, then this should be stated.
- It is important to research the expert before making an approach to find out something about the expert's interests, for example (this is where an approach from someone who knows him or her could also help).
- Experts are more likely to respond positively if they think that the project is worthwhile (e.g. it contributes to society or the development of their field, or sets new standards) so this should be stressed (many scientists these days are under pressure from their funders to demonstrate that their research as societal '*impact*').
- Make it clear what the outputs of the project will be and how they will be publicised (in particular, the expert's access to the outputs should be guaranteed). Meyer and Booker (2001) suggest that experts should be given the right to veto outputs they disagree with. Our view is that this is not a good idea because the elicitation process is expensive and should not be jeopardised by one or two people. Rather, experts should be told at the outset that they do not have right of veto and anyone who finds this unacceptable can opt out at this stage.
- Most of the experts EFSA is likely to be considering will be scientists, and therefore finding things out will be motivating in itself.
- Presenting the project as a challenge or as a novel experience could be motivating for many experts.

- Most experts would like public acknowledgement of their contribution (but not all, so it is important to sound out the expert on this) so this may also be offered.
- Letting the expert know that other experts he or she may hold in regard have already been recruited may also make the project more attractive. (Again, it is important to have a good knowledge of the expert and his or her interests. For example, simply naming another recruit with whom the expert may have a potential conflict might act as a deterrent but explaining how a balance of views is desired might solve this problem.)
- Explaining EFSA's role, reputation and achievements might also be an inducement (i.e. association with EFSA may be regarded as prestigious and important and will look good on the expert's CV).

At the initial encounter, potential disincentives to involvement should be minimised. In particular, worries about the workload involved (e.g. amount, level and timing of commitment) should be alleviated as far as possible, e.g. by specifying the maximum time commitment or by explaining that judgements can be given online at times convenient to the expert, and so on. In some cases, it might be an incentive to increase the extent of the involvement or significance the role. Being as specific as possible about what the expert will be expected to do is better than being vague (and may lead to greater engagement later). Experts should be told that they will not have to give judgements if they are not happy about them and that they can leave the project at any point (part of the principle of 'informed consent'). However, the induction process should lead the expert to be significantly invested in the project such that he or she is be unlikely to exercise this right.

Repeat attempts at recruitment might be required:

- Sometimes this could be more effective if done by different people.
- Sometimes different incentives could be stressed (e.g. money vs. prestige).

BUT must avoid harassment.

Other considerations:

- Is there likely to be travel involved? Some people may like this opportunity, whereas others will not.
- Is presence virtual or physical? Virtual presence reduces workload but may result in a less positive experience for the expert and, therefore, a higher dropout rate An alternative is to use a combination of virtual and physical participation. For example, a virtual presence may suffice at the scoping phase whereas a physical presence may be useful if the expert will be consulted on a more regular basis. This shows more commitment on the part of the expert and lets you know whom you are dealing with.
- What format should the elicitation take? How the judgements are to be elicited may have implications for the expert's desire to participate (e.g. if the process is familiar then he or she will be less likely to experience performance anxiety, will not need training and will be clearer about how long it will take, etc.). Some experts may prefer some methods to others (although ideally it would be desirable to tailor methods to the characteristics of experts, this is unlikely to be practicable).
- Is a payment to be made? There is a view that fees should be avoided—or minimised—except to cover expenses because it crowds out intrinsic motivation (e.g. Meyer and Booker, 2001). However, if the involvement of the expert is likely to require significant time and effort on his or her part then we believe that paying a fee is not only legitimate, but may also be necessary, particularly for recruitment purposes (i.e. before the 'warm glow' of intrinsic motivation can

have its effect). The crowding out effect is unlikely to occur to any significant degree for the amounts of money that will normally be on offer.

- Is fear of being replaced or criticised a consideration? The former seems unlikely to be a major consideration for our purposes; however, the latter could be a problem – stress that, under uncertainty, there are no correct answers, and that they will not be evaluated.

Once experts have agreed to participate, they might be persuaded to assist in the recruitment of further experts (including commenting on the selection criteria/process). This strategy might be useful if there is difficulty identifying relevant experts by other means, but the dangers of encouraging cronyism, building voting blocs and generally reducing heterogeneity of opinion need to be weighed against the advantages.

It is essential to maintain the interest of the experts. To this end it is important to make the expert feel an integral and vital part of the project thus long gaps between contact should be avoided. Further, the expert should be provided with regular feedback about the progress of the project, how their contribution is being used, and any future contribution that might be required from them.

A.2.3.4. Principles and practice of recruitment and retention

Principles

- Maximise factors that motivate experts (e.g. recognition, interest) and minimise those that demotivate (e.g. boring elicitation process, heavy time burden).
- Those who will be responsible for conducting the elicitation should be involved in the recruitment process (i.e. have an idea of what sort of expertise is needed).
- As for empirical research with human participants, the principles of informed consent, anonymity (if requested) and confidentiality should be followed.
- Balance the reduced workload and flexibility of virtual meetings against the efficiency (in some cases) and motivational aspects of face-to-face meetings.
- Extrinsic motivation may ‘crowd out’ intrinsic.

Practice

- Research beforehand and/or discuss with experts at the first meeting their likes and dislikes as there will be differences—some things that will be a positive incentive to one expert will be a negative for others.
- Clear information should be given to experts at the outset regarding what is expected of them and what they will get in return.
- Try to fit in with the experts’ schedule and preferences as much as possible.
- The experts should always be compensated for their expenses and we also recommend paying them a modest fee for any significant time commitment (very large fees are to be avoided, though, as they may crowd out intrinsic motivation).
- Give feedback during and after the project about the promotion and use of elicited expertise.

Once a list of experts is compiled, the Steering Group members should agree on this list, after which the experts can be invited.

The invitation letter checks for willingness to participate in principle and for availability. It does not provide any real details on the elicitation exercise itself. The invitation letter should contain the following information:

- What: state the nature of the problem and the motivation for the elicitation, as well as the conditions of the elicitation exercise (e.g. type of elicitation exercise, workload and compensation).
- When: provide date.
- Why: reasons that the expert was selected.

An example invitation letter can be found in part II, section 5.1.

Once the experts have agreed to participate, more detailed information should be provided. This letter should contain the following information:

- When: date, location, venue of elicitation.
- Procedure, agenda of meeting, incentives (getting paid/reimbursed).
- Provide details on: model, variables, parameters that will be asked about.
- Constitution of the expert group (in short).
- Confidentiality.
- Ask for available additional information (data) that experts may have/want to share in advance (if relevant).
- Ask about concerns that experts may have.

A.2.4. Constructing groups of experts

In almost all cases where we want to elicit substantive, declarative expertise there will be significant divergence of opinion (as opposed to normative and/or procedural expertise, where there may generally be consensus, or a few distinct positions, regarding both theory and practice). This being the case, adding more experts should lead to a better overall outcome than having fewer experts. However, there is reason to believe that potential gains from adding more experts might not be achieved (or even that there is a negative relationship between number of experts and quality of outcome) if certain conditions pertain. For example, Janis (1982) proposed that 'groupthink' can occur when:

- There is a high degree of consensus between experts.
- The group is highly cohesive.
- There is a strong leader (or particularly vocal/influential group member(s)).

Groupthink means that the group members concentrate on maintaining group identity and cohesion rather than on optimising the outcome of group processes (i.e. making good judgements or decisions), with the result that these outcomes can be poorer than those stemming from individuals (while confidence in those decisions may increase). Adding more experts to a group can increase the chance of conformity and groupthink (see, for example, Gerard et al., 1968). Likewise, increasing group size may not have any beneficial effects on the quality of outputs if there are information asymmetries that are not resolved (e.g. due to vested interests that restrict or bias the exchange of information between experts)

A.2.4.1. Heterogeneity of opinion

Heterogeneity of opinion is theoretically desirable as, in the extreme case, if each expert added to a group has identical opinions then there is no net gain to the quality of the output by adding that expert. In fact, as we have just noted above, there can be a disadvantage to adding more experts of similar opinion in that confidence will increase but accuracy will not. However, so long as there is some mechanism for opinions to be exchanged feely, or otherwise integrated (or aggregated) in a relatively unbiased manner (e.g. mathematically, or by a elicitor), then there is potential for net gain from the group combination process relative to polling individual experts when opinions are heterogeneous. Note that a secondary advantage to having heterogeneity of opinion within the expert group is that it may increase the credibility of the outcome for external stakeholders.

Note that the role and effectiveness of heterogeneity of opinion is dependent on the way that the group opinions are combined. In particular, we can contrast behavioural and mathematical aggregation. In the former case, experts will discuss their opinions, and the reasons for holding them, with each other and attempt to reach a consensus whereas, in the latter case, experts will give their opinions individually and these opinions will be combined mathematically (e.g. a weighted or unweighted average).

Where mathematical aggregation is used, heterogeneity of opinion can only be created through the selection process, although some opinions can be given greater weight than others at the aggregation stage, as in Cooke's method discussed in chapter A.4. For example, Aspinall (2010) observes the value of heterogeneity of opinion in one of his expert elicitation exercises—the opinions of a sub-sample of engineers proved more valid than those of a sub-sample of academics when predicting the failure rate of dams, perhaps because the former had more practical experience. The engineers generally performed better on a test composed of 'seed questions' for which the answers were already known; thus, this sub-group's opinions were given overall greater weight than those of the academics in the aggregation, thereby presumably increasing the validity of the final output (although this is not reported).

For behavioural aggregation, heterogeneity of opinions can either be achieved through selection or artificially, for example by giving a group member or the elicitor the role of a 'devil's advocate'.

It can be difficult for experts with minority opinions to express these in freely interacting groups as there may be pressures towards agreement (with either dominant individuals or group views). Further, information may not always be exchanged in a manner that maximises the potential of diverse views owing to censorship by group members. This censorship might by intentional, for example due to protection of the individual's interests, or unintentional, for instance due to incentive structures of which the expert might not be totally aware (e.g. certain organisational cultures—such as social and health services—may be risk averse, while others—such as financial service—may be risk seeking). Analysis of organisational culture, as well as audits of interests, might be of use here for debiasing judgement and/or freeing up information exchange.

Another source of unintentional censorship is a group bias towards discussing shared information over unshared (Stasser and Titus, 1985; Greitemeyer and Schulz-Hardt, 2003). This bias, of course, reduces the value of having multiple experts with heterogeneous opinions. Research has shown that unshared information is more likely to be tendered after the shared information has been given a thorough airing, suggesting that group discussion should be sufficiently long to ensure the latter occurs (Larson et al., 1998; Fraidin, 2004). If there is a diversity of expertise in a group, then encouraging each member to talk about his or her speciality can also help unshared information be considered by the group (e.g. Stasser et al., 1995). A recent unpublished study in the laboratory of one of the authors suggests that the bias towards discussing shared information can be ameliorated if the importance of private information is marked.

Heterogeneity or diversity of opinion is good, but ideally we would also like opinions to cover the entire spectrum of what is credible. Although efforts can be made in the selection process, this is unlikely to be achieved naturally. It is therefore suggested that brainstorming be used to establish boundary conditions (for example, best and worst possible cases), then these positions are considered by the group (e.g. probabilities assigned, as appropriate). Similarly, vested interests should be as wide as possible—this may be easier to arrange (given sufficient experts to draw upon, that is).

Heterogeneity of opinion can be also be created by devil's advocacy; in other words, the elicitor, or one of the experts in an interacting group, can express opinions that are contrary to the views held in common by the group. Schweiger et al. (1986, 1989) show how devil's advocacy can lead to improved group-decision outcomes while Bolger and Wright (2011) suggest how devil's advocacy can be incorporated into a Delphi procedure. However, in practice, whether a elicitor will have sufficient domain knowledge to be a credible devil's advocate is questionable. One of the group's experts might therefore be a better choice, but again there may be problems of credibility if it is known that the expert is presenting views contrary to what he or she actually believes. A solution to this might be anonymity. Another problem may be that experts will have difficulty expressing opinions that run contrary to what they think, or believe (referred to as 'belief bias' and 'myside bias' respectively; see Stanovich and West, 2007). There is some suggestion that giving explicit instructions to ignore prior knowledge and belief can reduce these biases, and that the ability to debias oneself is greater amongst those higher in cognitive ability (op. cit.).

In many instances, then, a more practical solution to the problem of homogeneity of expertise than devil's advocacy might be to ask experts to try to take account of opinions that they know are held in the community, but not present in the meeting, and to note in the formal record that the deficient range of expertise present is a limitation on the conclusions.

A.2.4.2. How many experts?

From a theoretical perspective, adding as many experts as possible would seem beneficial as it should increase the chance that the 'true' values of the variables of interest are amongst those elicited. Furthermore, if judgements are considered measurements of the uncertain quantity, and are subject to error, then a larger number of measurements should reduce this error. From this point of view we recommend recruiting as many as can be found, given time and budgetary constraints.

However, too many experts can be problematic too, depending on the type of elicitation process chosen. Questionnaires can permit the inclusion of many expert participants, but workshops or focus group sessions can easily have too many participants *per session* (so you would need more sessions, and additional elicitors or time). Also, there is a trade-off between quantity and quality of data—sometimes it may be more beneficial to spend an extended amount of time with each expert (probably face to face), at other times a quick straw-poll of many experts (probably by email) might serve your purposes better—a third option would be to combine both the 'broad' and 'deep' approaches.

A limiting factor here, though, is, of course, availability of experts. Given there are often difficulties in finding experts we may not be actually able to find sufficient numbers to use a broad approach. A possible solution to this availability problem is to make compromises in quality of expertise – perhaps a less experienced person may suffice (and in some cases might be a better choice in terms of, for instance, motivation, than a more established expert). Further, someone who is less skilled with the response mode than ideal could be given appropriate training.

A broad approach to elicitation will often mean that experts will not have to be interviewed in person and so can make their contributions at their own convenience in terms of time and place. This may mean that experts are more willing to take part. On the other hand, a deep approach might be more involving for the experts and thus may be better for keeping them on board and trying hard. Again a combination of the broad and deep approaches might be ideal (i.e. poll many remotely, interview a few in person), if it can be achieved.

Note that there may be diminishing returns on the number of experts used in an elicitation. Aspinall (2010) comments regarding expert risk elicitation:

> *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)

Another point regarding the question of 'how many experts?' concerns the goal of creating heterogeneity of opinion. As Meyer and Booker (2001) note:

> *Having less than five experts reduces the chances of providing adequate diversity or information to make inferences.*
>
> (p. 88)

A.2.4.3. Principles and practice of constructing expert groups

In sum, we need to identify experts of acceptably high quality (or trainable to the required standard)—of sufficient number and diversity—persuade them to participate, and give them incentives to contribute fully. These steps will mostly be sequential but iterations may be required; for example, persuasion and/or incentives may have to be repeated periodically to keep experts on board.

Principles

- In most cases, if it is principally quantitative judgements that are to be elicited, the more experts the merrier, with a minimum of five.
- Heterogeneity of opinion is preferable to consensus.
- Avoid creating the conditions that might lead to groupthink, e.g. picking groups of experts who all know each other well, or forming interacting groups around a leading figure in the field.

Practice

- Snowballing may lead to groups who all know each other and/or hold similar opinions, so try to set up 'adversarial collaborations', or at least ask experts to suggest others whom they know they disagree with.
- Freely interacting groups can produce biased judgment so either should be avoided (e.g. individual elicitation or nominal groups - Cooke and Delphi methods, respectively) or facilitated carefully by someone who is skilled at minimizing the effects of group biases (e.g. the Sheffield method). Note, however, that interacting groups may be useful for reasons peripheral to the central goal of EKE, such as motivating and retaining.
- It is recommended that experts give reasons for their judgments (see e.g. Bolger & Wright, 2011), and that, in the case of behavioral aggregation with nominal groups, these reasons are fed back to other experts
- experts, so may be complimentary to individual elicitation or nominal group techniques.If important for experts to interact and if heterogeneity is perceived to be low, consider creating heterogeneity through devil's advocacy. If this is not practical, we propose simply asking experts to consider opinions that are not represented in the group and/or formally noting the lack of dissent.

A.2.5. Conclusions

Although each project will be different, and it is essential to tailor selection and recruitment to individual project requirements, it is possible to identify a number of general principles and related 'good practices'. Some of the key goals are:

- Analyse tasks (both the experts' and the project's).
- Research the expert.
- Maximise both quantity and diversity of expertise.

- Maximise factors leading to process gain while minimising those leading to process loss.
- Value intrinsic motivation over extrinsic.

In identifying experts, the following steps need to be taken to ensure selection of relevant and high-quality experts.

- Define expertise and consider variations in its extent and quality.
- Select experts with appropriate expertise, both substantive and normative or select experts with appropriate substantive expertise and train them in the response mode or;
- Select experts with appropriate substantive expertise and use a response mode with which the expert is familiar (e.g. elicit verbal probability terms then map them on to a numeric scale, although—research shows that this is difficult to do reliably).
- Select experts with expertise appropriate to project.
- Where possible collect a 'CV' from prospective experts.
- Try to recruit a large a sample of experts as possible (e.g. use snowballing, set up 'adversarial collaborations' or at least ask people to suggest others whom they know they disagree with).
- Ask Working Group members to fill in a relevant expertise grid and then invite the experts identified to provide the names of additional experts.

Consideration of elicitation process

- If important for experts to interact and if heterogeneity is perceived to be low, create heterogeneity through devil's advocacy.
- Think of the elicitation method that will be used, in order to select the appropriate types (taking into account substantive vs. normative expertise) and number of experts.

It often helps to explain to the expert why he or she has been chosen and to provide details about the elicitation. In this way the expert knows the kind of exercise for which he or she selected and the reason for selection, and is better able to determine whether or not if he or she is knowledgeable. In addition, and if possible, mention a name known to the expert, so that he or she is more connected to you and your request.

In summary, expert selection involves the following steps (see part II, chapter 6 for specific examples):

- Once the question or questions to be answered have been defined to an acceptable level of specificity by the Steering Group, the Steering Group should identify the essential and desired characteristics of experts and build up **profiles of experts** who may be able to answer these questions (profile matrix, Table 9);

- Identify **types of expert** who may have suitable profiles;

- Create expert information matrix (Table 10)

Identify **potential experts** (long list) based on suggestions from Working Group members, EFSA expert database, institutional contacts and the internet.

- **Contact experts** by email invitation, letter or telephone (example in part II, section 4..3) and mention the following"

    - EFSA has an important problem and you have been identified as expert.

- Here is a brief overview of the problem.
- We are currently screening for experts.

If a positive response to contact letter is received then send out the **expertise questionnaire** (Table 11 for more informed selection)

- Invite experts to provide the names of additional experts.
- If an expert invites another, then ask the expert to complete the GEM for the nominee (Table 6).

SG agrees on expert shortlist and determines expertise indicators

- **Invite the experts** to participate in the elicitation exercise by sending them an invitation letter (example in part II, section 5.1) comprising:
  - What: inform on the nature of the problem, and the motivation for the elicitation, as well as the conditions of the elicitation exercise (e.g. type of elicitation exercise, workload and compensation).
  - When: provide date.
  - Why: reasons why the expert was selected.
- **Send out detailed information** letter—this will normally be sent out with the invitation letter outlined in (5) above but may also be sent out or after the expert has agreed to participate (for example if there are confidentiality issues). This letter should include:
  - When: date, location, venue of elicitation.
  - Procedure, agenda of meeting, incentives (getting paid/reimbursed).
  - Details on: model, variables, parameters that will be asked about,
  - Constitution of the expert group (in short).
  - Assurance of confidentiality.
  - Ask for available additional information (data) that experts may have/want to share in advance (if relevant).
  - Ask about concerns that experts may have.
- **Provide feedback** (some examples are given in part II, chapter 6) and exit questionnaire

A.3. PRINCIPLES AND PRACTICE OF ELICITING A PROBABILITY DISTRIBUTION

In this chapter we discuss how to elicit a probability distribution for an uncertain quantity. We suppose assume that this quantity is required as an input in a risk model and has a true value that, in theory, could be measured, but is uncertain to us. We consider eliciting from a single expert only, and for the most part consider a single uncertain quantity only. The case of joint probability distributions for multivariate variables is discussed briefly at the end of this chapter. We suppose that the elicitation will be conducted face to face, but also briefly discuss alternatives at the end of the chapter.

A.3.1. Elicitation roles

In addition to the expert(s), whose beliefs we wish to elicit, there will be a **elicitor**, also often referred to as a facilitator, who conducts the elicitation exercise. The elicitor must be very familiar with the elicitation process, and understand basic probability theory. Currently, there is little software available for implementing probability distribution elicitation methods, and elicitors may occasionally have to deal with situations that are not easily handled by existing packages (for example, an expert may judge that two uncertain quantities should not be modelled independently). The elicitor should also have knowledge of the various biases and pitfalls in probability elicitation that have been analysed in the psychology literature. It will also help if the elicitor has some knowledge of the subject matter for the elicitation, to aid communication with the expert.

A.3.2. Representing a probability distribution

As a running example, suppose we wish to elicit an expert's beliefs about the proportion of dairy cows in a particular region that are infected with a disease. We denote this uncertain quantity by X, and our objective is to elicit a probability distribution for X from the expert. There are different ways to represent a probability distribution. We discuss three ways, all of which can play a role in a single elicitation method. The first such way is the **cumulative distribution function (CDF)**, which we denote by $F_X(x)$, and define as

$$F_X(x) = P(X \leq x),$$

the probability that the uncertain quantity X is less than or equal to some particular value x. An example CDF is shown in Figure 17.

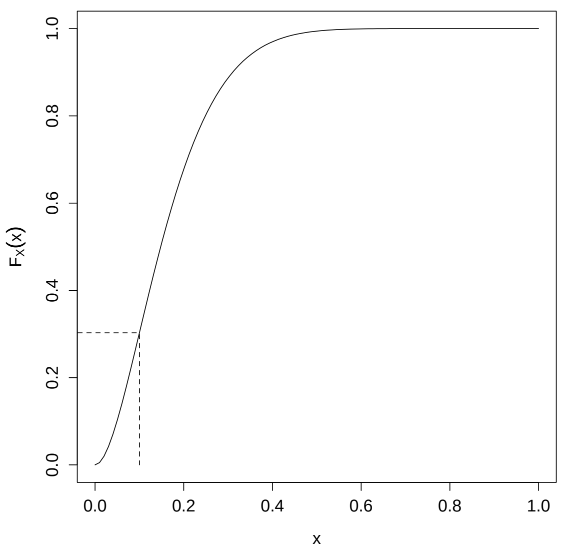


Figure 17: An example CDF. The dashed line indicates that the probability that X will be smaller than 0.1 is approximately 0.3

Reading off the graph (Figure 17), we can see that, for example, the probability that *X* will be smaller than 0.1 (i.e. 10 %) is approximately 0.3.

Specifying the probability for all possible choices of *x* completely determines the probability distribution of *X*. Hence, the task of eliciting a probability distribution from an expert can be thought of as asking an expert to specify the complete CDF: the expert's probabilities $P(X< \leq x)$ for all possible *x* (in this example, all values of *x* between 0 or 1).

A more helpful way to visualise a probability distribution (for a continuous uncertain quantity) is using the **probability density function (PDF)**, which we denote by $f_X(x)$. The value of a PDF at a single value of *x* is not particularly informative, but the probability that *X* lies between two values *a* and *b* is the area under the curve $f_X(x)$ between $x = a$ and $x = b$. An example PDF is plotted in Figure 18.

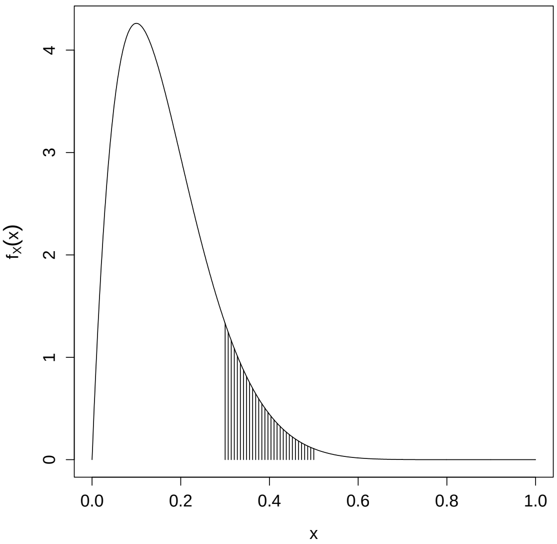


Figure 18: An example PDF. The probability that *X* lies between 0.3 and 0.5 is the area of the shaded region: the area under the curve between these two points

In this example the probability that *X* is between 0.3 and 0.5 is the area under the curve between these two values, which is approximately 0.1. The PDF in this example corresponds to the same probability distribution shown in Figure 17, but reveals judgements that are not immediately apparent from looking at the CDF. The PDF is greatest for values of *X* around 0.1, indicating where the most probable values of X are, but the non-symmetrical shape represents a belief that *X* is more likely to be greater than 0.1 than less than 0.1.

A third way to specify a distribution is to choose one of the standard families of distributions, with particular choices of parameters. For example, if we say that *X* has a Beta(*a*,*b*) distribution, this corresponds to the PDF

$$f_X(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1},$$

for values of *x* between 0 and 1, and $f_X(x) = 0$ otherwise. (This makes the Beta family of distributions particularly suitable for representing uncertainty about quantities that must lie between 0 and 1, such as proportions.) No simple formula exists for the CDF of the Beta(*a*,*b*) distribution though various statistical packages will calculate it. Figures 17 and 18 show, respectively, the CDF and PDF of the Beta(2,10) distribution, i.e. the Beta(*a*,*b*) distribution with $a = 2$ and $b = 10$.

The standard families (that one would find in a textbook on probability, or in a statistical software package) are not always adequate for representing uncertainty. In Figure 19, we show an example of a more complicated PDF, representing a judgement that *X* is most likely to be 'small', but is more likely to be 'high' than 'moderate'.

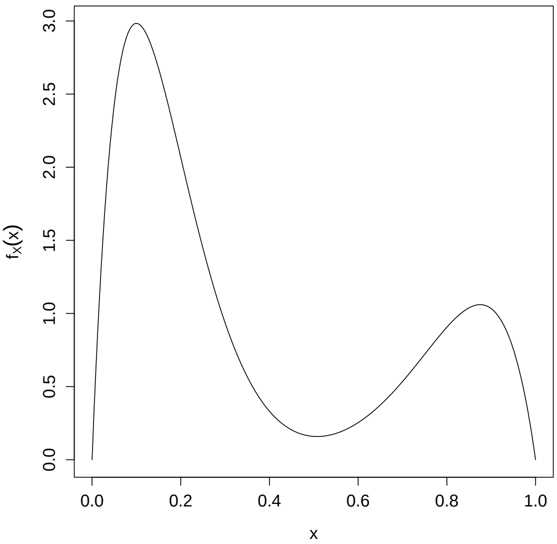


Figure 19: A PDF that does not correspond to one of the 'standard' parametric families of probability distributions

A.3.3. The strategy for eliciting a probability distribution.

In theory, we could ask experts to (1) specify the CDF, (2) specify the PDF or (3) choose a standard distribution and tell us the parameter choices. We immediately rule out option 3; this is obvious from looking at the equation for the Beta(a,b) PDF; it would be hopeless to provide experts with the equation and ask them to tell us their values of a and b.

In theory, instead of asking for parameter values, we could ask for **moments**, such as the expectation and variance of the uncertain quantity. The parameters can then be inferred from these. However, moments are difficult to judge. Consider, for example, the Beta(2,10) distribution plotted in Figures 17 and 6. With training, an expert should be able to understand and interpret the plots of the CDF and PDF, but, even after examining these plots, the expert may have some difficulty understanding the concepts of expectation and variance for this distribution (in particular, why the expectation is 0.17 and the variance is 0.01). An expert may also confuse the concept of expectation with that of an average.

> **Recommendation:** We recommend *against* using any elicitation method that involves directly asking the parameters of a distribution, or moments of a distribution such as a mean or a variance.

Conceivably, an expert could draw a PDF, as a plot of the PDF is easy to interpret, but there would be practical difficulties converting the drawing into a useable form, as well as concerns about how precisely the function is drawn. We are not aware of available software for implementing such an approach, and we do not discuss this further. This leaves option 1, but there remains the problem that an expert cannot provide the *entire* CDF. The solution, common to most elicitation methods, involves a combination of all three representations of a probability distribution.

1. The expert provides values of the CDF at a small number of points.

2. The elicitor chooses a probability distribution from a standard family of distributions

3. The elicitor (may choose to) plot the PDF corresponding to the choice in step 2, to show the result of the elicitation to the expert.

A.3.4. General principles

There are many choices regarding precisely what is asked for in step 1, as well as how to choose a distribution in step 2. Some general principles to consider are the following.

- An expert is likely to find the elicitation process difficult. An expert may have decades of subject matter experience, and yet never have had to make probabilistic judgements before.
- Experts are susceptible to biases. Two mathematically equivalent ways of asking for the same judgement may not result in the same response.
- It is quite possible that experts will have misconceptions of probability elicitation, for example:
    - that they will be expected to provide a point estimate of *X*, rather than to provide a distribution to describe their uncertainty about *X*;
    - that if they are very uncertain about *X*, it will not be possible to elicit a distribution to describe their uncertainty; elicitation 'does not work' if we are too uncertain.

Recommendation: As has already been stressed, training is important, and the elicitor should first give the expert training in eliciting probability distributions. The training should cover the idea of representing uncertainty about a fixed, unknown quantity using a subjective probability distribution, a discussion of potential psychological biases when making probability judgements and a practice elicitation exercise.

Recommendation: The elicitor should challenge any misconceptions, and explain that *the objective is to represent the expert's uncertainty about X*, not to obtain a spuriously precise estimate of *X*. Experts should be reassured that if they are genuinely very uncertain about *X*, the elicited distribution will reflect their uncertainty. Even if no objections are raised, a discussion of these issues can help the experts to better understand the process. Presenting an example such as the following can be helpful.

Example: *Suppose we wish to elicit an expert's beliefs about the European Central Bank's interest rate in one year's time. Two probability distributions are plotted in Figure 20. The top plot represents a belief that the rate will almost certainly be 1.5 %, and, for example, rates higher than 1.75 % or lower than 1.25 % are impossible. The bottom plot represents a belief that the interest rate is equally likely to lie anywhere between 0 % and 10 000 %. Realistically, no-one would claim to be as certain as that implied by the first distribution, or as uncertain as that implied by the second. The elicitation process can be thought of as establishing where an expert's uncertainty sits between these two extremes. We can be very uncertain about a quantity, yet instantly recognise probability distributions that do not represent our beliefs; no-one would say "I do not know much about interest rates, so I do not know whether or not either of these distributions represents my beliefs."*

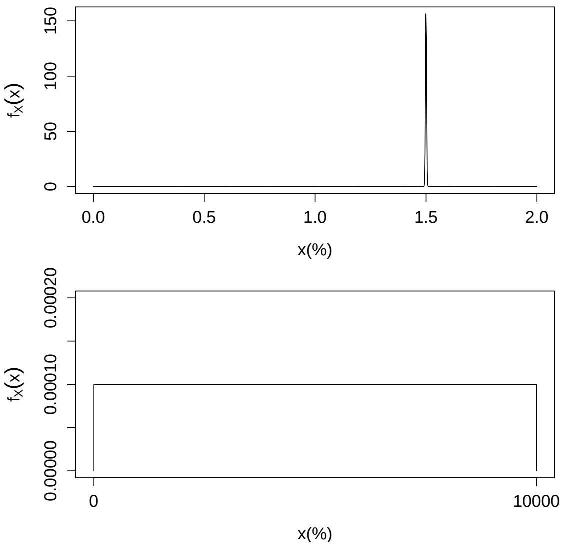


Figure 20: Two probability distributions for the European Central Bank's interest rate in one year's time. The top distribution represents near certainty; the bottom distribution represents extreme uncertainty. We can easily recognise distributions that do not represent our beliefs well

A.3.5. What questions to ask

A.3.5.1. Eliciting points on the CDF: the options

There are two ways to elicit a single point on the CDF. The first is to choose a value of *x*, and then ask for the probability (i.e. choose a value on the x-axis, and ask for the corresponding value on the y-axis). Such a question might be

What is your probability that X *is less than (or equal to) 20?*

or

What is your probability that X *lies between 20 and 40?*

This is sometimes referred to as a **fixed interval** approach and is illustrated in Figure 21.

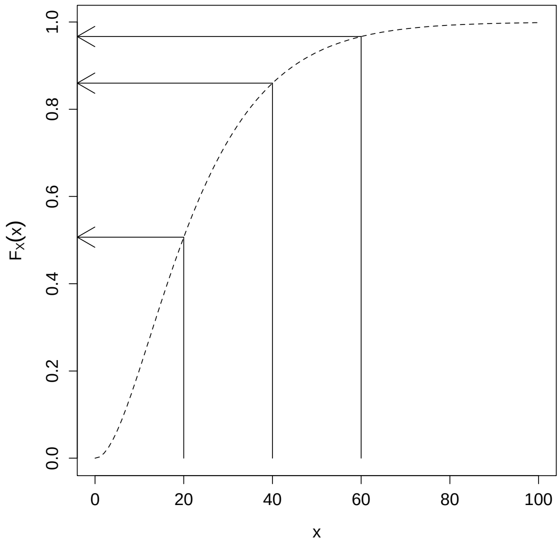


Figure 21: The fixed interval method. The elicitor chooses some values for the uncertain quantity on the x-axis (20, 40 and 60 in this example), and the expert provides the corresponding probabilities on the y-axis

When eliciting a probability, one visual tool that can be used is a sliding coloured bar, as shown in Figure 22. The expert specifies a probability by choosing what proportion of a bar to colour in. In the plot, the elicited probability can be interpreted as saying that the following two events are equally likely:

1. The true value of X lies between 0 and 40.

2. A point in the rectangle in Figure 22 is selected at random, and is found to lie in the yellow section.

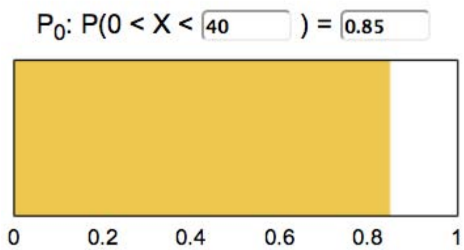


Figure 22: Representing a probability using a sliding bar. The expert chooses what proportion of the a rectangle to colour yellow

The second way to elicit a point on the CDF is to choose a value of the probability $P(X \leq x)$, and ask for the corresponding x value. Such a question might be

Suggest a value x *such that you are 90 % sure that* X *will be less than* x.

or

Suggest two values x_1 *and* x_2 *such that you are 90 % sure* X *will lie in the interval* (x_1, x_2).

This is sometimes referred to as a **variable interval** approach. Here, we are eliciting **quantiles** rather than probabilities. For example, if we ask for x such that $P(X \le x) = 0.9$, we are asking for the 0.9 quantile. This is illustrated in Figure 23.

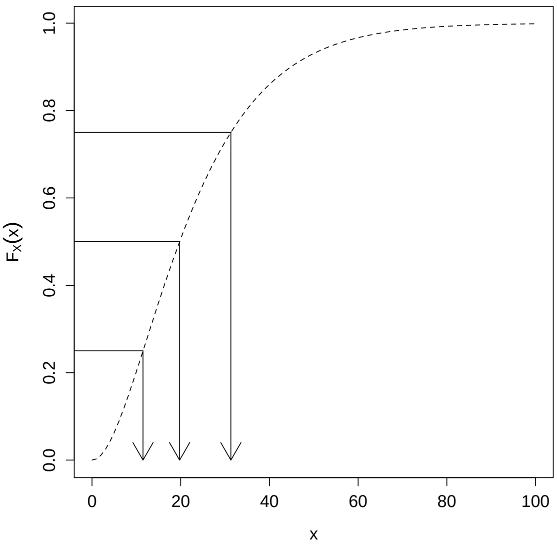


Figure 23: The variable interval method. The elicitor chooses some probabilities on the y-axis (0.25, 0.5 and 0.75 in this example), and the expert provides the corresponding parameter values on the x-axis

One special case of the fixed interval approach is the **roulette method**. This method was suggested by Gore (1987), and reviewed in Johnson et al. (2010). In this method, the expert is asked to distribute a number of chips into some bins, with the probability of X lying in a particular bin interpreted as the proportion of chips allocated into that bin. In this method, experts can see the shape of their distribution building up as they allocate the chips. We illustrate this in Figure 24.

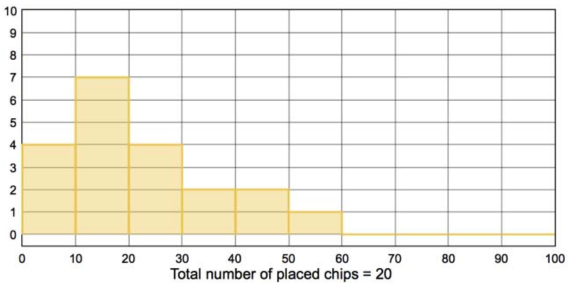


Figure 24: The roulette elicitation method. We interpret, for example, the four chips in the bin [0,10] as a probability of 4/20 that X will lie in that bin

There is no evidence to favour one method definitively over the others. However, it is unlikely that experts will find all methods equally easy to use. At the training stage, experts can be shown several methods, and can experiment to see which method they prefer. There are limits to how much flexibility the experts can be given; there are methods that either will not yield a probability distribution or are likely to give poor results, but, otherwise, the easier the experts find the method, the more likely they are to engage with the elicitation process.

> **Recommendation:** At the training stage, and if time allows, the elicitor should present the experts with alternative elicitation methods which the elicitor regards as appropriate to the elicitation task and the expert's background, to see which method the expert prefers (although the elicitor should be prepared to choose a method if the expert has no strong preference).

A.3.5.2. Practicalities

Within each method, there are choices to be made about how many and what probabilities/quantiles to elicit. If we are fitting a standard parametric distribution to the expert's judgements, then, as a minimum, we will need one distinct point on the CDF per parameter in the distribution (typically two, for most standard univariate probability distributions). Usually, we will need more than the minimum, so that we can identify which family of distributions is most suitable (e.g. whether to choose a symmetrical or a skewed distribution).

> **Recommendation**: When using the fixed or variable elicitation methods, ask for a minimum of three (and at least one more than the number of parameters in the chosen family) probabilities/quantiles. With suitable questions, this will usually establish what values of *X* the expert believes to be most likely, the magnitude of the expert's uncertainty about *X* and the shape of the expert's distribution.

A.3.5.3. Using the fixed interval and roulette methods

To use either the fixed interval or roulette method, it is first necessary to establish the expert's range for *X* (for example, to choose the locations of the bins in the roulette method). We denote this range by $[L,U]$.

> **Recommendation:** A useful way to think about this range is to imagine a plot of the expert's PDF, and to set *L* and *U* to be appropriate limits for the x-axis: what is the smallest range such that we can see the PDF clearly? The expert should be certain that *X* will lie inside this range, but the range should not be so wide that the shape of the PDF cannot be seen.

We are not aware of any research into the best probabilities to elicit, but, as an example, the MATCH elicitation tool (see section A.3.9) has defaults of $P(L < X < 0.25(U-L)$, $P(L < X < 0.5(L+U))$ and $P(0.75(U-L) < X < U)$.

> **Recommendation:** For the roulette method, we suggest 10 bins spanning the interval $[L,U]$. Unless *L* and *U* represent the physical limits of *X*, the end bins should be left empty; placing chips in the end bins suggests that the limits are not small/large enough. We suggest allocating 20 chips in total, but any number can be used, as it is the proportion of allocated chips that gives the probability.

A.3.5.4. Using the variable interval method

The variable interval method can be used without first setting a range, but we recommend eliciting the expert's range in any case. This will be necessary if using graphical methods for eliciting quantiles, and can also help reduce the risk of anchoring on a single value. In particular, beginning with both *L* and *U* means that subsequent judgements are anchored both sides, so that they tend to cancel each other out.

Some common options are

- Elicit the quartiles: the median (0.5), lower quartile (0.25) and upper quartile (0.75).
- Elicit the tertiles (0.33 and 0.66 quantiles).
- Elicit the median, the 0.05 quantile and the 0.95 quantile.
- Elicit the median, the 0.17 quantile and the 0.83 quantile.

When choosing which quantiles to work with, two issues to consider are interpretability and the risk of overconfidence. An attraction of using the quartiles or tertiles is that the process of specifying these values can be thought of as dividing the interval $[L,U]$ into equally likely regions. For example, suppose the expert has provided the following values for an uncertain parameter X:

- $L = 0$;
- $X_{0.25} = 10$;
- $X_{0.5} = 20$;
- $X_{0.75} = 40$;
- $U = 100$.

This can be interpreted as follows. Suppose the expert can bet on X lying in one of the four intervals [0,10], [10,20], [20,40] or [40,100]. If the expert chooses correctly, he or she gets a reward, but there is no penalty for choosing incorrectly. If the expert's quartiles are 10, 20 and 40, the probability of X lying inside any of these intervals is 0.25 = 25 %, and the expert should have no preference for which interval to bet on. If, on the other hand, the expert would rather bet on [10,20] than on [0,10], this is interpreted as a judgement that $P(0 \leq X \leq 10)$ is less than $P(10 \leq X \leq 20)$, so that 10 and 20 cannot be the expert's lower quartile and median. A similar interpretation can be given if the tertiles are elicited.

Various studies have investigated which quantiles are good or bad to elicit, in terms of reducing overconfidence. A review is given in O'Hagan et al. (2006, pp. 101–104), with the conclusion that it is better to ask for 'moderate' quantiles such as the quartiles or tertiles than more extreme quantiles such as the 0.05th and 0.95th. The 0.17th and 0.83rd are a compromise, and can be thought of as asking for a 'two-to-one' interval: X will be twice as likely to lie in inside these quantiles as outside them.

An experiment by Soll and Klayman (2004) compared asking directly for an 80 % interval with first asking for the 0.1 quantile, and then for the 0.9 quantile. Although this method is also asking for an 80 % interval, they found that it reduced overconfidence compared with the previous approach. The authors conjectured that asking for low and high values separately encouraged the participants to assess their uncertainty more thoroughly.

> **Recommendation:** We suggest asking for moderate quantiles, and asking for quantiles individually, rather than intervals.

A.3.6. Fitting and feedback

A.3.6.1. Fitting a distribution

Having elicited some points on the CDF, the next step is to propose a complete distribution to represent the expert's beliefs. For illustration, we suppose the expert has provided the following:

- $P(X \leq 0) = 0$;
- $P(X \leq 5) = 0.05$;
- $P(X \leq 20) = 0.5$;
- $P(X \leq 60) = 0.95$;
- $P(X \leq 100) = 1$

Perhaps the simplest option is to choose a piece-wise uniform distribution, obtained by connecting the points on the CDF with a straight line. We illustrate this in Figure 25.

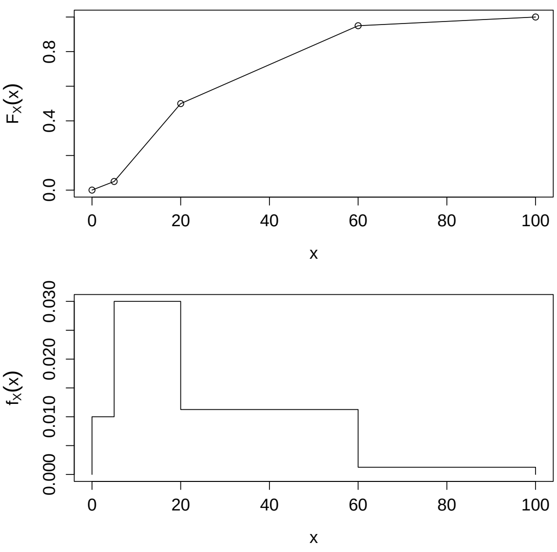


Figure 25: Fitting a piece-wise uniform distribution to the elicited points on the expert's CDF

Alternatively, we can choose a parametric family of distributions, and use a least squares procedure to choose the parameter values. For example, we could fit a gamma distribution to the points on the CDF. If we denote the CDF of a Gamma(shape = a, rate = b) random variable by $F_X(x;a,b)$, we find a and b to minimise

$$\{F_X(5;a,b) - 0.05\}^2 + \{F_X(20;a,b) - 0.5\}^2 + \{F_X(20;a,b)0.95\}^2 + \{F_X(100;a,b) - 1\}^2$$

This must be done using a numerical optimisation routine. In the example, this is minimised (approximately) at a = 2.2 and b = 0.09. We illustrate this in Figure 26.

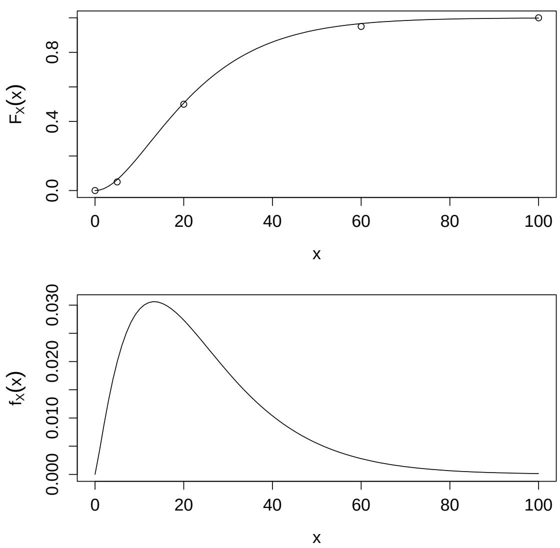


Figure 26: Fitting a gamma distribution to the elicited points on the expert's CDF

There is no formal justification for using least squares as opposed to some other criterion, but it can usually be relied on to identify a suitable fit, if one exists.

A.3.6.2. Choice of distribution

Three scenarios that often occur in univariate elicitation are the following.

1. X is a proportion, and so must lie between 0 and 1. The expert may judge, for example, that X could be close to 1, but obviously cannot exceed 1. A Beta distribution will usually be suitable for representing uncertainty about a proportion, and will always force X to lie between 0 and 1.

2. X has a lower limit, and may or not be close to this limit. Unlike the first scenario, the expert will not have beliefs of the form "X could be very close to an upper limit U, but X cannot *exceed* U". In this case, a gamma distribution may be suitable, as this will only impose a strict lower limit for X. An alternative to the gamma in this case is the log-normal distribution.

3. The expert's PDF will be symmetrical about his or her median value. A Beta distribution can be used if the expert wishes to impose lower and upper and limits, but alternatives are the normal distribution and the Student t-distribution.

Examples of each distribution are plotted in Figure 27.

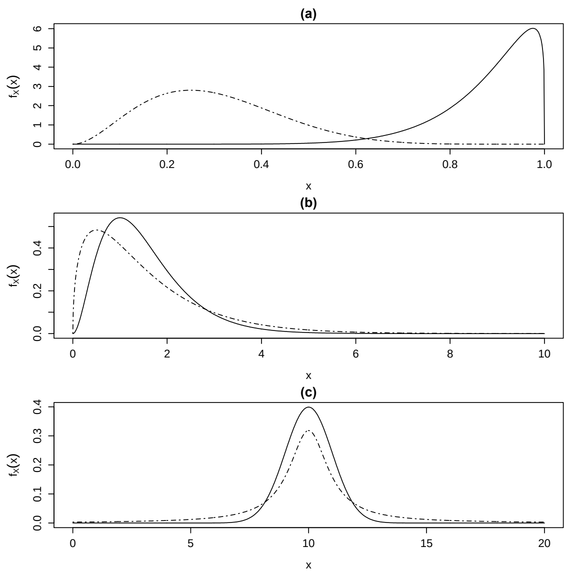


Figure 27: (a) two beta distributions, (b) two gamma distributions, (c) a normal distribution (solid line) and a Student t-distribution with two degrees of freedom (dashed line)

Recommendation: The elicitor should have the means to fit distributions for these three scenarios.

In some cases, standard families of distributions may not be flexible enough to fit an expert's beliefs. These can arise when there this is additional structure not represented by the risk model; typically judgements of the form "If scenario A occurs, then I expect X to be small, but if scenario A does not occur, then I expect X to be large". This can be handled by expanding the risk model and then eliciting three distributions:

- a distribution to represent uncertainty about X if it is known that scenario A has occurred;
- a distribution to represent uncertainty about X if it is known that scenario A has *not* occurred;

- a distribution to represent the proportion of occasions in which scenario A will occur (or a probability that scenario A will occur if this is a one-off event).

A.3.7. Checking the fitted distribution

Whatever distribution we fit, we will be assuming additional judgements that the expert has not made. We can check whether the assumed distribution is acceptable to the expert in one of two ways:

1. **Feedback**. We provide additional summaries from the fitted distribution, and invite the expert to comment.

2. **Overfitting**. We ask for extra judgements, which can be compared to those from the fitted distribution.

For example, in the Gamma(2.2,0.09) distribution, the 0.33 and 0.66 quantiles are 15 and 28 respectively, so we could feed these back to the expert, and see whether the expert is happy with the implication that the intervals [0,15], [15, 28] and [28,100] all have the same probability of containing *X*. Alternatively, we could ask the expert to provide three equally likely intervals, and compare them with the intervals from the fitted Gamma distribution. The risk with feedback is that the expert may simply accept what is suggested, without considering the additional judgements carefully. The risk with overfitting is that the expert may develop ‘elicitation fatigue’ from having to think about too may judgements.

Recommendation: The suitability of the assumed distribution must be checked, using either feedback or overfitting. If the expert is providing several distributions in one sitting, the easier option of feedback may be preferable.

Recommendation: We also suggest feeding back some extreme quantiles from the fitted distribution, e.g. the 0.01 and 0.99 quantiles. Questions the elicitor may ask are: “Would you be surprised to discover that *X* was this large/small?” or “Can you think of a reason why *X* would be this large/small?”. This may encourage the expert to consider whether he or she has been overconfident in the first instance.

At this stage, the expert may choose to modify his or her initial judgements, in which case the elicitor will need to re-fit the distribution, or the elicitor may need to try a different parametric family.

A.3.8. Dependence and multivariate elicitation

Risk models will usually have more than one uncertain input, and we must consider how to represent uncertainty about more than quantity *jointly*. The simplest option is to assume independence between the inputs, and elicit distributions for each input separately. If we have two uncertain quantities *X* and *Y*, we would judge them to be dependent if, after learning the true value of *X*, we would change are beliefs about *Y*, and vice versa; otherwise they can modelled independently.

As an example, suppose that *X* and *Y* are the proportions of dairy cows infected with a disease in two different regions, well separated so that we do not believe the value of *X* could *physically* influence the value of *Y*. Suppose also that our current belief is that the disease is not very infectious, and prevalence in any region should be low. If we then discover that *X* is much larger than expected, we may then wish to revise judgements of *Y* upwards, as our original assumptions now appear less valid.

Eliciting joint probability distributions to allow for dependence is a more complex task, and we do not discuss it any further here. Suggested reading is Kurowicka and Cooke (2006), O’Hagan et al. (2006) Daneshkhah and Oakley (2010) and O’Hagan (2012).

A.3.9. Elicitation software

A detailed review of elicitation software is reported in Devilee and Knol (2011). The authors have used the term "elicitation" in a broader sense, so that not all packages reviewed are designed for eliciting probability distributions. In their conclusions, they note that

> *... there is a lack of software that lowers the costs of expert elicitations in terms of travel and organising time and consequently money ... there is a need for more inexpensive software that enables a fast consultation of experts at different locations and at times that suit them ... elicitation software that can be used online or sent by email can provide a very fruitful additional contribution.*

Two recently developed software packages that can be used online, with elicitor and experts in different locations, are

- The MATCH Uncertainty Elicitation Tool, available at http://optics.eee.nottingham.ac.uk/match/uncertainty.phpThe Elicitator, available at http://elicitator.uncertweb.org/

A.3.10. Imprecision

It must be recognised that, in practice, all elicited judgements, and in particular the finally chosen probability distribution, are imprecise. The expert cannot make arbitrarily precise judgements of probability, for instance declaring that his or her probability that X is less than 1 is 0.23076. In practice, the expert will certainly not be able to discriminate between 0.23076 and 0.23. A related point is to remember that the expert's judgements will be influenced by the way that questions are posed (see the discussion of psychology in part I, chapter 2). Had the elicitation been performed with the questions phrased differently, in a different order, using an alternative protocol, or even using exactly the same protocol on a different day, the expert would be likely to give different answers. Good elicitation practice, as in the exemplar protocols described in part II, chapter 6, is intended to minimise the imprecision in elicited distributions, but imprecision cannot be eliminated.

Consequently, it is important for the client (i.e. the EFSA Steering Group) to make allowance for the imprecision in elicited probability distributions. The sensitivity of the resulting risk assessment should be evaluated. Although methods exist (see Oakley and O'Hagan, 2007) to assess elicitation imprecision formally, in practice it may be simpler just to evaluate sensitivity empirically by recomputing the risk assessment using some alternative distributions. These should deviate from the elicited distributions in ways that reflect reasonable variations of the expert's elicited judgements. (For instance, the distribution may be shifted to have a mean of 0.53 when the elicited distribution has a mean of 0.5.)

A.4. Principles and practice of handling multiple experts – Interactions and aggregations

A.4.1. Introduction

It is generally agreed that an elicitation procedure will always benefit from the involvement of more than one domain expert (see also section A.4.5). Group assessments of uncertainty have the final goal of obtaining a single distribution for each variable of interest. Experts might be asked to first discuss and share knowledge, and then provide the decision maker with a consensus distribution; or individual opinions might be first collected and then aggregated. As mentioned in the previous chapter, the presence of a elicitor is essential during a face-to-face elicitation procedure; hence we will further assume that during such elicitation a elicitor is always present.

Sometimes the individual expert opinions will be needed as such and the heterogeneity among experts will be used as means of quantifying and comparing sources of uncertainty about models' parameters (Hoffmann et al., 2006). Some argue that the range of expert opinion may be at least as valuable as the aggregate assessment (Keith, 1996).

Nevertheless, in most situations it is necessary (and advisable) to obtain a single distribution that captures the uncertainty derived from the expertise of several experts. Much of the literature on expert elicitation focuses on how to aggregate expert judgement to achieve a reliable consensus on distributions of interest. Broadly speaking, two approaches have been taken to aggregation: behavioural and mathematical.

The behavioural aggregation approach is based on interaction between the experts from the selected panel. Through this interaction/conferencing, the experts are expected to develop a shared understanding of the issues at hand. Some of the behavioural approaches expect a commitment of all experts to the same opinion. Other methods aim for a distribution that reflects to some extent all agreements and disagreements between experts. The experts do not necessarily commit to the same opinion, but rather contribute to it. Behavioural aggregation approaches are further detailed in section A4.2. We consider only face-to-face elicitation in the presence of a elicitor.

In contrast, in a mathematical aggregation approach, the experts do not interact. Each expert's distribution is elicited independently of the others. The individual distributions are then mathematically combined (see a more detailed discussion in section A4..3).

Combinations of the mathematical and behavioural aggregation approaches are also possible and they will be touched upon in the last section of this chapter.

More details, reviews and references for the above-mentioned approaches can be found in Cooke (1991), Clemen and Winkler (1999), French and Rios Insua (2000), Goosens and Kelly (2000), Garthwaite et al., (2005), Jacobs (1995) and O'Hagan et al. (2006).

How an elicitation exercise is best conducted can be critical to a decision process, as the differences in efficacy and robustness of the elicitation methods can be substantial. Although this research field has been increasingly active during the last 30 years, and various methods for combining expert opinion are available, to date there is no overall methodological consensus, let alone a "gold standard".

Empirical evidence and comparisons are available and in contradiction. In some studies behavioural aggregation is outperformed by mathematical aggregation (e.g. Lawrence et al., 1986), whereas in other studies the opposite finding is presented (e.g. Hastie, 1986). Some of the studies under investigation consider only certain (simple) methods of mathematical aggregation, or mainly use almanac questions and/or students as experts (e.g. Seaver, 1978). Others show that more complicated methods of mathematical aggregation might outperform simpler methods (Cooke, 2013a), however it has been argued that the additional effort required does not justify the use of such methods. When a method has been used long enough and the studies have been documented, statistical analysis of the results is possible and advisable.

Scientists agree to disagree when it comes to the choice of the aggregation method. This modus vivendi is not accepted for the sake of comfort, but rather is driven by the level of complexity, subjectivity and context dependency of the problem.

In many decision problems, where the risks are not life-threatening, or they do not relate to substantial losses, methodologies such as behavioural aggregation help to develop a shared understanding of the problem context and, in most instances, a commitment to the chosen action. Most behavioural approaches tend to encourage convergence of views. In risk analysis this may be extremely dangerous. Good risk management should keep track of potential threats and maintain an overview of all opinions by maintaining disagreements.

> **Recommendation:** Always take into account the context of a study and design the elicitation and aggregation processes such that it fits that particular context. Making a generalisation that mathematical aggregation is generally better or worse than behavioural may blind us in meeting the needs of a particular study.

A.4.2. Behavioural aggregation

In most behavioural aggregation approaches, the experts produce a consensus probability distribution. The elicitor aids the process of interaction and debate. The main goal of these approaches is the achievement of shared understanding and knowledge about the phenomena represented by the parameter elicited. Unanimity with respect to the form of the common probability distribution is an indication that all background information is assimilated, and similarly interpreted, by all experts.

Nevertheless, other practitioners' goal is for the experts to agree to a distribution that reflects the extent of agreement and/or disagreement between them at the end of the process. They are supposed to agree on what an *intelligent observer* should believe about the quantity of interest (see discussion on group elicitation in part II, section 6.2.4). Implicitly, the experts are expected to agree on a reasonable weighting of the various opinions. If they do not agree at the end of the process, mathematical aggregation may be applied, or the problem owner may simply be provided with more than one probability distribution for the quantity of interest.

Group elicitation poses more psychological, rather than mathematical, challenges. For this reason, perhaps the most demanding task is that of the elicitor. The heuristics and biases discussed in part I, chapter 2 acquire a new dimension:

- Anchoring effects might be reduced through different anchors (Sniezek,1992).
- The effects of the representativeness bias might get stronger (Kerr et al., 1996).
- The group overconfidence might increase due to "group polarisation" (Sniezek, 1992; Plous, 1993).

The elicitor should stay impartial and should react quickly when recognising potential biases.

A well-designed protocol for the elicitation could help the elicitor in what is a tremendous task. Moreover, giving constant feedback of the implications of the experts' judgements could ease the debate. Interactive software is then almost essential for the effective use of feedback.

One interactive software package that is design for eliciting experts' opinions separately and as a group is the SHeffield ELicitation Framework (SHELF) package of documents, templates and software. SHELF is delivered as a zip file that can be freely downloaded from http://www.tonyohagan.co.uk/shelf/

Recommendation: The elicitor must be able to encourage the sharing of knowledge without allowing the group to be dominated by the most confident and outspoken experts, recognise and correct potential biases, and use feedback to aid the debate.

Given the difficulty of controlling and correcting for biases in group interactions, behavioural aggregation, in which experts' interaction is reduced and controlled, is proposed. We give a short overview of such methods in section A4..4.

Nevertheless, sometimes behavioural aggregation does not succeed because experts strongly disagree. When they do, any attempt to impose agreement will promote confusion between consensus and certainty. And since the goal should be to quantify uncertainty, not to remove it from the decision process, mathematical methods of aggregation should be sought.

Recommendation: When experts strongly disagree, report the two (or more) elicited probability distributions representing the opinions of the dissenting groups, together with a mathematical aggregation of those distributions.

A.4.3. Mathematical aggregation

There are two approaches to mathematical aggregation that seem to have passed the test of time: the Bayesian and opinion pooling.

Bayesian approaches treat experts' opinions as data for the decision maker and try to develop appropriate likelihood functions to represent the decision makers' confidence in experts' judgements. Opinion pools simply weight together the experts' judgements. There are myriad models for combining opinions, expert or otherwise. Older methods, such as Delphi or the nominal group technique operate, on experts' point assessments of unknown quantities. More recent methods concentrate on the combination of expert subjective probability distributions. These methods may use an arithmetic or geometrically weighted mean or perhaps something more generalised.

A.4.3.1. Bayesian methods

In the case of Bayesian aggregation, the elicitor defines the likelihoods of the experts' judgements and treats these judgements as data for updating his or her prior belief to a posterior belief. Appropriate likelihood functions need to represent the decision maker's confidence in the experts' assessments. The decision maker needs to incorporate in the probabilistic model his or her understanding of the ability of experts to quantify their uncertainty.

The experts may be judged as independent or dependent. When dependence is considered, the model has to also incorporate the elicitor's opinion about the correlation between experts' assessments and the correlation between the experts' judgements and the elicitor's own judgements.

Taking into account all these levels of complexity adds to the elicitor's assessment burden, which becomes higher than that of any other experts. Moreover, modelling the calibration and correlation of experts' assessments poses theoretical challenges.

Many Bayesian models for combining expert assessments have been proposed in the literature, but few have been applied. Examples of applications of Bayesian methods can be found in Wiper and French (1995), Roosen and Hennesy (2001) and Szwed and van Dorp (2002).

The Bayesian approach, although flexible and appealing conceptually, remains impractical mainly because of the difficulty in developing the likelihood functions.

A.4.3.2. Weighted combinations of probabilities

In opinion pooling, a consensus probability distribution is obtained via a combination of individual distributions elicited from experts. Much has been learned about the mathematical properties of various rules for combining probabilities. Good summaries can be found in French (1985) and Genest and Zidek (1986).

General results that show how constraints on a combination rule determine its form are available. Different combination rules possess different properties, but unfortunately it is not possible to have all desirable properties in one combination rule. When one chooses a particular rule, one also accepts that some properties will be violated. The advantages and disadvantages of each rule are investigated and balanced.

One of the options is the linear opinion pooling. For a discussion of pros and cons of the linear pooling see Genest and Zidek (1986) and Cooke (1991).

We shall further restrict our attention to linear opinion pooling. Once it has been decided to use a linear opinion pool, the rule is completed by deciding on a set of weights. The weights may be simply chosen by the elicitor based on his or her judgement about the experts' expertise, usually without any operational definition for the concept of "expertise". Self weighting is sometimes proposed. Self weighting and confidence of the experts in their own ability to quantify uncertainty are often poor indicators of relative expertise (Burgman et al., 2011; Cooke 2013b). Alternatively, the equal weights may be preferred, by invoking the principle of indifference, or of equity. It has also been suggested that the weights might be defined from the social network of experts, one operationalisation being the relative frequency with which their work is cited in the literature (Cooke et al., 2008). Since the 1980s, Cooke's model has offered an alternative method of defining weights on the basis of the experts' relative performance on a calibration set of variables, whose realisations are unknown to the experts, but known to the elicitor/analyst (Cooke, 1991).

A.4.3.2.1 Cooke's Model

A proposed way of deriving weights indicating the performance of the experts is based on applying a set of principles formulated in Cooke (1991)' according to Cooke, these principles should be satisfied by any method warranting the predicate "scientific". These principles are: scrutability/accountability, empirical control, neutrality and fairness.

The group agrees on a method according to which a representation of uncertainty will be generated for the purposes for which the panel was convened, without knowing the result of this method. It is not required that each individual member adopt this result as his or personal degree of belief. A workable procedure that tries to satisfy the above principles and gives good results in practice (Cooke and Goossens, 2008) is embodied in Cooke's model. Cooke's model weighs the opinion of each expert on the basis of his or knowledge and ability to judge relevant uncertainties (translated into scores).

The weights are based on the theory of proper scoring rules and reward good calibration and high information content. Proper scoring rules ensure that the experts state their true opinion. This is important since advice during an emergency is usually equated with a huge responsibility. The pressure to be extremely cautious might influence the experts' willingness to state their true opinion.

Experts assess variables in their field of expertise for which the true values are known, in addition to the target variables. These are called seed, or calibration, variables; hence the method requires the elicitor to define seed variables.

> **Recommendation:** Define seed variables in a way that triggers in the experts the same heuristics as the target variables and such that they are representative of the variables of interest.

Recommendation: The representativeness of seed variables should always be scrutinised.

The need for seed variables implies that extra steps are necessary in the pre-elicitation stage: the identification of seed variables and a dry-run exercise to finalise the elicitation questions.

This is followed by a plenary meeting of all experts in which the issues are discussed, the study design is explained and a short elicitation exercise is carried out. This involves a small number of seed variables, typically five. Experts are shown how the scoring and combining works. Afterwards, the experts are elicited individually. An elicitation session should not exceed half a day. Fatigue sets in after two hours.

When experts are dispersed it may be difficult and expensive to bring them together. In such cases the training is given to each expert in abbreviated form. In general, it is not advisable to configure the exercise such that the presence of all experts at one time and place is essential to the study, as this makes the study vulnerable to last minute disruptions.

Let us return to the subject of "how the scoring and combining works".

In the quantile format, experts are presented with an uncertain quantity taking values in a continuous range, and they give predefined quantiles, or percentiles, of the distribution, typically 5 %, 50 % and 95 %.

Calibration is defined as the likelihood of a statistical hypothesis which is defined for each expert as:

The realisations might be regarded as independent samples from a distribution corresponding to the expert's quantile assessments.

In other words, the elicitor wants experts for whom the corresponding statistical hypothesis is well supported by the data collected from the seed variables. Although the calibration score uses the language of simple hypothesis testing, it must be emphasised that we are not rejecting expert hypotheses; rather we are using this language to measure the degree to which the data support the hypothesis that the expert's probabilities are accurate. Low scores, near zero, mean that it is unlikely that the expert's probabilities are correct.

The second scoring variable is information. Loosely, the information in a distribution is the degree to which the distribution is concentrated. Information cannot be measured absolutely, but only with respect to a background measure. Being concentrated or "spread out" is measured relative to some other distribution. Commonly, the uniform and log-uniform background measures are used (other background measures are discussed in (Yunusov et al., 1999)).

The weights are constructed as proportional to the product of the calibration and information scores.

The information is a slow function, i.e. large changes in the expert assessments produce only modest changes in the information score. This contrasts with the likelihood function in the calibration score, which is a very fast function. This causes their product to be driven by the calibration score.

Consider the following score for expert *e*:

$$w_\alpha(e) = 1_\alpha(\text{calibration score}) \times \text{calibration score}(e) \times \text{information score}(e) \, (*)$$

where $1_\alpha(x) = 0$ if $x < \alpha$ and $1_\alpha(x) = 1$ otherwise. Cooke (1991) shows that (*) is an asymptotically strictly proper scoring rule for average probabilities. The scoring rule constraint requires the term 1_α(calibration score), but does not say what value of α should be chosen. Therefore, we choose α so as to maximise the combined score of a resulting "combined" expert. Although this rule appears to discard poor assessors; it actually finds a spanning set.

There is no mathematical theorem stating that the above constructed weights outperform equal weighting or outperform the best expert. It is not difficult to construct artificial examples where this is not the case. Performance of these weighting schemes is a matter of experience. Of course, there may be other ways of defining weights that perform better, and indeed there might be better performance measures. Good performance on one individual dataset is not convincing. What is convincing is good performance on a large diverse dataset, such as the Technical University of Delft expert judgement database (Cooke and Goossens, 2008). Performance-based combinations do better (in general) than best expert and equal-weighted combinations. This conclusion was challenged as soon as the TU Delft database became available. Validation studies were performed in a few cases, where the variables of interest were later observed (e.g. real estate investment, options trading). In most cases the variables of interest are not observable on timescales relevant for the decision problem. Therefore, various forms of cross-validation have been suggested. Clemen (2008) looked at 14 studies from the TU Delft database and proposed a cross-validation technique which determined that the Cooke's performance-based combination of experts' opinions does not appear to outperform the equal-weighted combination. Lin and Cheng (2009) extended this analysis by conducting the same cross-validation tests on almost all of the studies acquired by Cooke and Goossens at the time. Their results showed that Cooke's model did maintain (only) a slight advantage over the equal weight method. Conversely, Cooke (2008) suggested that the cross-validation technique proposed by Clemen may (not intentionally) penalise the performance-based combination. Instead, Cooke presented a different cross-validation method. In 20 of 26 validation runs, the performance-based combination outperformed the equal-weighted combination. Flandoli et al. (2011) performs a similar analysis on data from five studies. Their results showed that Cooke's model provides the best indication of uncertainty. The most recent and comprehensive study (Eggstaff, 2014) indicates that Cooke′s model significantly outperforms equally weighting expert judgment.Cooke's model is possibly the most applied of all expert judgement methods. (Goossens and Kelly 2000; Cooke and Goossens, 2008; French 2011). It is implemented in the EXCALIBUR (acronym for EXpert CALIBRation) software, freely downloadable form http://risk2.ewi.tudelft.nl/oursoftware/6-excalibur

A.4.4. Mixed techniques

Behavioural and mathematical aggregation represent two separate approaches for combining the judgments of experts. A third class of approaches (discussed by Ferrell, 1985, and Rowe, 1992) are so-called 'mixed' approaches, which combine elements of these, i.e. they allow some degree of interaction between experts—with the hope that this might lead to enhanced judgement (allowing the experts to gain exposure to new arguments or facts about which they may not have been aware, and indeed, curing misconceptions)—while also using some form of mathematical combining of judgements, so pre-empting difficulties and biases that may arise from experts having to negotiate a group consensus.

Mixed approaches can vary in terms of how the behavioural/mathematical aspects are employed. Perhaps the most well-known of the mixed approaches is the Delphi technique. The Delphi method is essentially an iterative survey that has the advantage of providing feedback from the involved experts over successive questionnaire rounds. This feedback comprises details of experts' judgements from the previous round—group responses such as the mean or median, quartiles, and written arguments (the technique does not implicitly demand any one particular type of feedback). Anonymity of experts is a specific feature of Delphi, as the technique is intended to reduce the social and political pressures to accept judgements that can arise in interacting groups; by removing identifying information from feedback, it is thought that experts can/will concentrate on the merits of the feedback information itself without being influenced by potentially irrelevant cues. Furthermore, the end judgement is then obtained through a equal weighting of the experts' judgements on the final round. When using Delphi, the (usually equal weighted) judgements are aggregated from the responses of experts on the 'final' round. Evidence suggests that the *accuracy* of judgements (on judgement and short-term forecasting tasks—albeit often from studies using students rather than experts as subjects) does tend to increase over rounds (Rowe and Wright, 1999), and that this tends to occur because 'less expert' participants tend to change their judgements more often than more expert subjects, which has a corresponding tendency to shift the averaged judgement towards the 'true' answer on successive rounds.

Evidence from practical applications of Delphi has revealed increasing 'drop-out' rates over subsequent rounds. High drop-out rates tend to occur when the expert sample is large and experts are likely to feel that they are relatively insignificant to the whole process and so can quietly slip out of the process by not responding. However, it is likely that things can be done to counter this. For example:

- Recruit a relatively small sample of experts, such that the experts' relative uniqueness can be played upon (tell them, for example, that they are one of only eight 'specially chosen' experts from around the 'world'). Their professionalism (along with the fact that they will have less feedback information to consider) can help to ensure that they continue to respond. In addition, the elicitor can be sure to keep in regular contact with the expert, by phone if necessary, to ensure continued involvement.
- Ask a limited number of questions, for example an initial 'qualitative' round, and then up to three 'quantitative' rounds. If you have many experts/questions, then clearly this will reduce your capacity. Alternatively, by strong piloting (with some appropriate experts) you could potentially get rid of the first, more qualitative, round, and then you could, at a minimum, only have a couple of rounds.

For each of the items included in the Delphi survey it is important to collect rationales for the (quantitative) responses (rationales: request the expert participant to provide a reason for his or her response). This is particularly important in the case of relatively small expert samples and short questionnaires. Ideally, this should be done all the time, but it may not be logistically feasible (for all questions) to do so if you have a huge expert sample or many questions.

Although the mandate has its own timeline, remember that creating and sending out a survey requires its own timeline. This should take into account, for example, the time the expert respondent needs to respond.

In order to create a timeline you need to take into account several steps. It is difficult to provide a general timeline as time is influenced by many factors, for example the amount of time available for the mandate or the number of persons that can be (full-time) involved. The experience you have with executing the method is also a factor, as with more experience you can draw on example surveys and know what works (and what does not!) for your domain. The estimated times for executing some of these stages may vary widely depending on the above-mentioned factors; therefore, a time range is provided here.

Table 13 shows the estimated time for the steps that are involved in elicitation using a questionnaire or one round of Delphi.

Table 13: Rough timeline of one round of Delphi

Step	Estimation of time needed
Set estimated timeline	1 day
Survey development	1–2 weeks
Pilot of survey	1 week
Expert training on probabilistic judgements	1 day
Send out survey	1 day
Survey out with expert participants	2–4 weeks
Send out participant reminder for survey	1 day
Closure of survey and data collation	1 day—2 weeks
Data analysis	1 week

Alternative versions of Delphi, such as ‘real-time’ Delphis, are not discussed here. It should be noted that their merits are not clearly established. Such Delphi processes would require greater specific experience from the elicitor to run the exercise, with appropriate software. In addition, it needs a different type of commitment for the expert to participate. With a ‘normal’ Delphi the expert can chose when to complete the survey (in one go or over multiple sessions), whereas a ‘real-time’ Delphi requires all experts to be present on a certain date and at a particular time.

A second mixed method is the ‘nominal group technique’ (NGT). This differs from Delphi in that the experts do meet physically in one room, and are allowed to discuss the problem (with the help of a elicitor) in a way similar to most behavioural approaches. However, rather than allowing the experts to then come to a group consensus, experts are asked to provide their judgements individually after the discussion, and these judgments are mathematically aggregated to provide a group response (again, usually through an equal-weighting process). This is done because—it is argued—it is at this last stage that inappropriate social and political pressures often come into play, leading to biases in the process (Van de Ven and Delbecq, 1971).

These approaches may have considerable merit, but they also have limitations. The Delphi approach, for example, is a remote process, and it is unclear how powerful is the limited feedback often provided (though evidence does suggest that aggregated judgments improve in accuracy from first to final rounds—e.g. Rowe and Wright, 1999), while the NGT is perhaps more a set of ideas than a well-defined technique per se, and it has not been the source of much empirical research.

A.4.5. Final remarks

In practice, a method of aggregation should be easy to apply and easy to explain, and should never do something ridiculous.

If a structured method of aggregation other than the equal weighting scheme is preferred, then the equally weighted combination of opinions can be used as a benchmark, in the hope that the chosen aggregation method outperforms it.

A.5. DOCUMENTING THE METHOD APPLIED AND THE RESULTS

Full public documentation is one of the fundamental characteristics of EFSA's work. This ensures that the risk assessment procedures are done in a transparent manner, an obligation which is already mentioned in the founding regulation of EFSA (EC 178/2002[36]). In 2006 and 2009, EFSA published two guidance documents on procedural (EFSA SC, 2006a) and scientific (EFSA SC, 2009) aspects of transparency in risk assessment.

Moreover, some guidance might be necessary on how to apply the general principles to the process of EKE. This chapter proposes the structure and details of the documentation of an EKE exercise for publishing a complete set of information in a harmonised way. Specific formats are given as examples in part II, chapter 6.

The EFSA guidance on procedural aspects of transparency (EFSA SC, 2006a) sets the general conditions for the involvement of civil society stakeholders. It is mentioned there that it is of crucial importance that the topics and stages in the risk assessment procedure in which stakeholders are involved are predefined and clearly stated. The previous chapters of this guidance also showed that EKE itself is rather a full process than a single method. Therefore, the documentation has to summarise all steps and decisions taken from the initiation until the final result. In this way different stages become the responsibility of different participants in the process: reasoning of the need/use of expert elicitation is carried out by the risk assessors, definition and evaluation of the protocol by the Steering Group and documentation of conformity with the protocol and the results by the Elicitation Group.

Nevertheless, the process will serve only the results into the actual risk assessment procedure. Taking this into account, the documentation should be divided into three types of reports (Table 14), each adapted to the needs of the task and audience.

Table 14: List of reports to document an expert knowledge elicitation

Type of report	Content/audience	Author
Result report	Summarises the results and will be used and published in the risk assessment procedure	Elicitation Group
Technical support document	Includes a full description of the process and enables the public to review the study	
	Decision for expert knowledge elicitation	Working Group
	Definition of the elicitation protocol and selection of experts	Steering Group
	Execution and documentation of the elicitation phase	Elicitation Group
Expert feedback	Confidential documentation for the individual expert summarising the input from each expert	Elicitation Group

A.5.1. The result report

The result report should provide the risk assessors with all information necessary to use and interpret the results of the EKE correctly. The report should be clear and understandable with an appropriate degree of detail. In the final risk assessment the result report will be one information source among many others. Therefore, an unambiguous answer to the risk assessment question is essential, as is discussion of constraints in the use of the results. As part of the final risk assessment the result report will also be recognised by the wider public. It should be written in a way that also enhances the general understanding of the risk assessment.

[36] See footnot 5, page 13.

The report should be usable as stand-alone document which answers the question of interest. In the case of potential conflicts, the steering and factual expert elicitation can be carried out by external contractors, to show the independence of the information generation from the risk assessment. Typically the elicitor conducting the elicitation is the author of the result report.

The report should include the following:

- *The constitution of the Steering Group responsible for the process*

  Expert elicitation is one method to gather quantitative information for the use in risk assessment. Even if the initiation and the use of expert elicitation depend on the risk assessment question, both processes should be separated in the conduct and documentation. The Working Group handling the EFSA mandate will therefore not automatically steer the expert elicitation. The result report names clearly the Steering Group responsible for the management and documentation of the expert elicitation process.

- *Summary of the risk assessment context as given to the experts*

- *The questions finally asked to the experts*

  Between the initiation and the final elicitation several adjustments might be made to the concrete question asked to the experts. But the correct interpretation of the results depends on the exact context (wording) given to the experts. The result report should clearly state which question was definitely and finally asked to the experts. This also includes different translations, if applicable.

  If the question was substantially modified at the request of the experts, for example to include or exclude specific conditions, the modification should be explained and discussed in view of the use of the results in the risk assessment.

- *A concise description of the method finally applied to gather and analyse the answers*

  The elicitation method and data analysis should be described in a clear and understandable way to explain the advantages and limitations for the use of the results. Special emphasis should be given to reason the choice of this method and possible modification to standard procedures.

- *The criteria used to identify necessary expertise to answer the questions*

  The expertise necessary to answer the questions should be defined before the selection of the experts. The explicit criteria used to evaluate the results of expert selection should be mentioned.

- *A list of experts who participated in the elicitation process and the elicitors involved including the reasoning for deciding that these individuals had the necessary expertise*

  All experts who participated in the elicitation procedure should be listed along with a short description of their affiliation or profile. If experts were only partly involved or declined to participate in the elicitation, this should be mentioned and discussed with regard of the use of the results. If additional elicitors were involved, these should be listed too together with their tasks. The final constitution of the expert panel should be evaluated against the criteria of necessary expertise.

- *The timeline and duration of the elicitation process*

To indicate the up-to-dateness and intensity of the elicitation process, the timeline and duration should be summarised. This includes training and information sessions.

- *The result for use in risk assessment*

  The final results will be documented using the probabilistic approaches to express uncertainties as given in chapter A.3. To avoid any confusion, the result report should clearly mention, which result should be used in the risk assessment. The physical units and necessary conversions should be explicitly named.

- *Discussion of assumptions, qualitative uncertainties and constraints of the result*

  Even if the elicitation itself quantifies the uncertainties, procedural aspects can only be discussed in a qualitative manner. This includes

  - all assumptions made to run the elicitation process which might have an influence on the result;
  - discussion and reasoning of decisions made in the process which might have had an influence on the result;
  - irregularities or deviations from initially planned process which might have an influence on the result.

  Special emphasis should be given to resulting constraints for the use of the result in the risk assessment procedure.

- *Any complaints to the result, if declaimed by a participant*

  Finally, all feedback from the participants which can influence the use of the result should be mentioned and discussed.

A.5.2. The technical support document

The technical report focuses on the procedure, reasoning, data pool and analysis of the EKE process. It can be assumed that the audience of the technical report is familiar with this guidance document and the standard procedures.

The main task of the technical support document is to enable interested parties to review the process and describe responsibilities and decisions made during the process. The document should be so detailed that a reader is, in principle, in a position to re-run the process and test the sensitivity by changing some decisions, e.g. re-do with a new set of experts.

There are three milestones:

1. decision to carry out EKE;
2. definition of the elicitation protocol and selection of experts;
3. execution and documentation of the elicitation process.

Because these milestones are the responsibility of different groups/individuals, the documentation will be provided by different authors: the EFSA Working Group, the Steering Group and the elicitor.

Milestone 1

The decision to undertake an EKE exercise is normally taken by an EFSA panel, Working Group or responsible unit. This decision should be supported by the following documentation:

- *Description of the risk assessment context, the quantitative parameter and model of interest and the intended use of the result*

  The whole following procedure is dependent on and restricted to the intended use of the results. Therefore, EFSA should provide all selected participants as well as the later user of the results with a comprehensive description of the context, the concrete model, information on the other parameters in the model and the intended use of the final results.

  This report is normally also part of the introductory information given to the experts before the elicitation.

- *Documentation of the key search terms and sources evaluated to retrieve information on the parameter and on the information gaps providing the rationale for expert elicitation*

  The use of published information on the parameter of interest is generally preferred in the EFSA's risk assessment. To justify EKE, an extensive search on the topic should be performed and fully documented in accordance with the related guidance document (EFSA, 2010a). Depending on the outcome the Working Group makes the proposal to perform an expert elicitation. Both the reasoning and all information retrieved on the parameter of interest should be documented.

  The summary of existing information is normally also part of the information given to experts before the elicitation exercise.

- *Conclusion on the need for expert elicitation and resources allocated for the process*

  The reasons why EKE is necessary and the limitations of the timeline of the mandate and of available resources should be documented. It should be made clear that EFSA decides the timeline, staff resources and the make-up of the Steering Group as well any involvement of external contractors.

- *Final outcome: constitution of the Steering Group, additional resources for the elicitation, documentation of the existing information about the quantitative parameter of interest*

  To initiate the process following information should be provided:

  - background information on the risk assessment model;
  - existing information on the parameter of interest;
  - preliminary timeline of the project;
  - members of the Steering Group;
  - personal resources within EFSA;
  - tasks and resources and for external contracts.

The results of the initiation phase might be internally or externally reviewed before entering into the next step, the definition phase.

Milestone 2

The definition of the elicitation protocol and the selection of experts will be carried out by the Steering Group responsible for supervising the elicitation process. The following steps should be reasoned and documented:

- *Development and framing of the questions for EKE*

  For model parameter in a risk assessment to be assessable through an EKE exercise, the problem has to be reframed. In the documentation possible differences between the initial problem and the framing for expert elicitation should be described and discussed, and alternative framings should also be considered.

- *Discussion and definition of necessary expertise, including the development of evaluation criteria*

  Before the expert selection the necessary expertise to answer the questions has to be defined. The documentation should clearly distinguish between the definition determined before expert selection and the expertise achieved in the final expert group. To allow an independent evaluation, a set of operational criteria should be predefined and documented.

  A typical way to define the necessary expertise is to describe appropriate profiles and roles of the intended experts and the number of requested experts per profile/role. Concrete evaluation criteria should be explicitly given to review the expert group, but also to enable the public to re-do the exercise with another group of experts fulfilling these criteria.

- *Procedure to identify and contact possible experts and elicitors*

  There is no one ideal way to identify possible experts. However, the identification process is likely to involve contacting or evaluating appropriate resources, such as Member State administrations, scientific networks, knowledge databases and stakeholder organisations, reviewing the responses and repeating the exercise with new contact points. This process should be fully documented and the contact points, questions and a summary of responses clearly described.

  To ensure the confidentiality of personal data, the expert names in connection with the results of individual evaluations will not be published. Nevertheless, the identification and contacts of possible experts should be thoroughly documented enable selection biases to be identified, for example arising from the refusal of some stakeholder groups to participate

  In addition, the decision of the Steering Group on the elicitation method has to be documented as well as the selection of the Elicitation Group to execute the elicitation protocol.

- *Selection of experts and evaluation of the necessary expertise and elicitors*

  At the end of the selection process the final expert panel should be evaluated against the predefined criteria. The impact of differences on possible restrictions of the elicitation result should be discussed.

  The evaluation can be based on general descriptions, which avoids a direct assignment of the experts to the predefined profiles. The evaluation of expertise should not be reported on an individual level. Furthermore, the proposed membership of the Elicitation Group should be documented and justified.

- *Decision on the elicitation protocol*

The Steering Group, including the elicitor, agrees, documents, and justifies the final protocol of the elicitation.

- *Final outcome: list of experts and elicitors and the expert elicitation protocol*

  To define the elicitation protocol, the following outcomes are given:

  - elicitation question;
  - list of selected experts (with possible substitutes, shortlist);
  - selected Elicitation Group;
  - detailed elicitation protocol with timeline and responsibilities.

  A list of experts is often given as acknowledgement of their participation in the documentation of the elicitation process.

Milestone 3

The final execution of the elicitation will be carried out by an internal or external qualified Elicitation Group.

- *Training, instructions to the expert and factsheets on the questions*

  The training and background information given to the experts should be fully documented.

- *Preparation of the elicitation process*

  The preparation of the elicitation should be documented along with to the background information, e.g. the invitation letters, attendance lists, clarifications, etc.

- *Documentation of the elicitation process*

  Finally, the elicitation process should be documented according to the structured protocol for each question.

  The report should mention if the result was signed off by the expert or if some experts disagree with the result. In particular, discussions on the validity of the elicitation exercise should be noted.

- *Retrieved answers and data analysis, including intermediate results*

  For most protocols, intermediate results will be retrieved before getting the final result. These results should be also documented in an anonymous way to show the performance of the protocol, deviations from the presettings and irregularities in the execution.

- *Final outcome: result report, experts' feedback*

  Finally, the result of the elicitation procedure should be documented, including the process evaluation of the experts. This is usually achieved using the result report and the expert feedback.

During the elicitation process the experts involved are several times invited to provide personal data and individual judgements. This starts in the expert selection phase to identify appropriate expertise and continues during calibration and elicitation of the topics of interest.

Disclosing these personal data will provide the public with information on individual performance, which is neither an objective of the EKE process nor necessary to fulfil transparency. On the contrary, disclosing personal data may discourage experts from taking part in the process or influence their responses, because of possible pressure to justify their answers to related interests groups (e.g. companies, administrations). "*Experts may feel they can respond more freely and avoid some of the motivational biases*" (RIVM, 2008). Therefore, participating experts should be assured on the confidential treatment of their individual answers (Acera, 2010; Chatham House Rules[37]). This is in accordance with EFSA's management board decision concerning measures of transparency and confidentiality requirements (EFSA MB, 2005) and the guidance on transparency (EFSA SC, 2006a) to protect the privacy and integrity of the individual.

Personal responses (other than a list of names and affiliations of all experts involved) will be reported in such a way that anonymity is guaranteed. This is normally done by giving experts a neutral identifier and reporting only datasets which do not allow the retrieval of responses of individual experts, e.g. not naming the origin of experts when only one or a few experts sharing this characteristic.

To ensure confidentiality in sensitive cases it is recommended that steering and elicitation are carried out by external contractors, who are contractually committed to confidentiality by EFSA.

A.5.3. The expert feedback

Nevertheless, individual experts, on request, should be provided with a written summary of the information they input to the process and all supplementary details of the analysis which are needed to show how this information was used to generate the final results. If the analysis shows that experts had a tendency towards bias, under- or overconfidence, this should be passed back to them (Cooke and Goosens, 1999). As the experts involved will have allocated time and resources to the process, it is important to give individuals feedback on their input and a brief description on the overall results of the risk assessment. In some circumstances individual feedback by phone or other means might also be necessary, to answer individual questions, address concerns or listen to the reactions of the experts to the expert elicitation.

This report fulfils EFSA's obligation of transparency to the individual experts involved in the expert elicitation process. It provides assurance that EFSA has used the correct input data and should build trust that individual inputs were treated in an appropriate manner. A simple realisation of the expert feedback is the disclosure of the expert's neutral expert identifier. It is recommended that the full technical support document should also be made available to the experts.

The expert feedback should be provided by the elicitor, as this individual will have had direct contact with the participating experts.

[37] Chatham House Rule: http://www.chathamhouse.org/about-us/chathamhouserule

B. EXTENSIVE WEB SEARCH ON GUIDANCE DOCUMENTS ON EXPERT KNOWLEDGE ELICITATION

B.1. OBJECTIVES AND SCOPE OF THE LITERATURE SEARCH

The literature search aimed to produce an extensive[38] list of reports on guidance on expert knowledge elicitation in food and feed safety risk assessment, which were subsequently reviewed and taken into consideration for the development of the EFSA Guidance.

The overall scope was to ensure that for the development of the EFSA Guidance the working group experts could draw upon relevant and broad information on already existing methods for expert knowledge elicitation.

B.2. THE LITERATURE SEARCH AND SELECTION PROCESSES

The methodology applied for searching and selecting the relevant reports was defined via an iterative and consultative process which implied scoping the literature (using an unstructured search) and working group discussion.

The results of the scoping exercise showed that existing guidelines for expert knowledge elicitation often are not referenced in typical bibliographic databases. Thus it was assumed that relevant guidelines are published on institutional websites as "grey" literature[39].

The method reported and documented in the following sessions is the definitive method applied.

B.2.1. Criteria for assessing report relevance

As first step the working group agreed that the literature search should indentify reports/guidance/guidelines covering the full expert knowledge elicitation method and not scientific evidence on specific aspects of the expert knowledge elicitation process.

The agreed criteria for relevance were:

- Content of the record: expert knowledge elicitation method(s) (full process) relevant for risk assessment in food and feed safety;
- Publication type: report/guidance/guideline/review (no presentations, proceedings, or websites);
- The authorship should be clearly specified (e.g. author(s) or institution(s) names).

In addition:

- Language restrictions: only reports in English;
- The records would be assessed for relevance up to page 10 of the search engine results (see below the section on information sources);
- No date limits applied to the search.

[38] Comprehensive literature searches are rather difficult achieve because of the number of information sources in different languages available to be searched. Therefore, this searched aimed to produce a set of relevant reports as extensive as possible.

[39] Grey literature is defined as types of publication which are less systematically recorded in bibliographic tools such as catalogues and databases than journals and books (EFSA, 2010).

B.2.2. Information sources searched

A restriction of the search to few institutions seemed not appropriate and it was preferred to run the search in the following search engine:

1. Google (UK): www.google.co.uk

The reference lists of the reports deemed to be relevant after the selection process were also used as information sources.

The search proved to be very time consuming by individual check of each website, and the typical unstructured information on it. Therefore alternative searches using other search engines, as Yahoo (UK, www.uk.yahoo.com) or Bing (IT, www.it.bing.co) was done only as a feasibility check.

B.2.3. Search terms

The synonyms were restricted only to alternative wordings. No wild-cards were used to define full sets of wordings, due to the restrictions of the search engines.

A sensitive search strategy (i.e. combination of terms) was used, capturing four key concepts via all relevant search terms and possible synonyms, as illustrated below:

Expert	AND	(Knowledge) OR Judgement OR Opinion	AND	Elicitation OR Involvement OR Consultation	AND	Guide (Guideline) (Guidance) OR Methods OR Techniques OR Procedures OR Protocol OR Review

- The term “Knowledge” was disregarded from the search due to irrelevant results
- The term “Guide” seemed to cover “Guideline” and “Guidance”. But the concrete algorithm used by the search engines is unknown to us.

At the end six phrases were searched:

1. **Expert Judgement Elicitation Guide**
2. **Expert Judgement Elicitation Methods**
3. **Expert Judgement Elicitation Procedures**
4. **Expert Opinion Elicitation Guide**
5. **Expert Opinion Elicitation Methods**
6. **Expert Opinion Elicitation Procedures**

The final search was performed from 31st January 2012.

Per search the first 10 result pages, containing about 100 links, were examined.

B.2.4. Exclusion of web pages

In total 601 web pages were screened. 343 (=57.9%) web pages were excluded by at least one criteria defined in section 1.1. Within the excluded web sites a “restricted view of only specific applications” was the most important reason for exclusion (189 = 55.1%). This was followed by web sites with wrong publication types: “Pure web sites” (51 = 14.9%), “Presentations” (32 = 9.3%).

Code	*Description*	*Count*	*Percentage*
Ex10	Unknown author	6	1.7%
Ex11	Wrong publication type: presentation	32	9.3%
Ex12	Wrong publication type: proceedings	4	1.2%
Ex13	Wrong publication type: website	51	14.9%
Ex14	Review	2	0.6%
Ex21	Wrong content: No expert knowledge elicitation method(s)	18	5.2%
Ex22	Wrong content: Not relevant for RA in food and feed safety	10	2.9%
Ex31	No guideline: Scientific note	9	2.6%
Ex32	No guideline: Specific topic	189	55.1%
Ex90	Page no more exists	19	5.5%
Ex98	No program to open	2	0.6%
Ex99	Covered by article from same author	1	0.3%
Total		**343**	**100%**

B.2.5. Identification of relevant publications

From the 601 reviewed web pages 258 (= 42.9%) passed the screening. From we were able to identify 86 individual publications. The most frequent referenced publication was the book of Meyer and Booker (1991) (34 web pages), followed by the book of Ayyub (2001) (15 web pages).

B.3. RESULTS OF THE LITERATURE SEARCH AND SELECTION PROCESSES

The following publications were identified via the web search and further reviewed for the development of the guidance document. 8 publications (updated and/or published versions) were added by the retrieval of the bibliographic data and full text publications.

The following list is ordered by time and "research groups" / topics:

"RAND / Delphi (1968)"
Galway 1950 LA Galway 1950 Subjective probability distribution elicitation in cost risk analysis: a review. RAND corporation 2007 http://www.rand.org/content/dam/rand/pubs/technical_reports/2007/RAND_TR410.pdf
Brown 1968 BB Brown 1968. Delphi process: A methodology used for the elicitation of opinions of experts. **RAND corporation** http://www.rand.org/content/dam/rand/pubs/papers/2006/P3925.pdf
"Morgan, Henrion (1990)"
Morgan, Henrion 1990 MG Morgan, M Henrion 1990: Uncertainty. **Cambridge Uni Press** http://books.google.vg/books?id=ajd1V305PgQC&source=gbs_similarbooks
"Meyer, Booker (1991)"
Meyer, Booker 1991 Mary A Meyer / Jane M Booker: Eliciting and Analysing Expert Judgement: a Practical Guide, **ASA-SIAM**. London, 1991 http://books.google.vg/books/about/Eliciting_and_analyzing_expert_judgment.html?id=Wwz2sRcmqwkC http://epubs.siam.org/ebooks/siam/asa-siam_series_on_statistics_and_applied_probability/sa07
Booker, Meyer 1996 JM Booker, MM Meyer 1996. Elicitation and Analysis of Expert Judgement http://www.armyconference.org/ACAS00-02/ACAS00ShortCourse/ess.pdf
Meyer et al 2002 MM Meyer et al 2002: Guidelines for Expert Judgment as Probabilities or Fuzzy Logic. In: Roos, Booker, Parkinson: Fuzzy Logic and Probability Applications – Bridging the Gap, Chapter 2 SIAM 2002 http://www.armyconference.org/ACAS00-02/ACAS00ShortCourse/ch5.pdf
"Cooke (1991)"
Cooke 1991 RM Cooke 1991: Experts in Uncertainty. **Oxford Uni Press** http://books.google.vg/books?id=4taZBr_nvBgC&source=gbs_similarbooks
Goossens, Cook 1997 LHJGoossens, RM Cooke 1997. Applications of some risk assessment techniques: Formal expert judgement and accident sequence precursors. **Safety Science 26, 35-47** http://www.sciencedirect.com/science/article/pii/S0925753597000271
Cooke, Goossens 1999 RM Cooke, LHJ Goossens 1999: Procedures Guide for Structured Expert Judgement Brussels: **EC DG Environment** Report EUR 18820, 1999. ftp://ftp.cordis.europa.eu/pub/fp5-euratom/docs/eur18820_en.pdf
Cooke, Goossens 2000 RM Cooke, LHJ Goossens 2000: Procedures Guide for Structured Expert Judgement in Accident Consequence Modelling**.** **Radiat Prot Dosimetry 90, 303-309** http://rpd.oxfordjournals.org/content/90/3/303.full.pdf
GoossensEtAl 2008 LHJ Goossens et al 2008. Fifteen years of expert judgement at TUDelft. **Safety Science 46, 234-244** http://www.sciencedirect.com/science/article/pii/S0925753507000495
van der Fels-Klerx et al. 2002 HJ van der Fels-Klerx, LHJ Goosens, HW Saatkamp, SHS Horst (2002) Elicitation of Quantitative Data from a Heterogeneous Expert Panel: Formal Process and Application in Animal Health. **Risk Analysis 22 67-81** http://onlinelibrary.wiley.com/doi/10.1111/0272-4332.t01-1-00007/abstract?
RIVM 2008 **RIVM** 2008: Expert Elicitation http://www.nusap.net/downloads/reports/Expert_Elicitation.pdf
Knol et al. 2010 AB Knol, P Slottje, JP van der Sluijs, E Lebret 2010: The Use of Expert Elicitation in Environmental Health Impact assessment: A Seven Step Procedure. **EnvHealth 9, 19** http://www.ehjournal.net/content/pdf/1476-069X-9-19.pdf
INTARESE 2007 Integrated assessment of Health Risks of Environmental Stressors in Europe (**INTARESE**) 2007. D15 Heath Effects Methodology – Protocol and Guidelines http://www.integrated-assessment.eu/sites/default/files/Health%20Effects%20Methodology.pdf
Aspinall 2008 W Aspinall 2008. Expert Judgment Elicitation using the Classical Model and EXCALIBUR http://risk2.ewi.tudelft.nl/extrafiles/EJcourse/Sheets/Aspinall%20Briefing%20Notes.pdf

Aspinall 2011 Aspinall 2011: Expert Elicitation and Judgement Pooling using the Cooke Classical Model and EXCALIBUR. http://risk2.ewi.tudelft.nl/ejwshop_2011_files/Aspinall_Briefing_Note_on_Classical_Model_and_EXCALIBUR_Jul_2011.pdf (http://risk2.ewi.tudelft.nl/workshops/33-ejworkshop2011)
Ouchi 2004 A Literature Review on the Use of Expert Opinion in Probabilistic Risk Analysis. **World Bank** Policy Research Paper 3201, Feb. 2004 http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2004/04/15/000009486_20040415130301/additional/115515322_20041117173031.pdf
"Potthoff(1991)"
Potthoff 1991 S J Potthoff 1991. A comparison of two elicitation methodologies for modelling expert judgements. **PhD Thesis**. Madison: Univ. of Wisconsin,1991 http://www.worldcat.org/title/comparison-of-two-elicitation-methodologies-for-modeling-expert-judgment/oclc/24571287
"Burge (1998)"
Burge 1998 **JE Burge 1998. Knowledge Elicitation Tool Classification. Worcester Polytechnic Institute: Artificial Intelligence Research Group** **http://**web.cs.wpi.edu/Research/aidg/KE-Rpt98.html
"Ayyub (2001)"
Ayyub 2001 BM Ayyub 2001: Elicitation of Expert Opinions for Uncertainty and Risks. **CRC Press** 2001 http://www.crcnetbase.com/doi/pdf/10.1201/9781420040906.fmatt http://books.google.vg/books?id=VFH60G8JzYYC&source=gbs_similarbooks
Ayyub 2000 BM Ayyub 2000. Methods for Expert-Opinion Elicitation of Probabilities and Consequences for Corps Facilities. IWR-Report-00-R-10 http://www.iwr.usace.army.mil/docs/iwrreports/00-R-101.pdf
Ayyub 2001b BM Ayyub 2001: A Practical Guide on Conducting Expert-Opinion Elicitation of Probabilities and Consequences for Crops Facilities. http://www.iwr.usace.army.mil/docs/iwrreports/01-R-01.pdf
"NRCP (1996)"
NCRP 1996 National Council on Radiation Protection and Measurements (**NRCP**) 1996. A Guide for Uncertainty Analysis in Dose and Risk Assessments to Environmental Contamination: Section 4: Elicitation of Expert Judgement. Commentary No. 14. http://www.knovel.com/web/portal/browse/display?_EXT_KNOVEL_DISPLAY_bookid=2024
"Rosqvist (2003)"
Rosqvist 2003 T Rosqvist: On the use of expert judgement in the qualification of risk assessment. Espoo (FIN), Helsinki University of Technology, **Dissertation**, 2003. http://lib.tkk.fi/Diss/2003/isbn9513862445/isbn9513862445.pdf
"JRC (2005)"
Simola et al 2005 K Simola, A Mengolini, R Bolado-Lavin 2005: Formal Expert Judgement – An Overview. **JRC Inst.f. Energy**, EUR 21772 EN. http://ie.jrc.ec.europa.eu/publications/scientific_publications/2005/EUR21772EN.pdf
Bolado, Badea 2009 R Bolado, A Badea M Poole 2009:Review Expert Judgement Methods for Assigning PDFs. **PAMINA** Milestone M2.2.A.3 http://www.ip-pamina.eu/downloads/pamina.m2.2.a.3.pdf
Grupa 2009 JB Grupa 2009. Trial of formal use of Expert Judgement for scenario conceptualisation. **PAMINA** Milestone M2.2.C.3 http://www.ip-pamina.eu/downloads/pamina.m2.2.c.3.pdf
"O'Hagan (2006)"
O'Hagan et al 2006 A O'Hagan et al 2006: Uncertain Judgements: Eliciting Experts' Probabilities. **Wiley**. http://books.google.vg/books?id=tl7CDpmlgFwC&source=gbs_similarbooks
Daneshkhah 2004 AR Daneshkhah 2004. Psychological Aspects Influencing Elicitation of Subjective Probability http://www.sheffield.ac.uk/content/1/c6/03/09/33/Psychologypapers.pdf
Garthwaite et al. 2005 PH Garthwaite et al 2005. Statistical Methods for Eliciting Probability Distributions. JASA 100 680-701 http://www.stat.cmu.edu/tr/tr808/tr808.pdf http://www.tandfonline.com/doi/pdf/10.1198/016214505000000105
Garthwaite et al. PH Garthwaite, JB Kadane, A O'Hagan: (Chapter 1) Elicitation http://www.shef.ac.uk/content/1/c6/03/09/33/elichand.pdf

Jenkinson 2005 D Jenkinson 2005. The Elicitation of Probabilities – A Review of Statistical Literature www.shef.ac.uk/content/1/c6/03/09/33/review3.pdf
Oakley 2010 J Oakley 2010. Eliciting Univariate Probability Distributions http://www.jeremy-oakley.staff.shef.ac.uk/Oakley_elicitation.pdf
O'Hagan 1998 A O'Hagan 1998. Eliciting Expert Beliefs in Substantial Practical Applications. The Statistician 47, 21-35 http://www.ime.unicamp.br/~lramos/mi667/ref/2ohagan98.pdf
O'Hagan 2005 A O'Hagan 2005. Research in Elicitation http://ryle-if-a.mirrorservice.org/sites/lib.stat.cmu.edu/DOS/general/first-bayes/pdf/ElicRes.pdf
O'Hagan 2011 A O'Hagan 2011. Probabilistic uncertainty specification: Overview, elaboration techniques and their application to a mechanistic model of carbon flux. **http://www.tonyohagan.co.uk/academic/pdf/probspec.pdf**
"INL (2005)"
INL 2005 **INL** 2005: Simplified Expert Elicitation Guideline for RA of Operating Events http://www.inl.gov/technicalpublications/Documents/3310952.pdf
"Hoffmann (2006)"
Hoffmann et al. 2006 S Hoffmann, P Fischbeck, A Krupnick, M Williams 2006: Eliciting Information on Uncertainty from Heterogeneous Expert Panels, **Res.f. Future** Discussion Paper DP 06-17-Rev http://ageconsearch.umn.edu/bitstream/10444/1/dp060017.pdf
Hoffmann et al 2007 Elicitation from large, heterogeneous expert panels: using multiple uncertainty measures to characterize information quality for decision analysis. Decision Analysis, June 2007 http://business.highbeam.com/435294/article-1G1-167026973/elicitation-large-heterogeneous-expert-panels-using
Hoffmann et al. 2008 S Hoffmann, P Fischbeck, A Krupnick, M Williams updated 2008: Foodborne Illness to Their Food Sources – Using Large Expert Panels to Capture Variability in Expert Judgement, **Res.f. Future** DP 06-17-Rev http://www.rff.org/RFF/Documents/RFF-DP-06-17-REV.pdf
"Bedford (2006)"
Bedford et al 2006 Bedford et al 2006Expert Elicitation for Reliable system Design. **Statistical Science 21, 428-450** http://projecteuclid.org/DPubS/Repository/1.0/Disseminate?view=body&id=pdfview_1&handle=euclid.ss/1177334515
"Accera (2006)"
Acera 2006 **Acera**2006_ElicitingExpertJudgements_LitReview http://www.acera.unimelb.edu.au/materials/endorsed/0611.pdf
Acera 2008 A Speirs-Bridge et al. 2008. Reducing Overconfidence in the Interval Judgment of Experts. **ACERA** project 0611 http://www.acera.unimelb.edu.au/materials/endorsed/0611_Spiers-Bridge.pdf
Acera 2010 **Acera** 2010: Process manual – Elicitation tool. http://www.acera.unimelb.edu.au/materials/endorsed/0611-process-manual.pdf
"Defence Canada (2007)"
Defence Canada 2007 K Leung, S Verga: Expert Judgement in Risk Assessment. **Defence R&D Canada,** Centre for Security Studies, DRDC CORA TM 2007-57. http://pubs.drdc.gc.ca/PDFS/unc88/p529083.pdf
"Hora / Encyclopedia (2008)"
Hora 1992 S Hora 1992. Acquisition of Expert Judgment: Examples from Risk Assessment. **J.Energy Eng 118, 136-148**
Hora 2009 SC Hora 2009. Expert Judgement in Risk Analysis. CREATE Research Archive
Hora 2008 SC Hora 2008. Expert Judgment Elicitation Methods. CREATE http://create.usc.edu/research/past_projects/Hora2008-ExpertJudgmentElicitationMethodsandTools.pdf
Sims et al 2008 BH Sims et al 2008. Expert Opinion in Reliability. Encyclopedia of Statistics in Quality and Reliability. **Wiley,** Online DOI: 10.1002/9780470061572.eqr110
Wiedlea 2008 ACK Wiedlea 2008. Expert Elicitation for Risk Assessment. In: Encyclopedia of Quantitative Risk Analysis and Assessment. **Wiley** Online DOI: 10.1002/9780470061596.risk0490

(Hora 2008) SC Hora 2008 Expert Judgement In: Encyclopedia of Quantitative Risk Analysis and Assessment. Wiley Online DOI: 10.1002/9780470061596.risk0525
"Hukki (2008)"
Hukki 2008 K Hukki 2008: A formal Process for Elicitation and Validation of Expert Judgements for Safety Case in the Context of Spent Nuclear Fuel Management. **Posiva** http://www.posiva.fi/files/508/WR_2008-60web.pdf
"Dieste (2011)"
Dieste, Juristo 2011 O Dieste, N Juristo. 2011: Systematic Review and Aggregation of Empirical Studies on Elicitation Techniques. **IEEE Transactions on Software Engeneering 37, 283-304**
Carrizo et al 2008 D Carrizo, O Dieste, N Juristo 2008. Study on Elicitation Adequacy. http://main.grise.upm.es/reme/publicaciones_download.aspx
Dieste e tal 2009 O Dieste M Lopez, F Ramos 2009. Obtained Well-Founded Practices about Elicitation Techniques by Means of na Update of a Previous Systematic Review http://main.grise.upm.es/reme/publicaciones_download.aspx
"US EPA (2009)"
US Government 2009 US Government 2009. Code of Federal Regulations §194.26 Expert Judgement. **US Governmental Printing Office**, 01/07.2009 http://www.gpo.gov/fdsys/pkg/CFR-2009-title40-vol24/xml/CFR-2009-title40-vol24-sec194-26.xml
US EPA 2009 US Environmental Protection Agency (**EPA**) 2009. Expert Elicitation Task Force White Paper – draft, January 2009 http://www.epa.gov/osa/pdfs/elicitation/Expert_Elicitation_White_Paper-January_06_2009.pdf
US EPA 2011 US Environmental Protection Agency (**EPA**) 2011. Expert Elicitation Task Force White Paper, August 2011 http://www.epa.gov/stpc/pdfs/ee-white-paper-final.pdf
"Critics"
Mosleh et al 1988 A Mosleh, VM Bier, G Apostolakis 1988. A critique of current practice for the use of expert opinions in probabilistic risk assessment. **Reliability Engineering and System Safety 20, 63-85** http://dx.doi.org/10.1016/0951-8320(88)90006-3
Otway et al 1992 H Otway, D von Winterfeldt 1992: Expert Judgement in Risk Analysis and Management: Process, Context and Pitfalls. **Risk Analysis 12, 83-93.** http://onlinelibrary.wiley.com/doi/10.1111/j.1539-6924.1992.tb01310.x/abstract
Kadane, Wolfson 1998 JB Kadane, LJ Wolfson 1998 Experiences in Elicitation. J**.Roy.Stat.Soc.Ser.D (The Statistician) 47, 3-19** http://yaroslavvb.com/papers/kadane-experiences.pdf
Roest 2002 I Roest 2002. Expert Opinion-Use in Practice http://www.few.vu.nl/en/Images/werkstuk-roest_tcm39-91422.doc
Loveridge 2004 D Loveridge 2004. Experts and foresight: review and experience. **Int.J.of Foresight and Innovation Policy 1, 33-69**
Hjorteland, Aven 2007 A Hjorteland, T Aven 2007. On how to use expert judgements in regularity analysis to obtain good predictions. **Zeszyt 3, 158-169** http://t.tribologia.eu/plik/spm/spmom-07v42n3_p-157.pdf
Aspinall 2010 W Aspinall 2010. A route to more tractable expert advice. Nature 463, 294-295
Curtis 2012 **A Curtis 2012. The Science of Subjectivity. Geology 40, 95-96** **http://geology.gsapubs.org/content/40/1/95.full.pdf+html**
"Bias"
Baddeley et al 2004 MC Baddeley, A Curtis, RA Wood 2004. An introduction to prior information derived from probabilistic judgements: Elicitation of knowledge, cognitive bias and herding. http://www.eeo.ed.ac.uk/homes/acurtis/Papers/BaddeleyCurtisWood2004.pdf
Kynn 2007 M Kynn 2007: The 'heuristics and biases' bias in expert elicitations. **J.Roy.Stat.Soc: Ser.A 171, 239-264.** http://onlinelibrary.wiley.com/doi/10.1111/j.1467-985X.2007.00499.x/full
Cleaves 2008 DA Cleaves 2008. Cognitive biases and corrective techniques: proposals for improving elicitation procedures for knowledge-based systems. **Int.J.of Man-Machine studies 27, 155-166**

Speirs-Bridge et al 2009.
A Speirs-Bridge et al 2009 Reducing Overconfidence in the Interval Judgements of Experts. Risk Analysis
http://www.botany.unimelb.edu.au/botany/aboutus/staff/pdf/Speirs_Bridge-et_al-(2010).pdf

"Aggregation"

Clemen,Winkler 1999
RT Clemen, RL Winkler 1999 Combining Probability Distributions from Experts in Risk Analysis. **Risk Analysis 19, 187-203**
http://onlinelibrary.wiley.com/doi/10.1111/j.1539-6924.1999.tb00399.x/pdf

Monti,Carenini 2000
S Monti, G Carenini 2000. Dealing with the Expert Inconsistency in Probability Elicitation. IEEE Transactions on Knowledge and Data Engineering 4, 499-508
http://www.cs.ubc.ca/~carenini/PAPERS/IEEE00.pdf

Albert et al 2007
I Albert 2007. Combining expert opinions in prior elicitation. **Biostatistics 8, 1-33**
http://eprints.qut.edu.au/40812/1/40812.pdf

Predd et al 2008
JB Predd et al 2008. Aggregating Probabilistic Forecasts from Incoherent and Abstaining Experts. Decision Analysis 5, 177-189
http://www.princeton.edu/~osherson/papers/preddAgg.pdf

Weiss 2009
R Weiss 2009 Optimally Aggregating Elicited Expertise: A Proposed Application of the Bayesian Truth Serum for Policy Analysis. **MSc Thesis** MIT
http://dspace.mit.edu/bitstream/handle/1721.1/53066/501829428.pdf?sequence=1

French 2011
Simon French 2011. Aggregating expert judgement. **RACSAM 105, 181-206**
http://www.springerlink.com/content/d366818p21228q1u/

"Qualitative"

Coutts et al 2003
DA Coutts et al 2003. How to Use Qualitative Analysis to Support a DSA. Aitken: Westinhouse Safety Management Solutions LLC
http://sti.srs.gov/fulltext/ms2003376/ms2003376.html

Dalton et al 2010
A Dalton et al 2010. Expert Elicitation Method Selection Process and Method Comparison. LabSi Working Papers N.30/2010
http://www.labsi.org/wp/labsi30.pdf

"BBN"

Wiegmann 2005
DA Wiegmann 2005. Developing a Methodology for Eliciting Subjective Probability Estimates During Expert Evaluations of Safety Interventions: Application for Bayesian Belief Networks. Savoy: University of Illinois 2005
http://www.humanfactors.illinois.edu/Reports&PapersPDFs/TechReport/05-13.pdf

"MicroOrganism"

Costard et al
S Costard et al. Expert opinion in a data-sparse environment: elicitation of parameter values for a risk assessment model on the introduction of African swine fever virus into pig farms in Madagascar

Gallagher et al 2002
E Gallagher et al 2002. Estimating the risk of importation of foot-and-mouth disease into Europe. **Veterinary Record 150, 769-772**
doi:10.1136/vr.150.25.769

Cross et al 2012-07-31
P Cross et al 2012 Eliciting expert opinion on the effectiveness and practicality of interventions in the farm and ruaral environment to reduce human exposure to Escherichia coli O157. Epidemiol. Infect 140, 643-654

Horst et al. 1998
HS Horst et al 1998. Introduction of contagious diseases into the Netherlands: elicitation of expert opinions. Livestock Production Science 53, 253-264
http://www.journals.elsevierhealth.com/periodicals/livest/article/S0301-6226(97)00098-5/fulltext

"Exposure Assessment"

Walker et al 2001
KD Walker, JS Evans, D MacIntosh 2001: Use of Expert Judgement in Exposure Assesment Part 1. Characterization of Personal Exposure to Benzene. **J.Exposure Analysis and Env.Epidemiology 11, 308-322.**
http://www.nature.com/jes/journal/v11/n4/pdf/7500171a.pdf

"Ecology"

Low Choy et al 2009
S Low Choy et al 2009. Expert elicitation and its interface with technology: a review with a view to designing Elicitator. IMACS/MODSIM Congress

Low Choy et al 2009b
S Low Choy et al 2009. Elicitation by design in ecology: using expert opinion to inform priors for Bayesian statistical models. **Ecology 90:265–277.**
http://dx.doi.org/10.1890/07-1886.1

James et al 2010 A James et al 2010. Elicitator: Na expert elicitation tool for regression in ecology. Environmental Modelling & Software 25, 129-145 http://www.sciencedirect.com/science/article/pii/S1364815209001704 http://eprints.qut.edu.au/30011/1/c30011.pdf
O'Leary et al 2009 RA O'Leary et al 2009. Comparison of four expert elicitation methods: For Bayesian logistic regression and classification trees. 18th IMACS/MODSIM Congress. http://www.mssanz.org.au/modsim09/J2/oleary_J2.pdf Ecography 2009, 4276-4282
Kuhnert et al 2010 PM Kuhnert et al 2010. A guide to eliciting and using expert knowledge in Bayesian ecological models. **Ecology Letters 13, 900-914**
Kuhnert et al 2009 PM Kuhnert et al 2009.Expert opinion in statistical models IMACS/MODSIM Congress **http://www.mssanz.org.au/modsim09/J2/kuhnert_J2.pdf**
"Geological"
Curtis, Wood 2004 A Curtis, R Wood 2004. Optimal elicitation of probabilistic information from experts. **Geological Society London, Special Publications, 1, 127-145** http://www.geos.ed.ac.uk/homes/acurtis/Papers/CurtisWood2004.pdf
"Human Failure"
Forester et al. 2004 J Forester et al. 2004. Expert elicitation approach for performing ATHEANA quantification. **Reliability Engineering and System Safety 83, 207-220** http://hera.ugr.es/doi/15076878.pdf
"Model"
Heuvelink et al 2011 G Heuvelink et al 2011. Expert Elicitation of input and model uncertainties. Deliverable 3.1 of Uncert Web http://ec.europa.eu/information_society/apps/projects/logos/8/248488/080/deliverables/001_UncertWebD31ARES20111005733.pdf

B.4. Discussion

As assumed the web search was able to retrieve Guidance documents published as "grey literature" from important research groups and institutions in risk assessment, e.g. JRC, US EPA. Nevertheless the unstructured information on web sites and the excessive duplication of information (more than 13% of the included web sites referring to only one book) make a simple web search inefficient. The use of more advanced web search techniques is therefore highly recommended.

B.5. References

EFSA (European Food Safety Authority), 2010. Application of systematic review methodology to food and feed safety assessments to support decision making. EFSA Journal 2010; 8(6):1637. [90 pp.]. doi:10.2903/j.efsa.2010.1637. Available online: www.efsa.europa.eu.

C. CASE STUDY IN ANIMAL HEALTH AND WELFARE

C.1. RISK OF INTRODUCTION OF RIFT VALLEY FEVER INTO THE SOUTHERN MEDITERRANEAN AREA THROUGH MOVEMENT OF INFECTED ANIMALS

The results of the case study are documented in a separate report:

European Food Safety Authority; 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, 2013:EN-416. [24 pp.].
Available online: http://www.efsa.europa.eu/en/efsajournal/doc/416e.pdf

C.2. LESSONS LEARNT

The purpose of the workshops was to assess the risk of introduction of Rift Valley Fever virus (RVFV) into countries of North Africa and the Near East (Morocco, Algeria, Tunisia, Libya, Jordan, Israel, The Palestinian Territories, Lebanon, Syria; referred to as "region of concern" (RC)) through the undocumented movement of live animals.

The group of experts consisted mainly of epidemiologists and virologists from the RC and Mediterranean EU Members States with large expertise on RVF from research activities, prevention and control of RVFV or networking activities in the RC. The elicitation process followed the Sheffield method.

Prior to the workshop the experts received a briefing document explaining the purpose of the elicitation workshop and a brief explanation of the tasks that experts would be asked to perform. In the workshop, the facilitator gave a presentation in which the task of judging probabilities was explained in more detail. The experts then carried out a practice elicitation, followed by further training on the nature of the aggregated distribution that is the result of an elicitation. The experts were then given an outline of the EFSA risk model and the parameters that they would be asked to elicit.

Overall, the experts felt comfortable with their tasks. However, they expressed their concerns about the difficulty to make judgements on the probabilities of certain parameters. Specifically, the number of animals moved into the RC was considered difficult to quantify. Including animal traders into the group of experts could potentially have enhanced this capability, although it would have been difficult to identify and invite traders involved in undocumented transport of animals.

Nonetheless, the participants showed to be very well acquainted with the epidemiology of RVFV in their area, and knew well the current practices related to trade of animals and traditional farming in their region. The EKE methodology was perceived as a good methodology to elicit the probabilities of parameters needed for the import risk assessment as it is transparent in all its steps and highlights the uncertainties around values of parameters. Furthermore, experts felt that it gives outcomes that are more meaningful than qualitative scores of likelihoods, which may have different meanings for different people.

For the introduction risk model, simulations on the numbers of infected animals entering the RC were derived by using randomly drawn values from the parameter distributions elicited from the experts. Working with distributions assists in avoiding focusing on precise numerical values of individual estimated parameters and better visualises the importance of uncertainty (the range of the distribution) related to the parameters.

The conclusion of the experts' analysis of the risk for introduction of RVFV into the RC through movement of infected animals is that it is likely that infected animals will be imported in outbreak years. Trade flows from RVF endemic areas towards countries adjacent to the RC (i.e. towards Mauritania, Egypt, Yemen and Saudi Arabia) have already led to introduction of RVFV into these countries in the past, and parallel, undocumented trade flows of ruminants towards the RC can be assumed to exist.

D. Case Study in Plant Health

D.1. Technical Documentation of an Expert Knowledge Elicitation on Survival Parameters of Pomacea

Table of contents

D. Case Study in Plant Health 202
D.1. Technical Documentation of an Expert Knowledge Elicitation on Survival Parameters of Pomacea 202
Table of contents 202
D.2. Introduction 204
D.3. Iniation phase 204
D.3.1. Problem definition of the initiation phase 204
D.3.2. Constitution of the Working Group 204
D.3.3. Background report including the risk assessment model 204
D.3.4. Justification and necessary conditions for EKE 204
D.3.5. Evaluation by the corresponding panel and EFSA administration 205
D.4. Elicitation protocol of the pre-elicitation phase 206
D.4.1. Constitution of the Steering Group 206
D.4.2. The final elicitation question 206
D.4.3. Description of the expert selection procedure 207
D.4.4. The decision on the elicitation method including the selection of the Elicitation Group 207
D.4.5. The final project plan for elicitation 207
D.4.6. Evaluation of the elicitation protocol 207
D.5. Result report of the elicitation phase 208
D.5.1. Constitution of the Elicitation Group 208
D.5.2. Documentation of the background material and training sessions given to the expert panel 208
D.5.3. Documentation of the elicitation process 208
D.5.3.1. Time line 208
D.5.3.2. Participant list 209
D.5.3.3. Documentation of Delphi study 209
D.5.3.3.1 First round questionnaire 209
D.5.3.3.2 Second round questionnaire 210
D.5.4. Documentation on the data analysis 210
D.5.5. Anonymised version of expert rationales for their judgements 210
D.5.6. Results for use in risk assessment 210
D.6. Evaluation of the case study experience and lessons learned 210

D.7. Appendix 1: First message to the experts 214

D.8. Appendix 2: Questionnaire on job description and characteristics of work 216

D.9. Appendix 3: Second message to the experts 218
Contacts 218
Letter from the Elicitor 219
Table of contents 220
1. Summary of the Risk assessment Problem 221
Background on the biology of the apple snail 221
1.1. Taxonomy 221
1.1.1. Current status of the taxonomy of *P. canaliculata* and *P. insularum* 221
1.1.2. Taxonomic peculiarities of apple snails found in the Ebro Delta 222
1.2. Life history 222
1.2.1. Definitions 222
1.2.2. Development for *P. canaliculata* (unless otherwise indicated) 223
1.2.3. Survival 224
1.3 References 226
2. Elicitation: Guidance for Experts 229
2.1. Uncertainty and Expert Judgement 229
2.2. Is Elicitation Scientific? 229
2.3. An Overview of the Process 230
2.4. Upper and Lower Bounds 230
2.5. The Median 231
2.6. The Quartiles 232
2.7. What If You Have No Knowledge? 233
2.8. Order Probabilities 233
2.9. Summary 234

D.10. Appendix 4: First Delphi round 235
D.10.1. Sheet 1: “Introduction” 235
D.10.2. Sheet 2: “SummaryDefinitions” 238
D.10.3. Sheet 3: “SummaryTraining” 240
D.10.4. Sheet 4: “Delphi_Round_No1” 243

D.11. Appendix 5: Second Delphi round 250
D.11.1. Summary tables 252

D.2. INTRODUCTION

This appendix reports the experience gained in EFSA on the use of the Delphi method by written procedure. The case study on which this EKE has been applied is a mandate of the Plant Health Panel. It is maybe worth reminding the reader that the aim of this appendix is not to be used as a reference and template for format and approach but as an example and a starting point to help implementing the use of Delphi in an effective and appropriate manner

D.3. INIATION PHASE

D.3.1. Problem definition of the initiation phase

The PLH mandate identified as suitable for the testing phase was a self-task mandate (M-2012-0351; EFSA-Q-2012-0099) devoted to the assessment of the potential establishment of the apple snail (*Pomacea* sp.) in the EU using modelling approaches. The apple snail is one of the largest freshwater snails and is highly invasive outside its native distribution range (South America), as a possible result of its polyphagy and high reproductive rate. Apple snails are considered to be serious rice pests and can cause devastating effects on the flora and fauna of natural freshwater wetlands. In 2010 this species was for the first time found in the European Union (EU), invading rice fields of the Ebro delta (North East Spain). Despite the control measures put in place to eradicate and/or contain the snail in the rice paddies, it is currently spreading not only in the cultivated areas but also in some nearby wetlands, and it has been found moving upwards along the Ebro riverbeds. The PLH Pomacea WG, in order to assess the potential establishment of apple snails in the EU, developed a population dynamics model which included the (air/water) temperatures necessary to the survival of the snail at different life stages. However, empirical information concerning the optimal, minimal and maximal temperature for the development and survival of egg/juvenile/adult stages of different defined for each Pomacea species (in particular *Pomacea canaliculata* and *Pomacea insularum*) is not available. For this reason, this case study was considered suitable for testing the EKE methodology.

D.3.2. Constitution of the Working Group

The Working Group corresponds to the EFSA Plant Health working group on ERA Pomacea: Nils Carlsson, Gianni Gilioli, Johan Coert van Lenteren, Pablo Rafael Martin, Sara Pasquali, Trond Rafoss, Gritta Schrader, Sybren Vos..

D.3.3. Background report including the risk assessment model

The background information have been included in the document initially provided to the elicited experts via email: the full text is available as appendix D.9 to this report. That information can be integrated with some clarifications on the mathematical model included in the PLH opinion, which was designed to *simulate the population dynamics in a spatial unit. In the model, a spatial unit comprises a spatially defined portion of land characterized by physical (meteorological and hydrological variables), ecological (plants, herbivores and predators) and management subsystems (Figure 28). All these subsystems are viewed in their contribution to facilitate / limit establishment, persistence and growth of the snail populations* (EFSA PLH Panel, 2013).

D.3.4. Justification and necessary conditions for EKE

The temperature estimation plays an important role in the population dynamics model included in the risk assessment exercise for *Pomacea* sp. Therefore, the purpose of the elicitation is to obtain better estimates (and the uncertainty around them) for those parameters that could then be used in the model. However, as the experts are asked to give a range of values for each parameter and to indicate which values are more plausible within that range, a condition for them to be able to perform EKE on temperature ranges for *Pomacea* sp. survival is to understand and to perform probabilistic judgements.

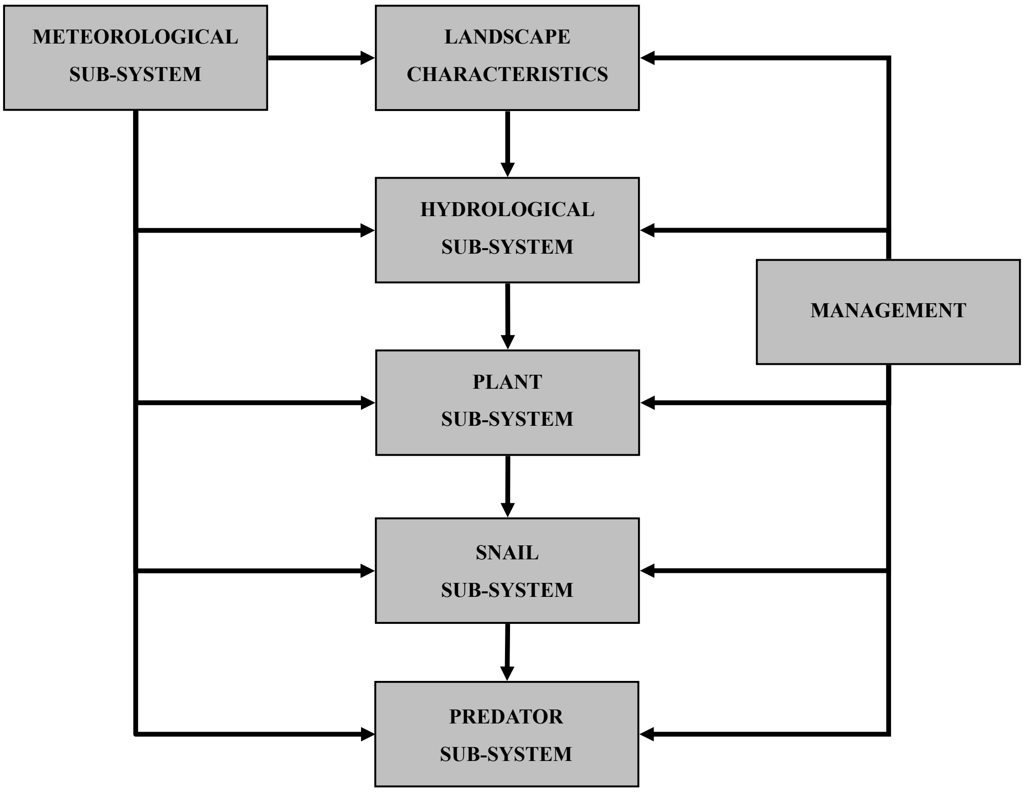


Figure 28: A schematic representation of the *Pomacea canaliculata* population system.

D.3.5. Evaluation by the corresponding panel and EFSA administration

This paragraph should document the evaluation and approval of the project proposal by Panel chair and, usually, head of unit. For this specific case, due to the aim of this exercise (i.e. case study for EKE guidance) the Panel chair was not involved.

D.4. ELICITATION PROTOCOL OF THE PRE-ELICITATION PHASE

D.4.1. Constitution of the Steering Group

The Steering Group (here ahead SG) is composed by:

- Elicitor: Meike Wentholt (EFSA working group on Expert Knowledge Elicitation)
- Trainer: Tony O'Hagan (EFSA working group on Expert Knowledge Elicitation)

The elicitor and trainer are experienced in the elicitation of expert knowledge.

- Plant Health: Trond Rafoss (EFSA Panel on Plant Health)
- Administration: Olaf Mosbach-Schulz (EFSA Scientific Assessment Support Unit), Sara Tramontini and Sybren Vos (EFSA Plant Health Unit)

D.4.2. The final elicitation question

The SG identified among the different temperature ranges needed for the model the temperature for survival (so not growing, reproduction, etc) as the most relevant ones. However, the temperature for survival of Pomacea individuals was expected to differ among eggs, juveniles and adults and the values to be identified should have been minimal, optimal and maximal temperature. Finally, differences between the two main species (*P. canaliculata* and *P. insularum*) should have been highlighted. The combination of all these factors (3 life stages × 3 values of temperature × 2 species) produced the following 18 specific questions:

1. What is the minimal air temperature for eggs survival of *P. canaliculata* species?
2. What is the optimal air temperature for eggs survival of *P. canaliculata* species?
3. What is the maximal air temperature for eggs survival of *P. canaliculata* species?

4. What is the minimal water temperature for youngs survival of *P. canaliculata* species?
5. What is the optimal water temperature for youngs survival of *P. canaliculata* species?
6. What is the maximal water temperature for youngs survival of *P. canaliculata* species?

7. What is the minimal water temperature for adults survival of *P. canaliculata* species?
8. What is the optimal water temperature for adults survival of *P. canaliculata* species?
9. What is the maximal water temperature for adults survival of *P. canaliculata* species?

10. What is the minimal air temperature for eggs survival of *P. insularum* species?
11. What is the optimal air temperature for eggs survival of P. *insularum* species?
12. What is the maximal air temperature for eggs survival of *P. insularum* species?

13. What is the minimal water temperature for youngs survival of *P. insularum* species?
14. What is the optimal water temperature for youngs survival of *P. insularum* species?
15. What is the maximal water temperature for youngs survival of *P. insularum* species?

16. What is the minimal water temperature for adults survival of *P. insularum* species?
17. What is the optimal water temperature for adults survival of *P. insularum* species?
18. What is the maximal water temperature for adults survival of *P. insularum* species?

D.4.3. Description of the expert selection procedure

The experts on the biology of *Pomacea* sp. are very few in the world. For this reason, the SG decided not to apply a full selection procedure but to recruit all those previously identified by the PLH Pomacea working group via direct contact and literature screening. The identified experts composing the experts panel (EP) are listed below:

- Romi Burks (US),
- Nils Carlsson (SE), already member of PLH Pomacea WG,
- Miguel Angel Lopez Robles (ES), already hearing expert of PLH Pomacea WG,
- Pablo Rafael Martín (AR), already member of PLH Pomacea WG,
- Takashi Wada (JP).

D.4.4. The decision on the elicitation method including the selection of the Elicitation Group

The SG could only consider the option of involving the EP remotely, either via webconference or email. However, the presence of experts located in areas with incompatible time zones (Argentina, Japan, Spain, Sweden, US) did not allow the organization of a single webconference. In addition, the limited confidence that some of the experts had on their own understanding of spoken English, drove the SG to opt for a questionnaire delivered by email. The most suitable method appeared to be the Delphi's, through which the experts can individually and according to their needs reflect on the problem and write down their answers and reasoning without direct interactions with the others.

D.4.5. The final project plan for elicitation

The initially defined plan included training on probabilistic judgements and at least two Delphi rounds to be completed in three months. However, the fact to have started the exercise during summer holidays, the time needed to develop a specific tool to carry out the elicitation (Excel file), and the reluctance of certain experts in delivering their replies, slowed down the process.

D.4.6. Evaluation of the elicitation protocol

Due to the fact that this was a case study, and in spite of participation of part of Pomacea WG to the SG, an open evaluation of the protocol developed by the SG was not requested by the Pomacea WG. .Therefore, the step concerning decisions on the allocation of financial and human resources was skept in this occasion though, in a real context, this decision phase would be essential.

D.5. RESULT REPORT OF THE ELICITATION PHASE

D.5.1. Constitution of the Elicitation Group

The experts included in the Elicitation Group (EG) were in this case part of the members of the SG: Olaf Mosbach-Schulz (preparation of the questionnaire template), Tony O'Hagan (analysis of the answers), Sara Tramontini (contact with EP), Meike Wentholt (elicitor).

D.5.2. Documentation of the background material and training sessions given to the expert panel

A first message was sent to the five members of the EP, inviting them to take part to the EKE exercise. The message informed them about the project and included a "Questionnaire on job description and characteristics of work" (Appendix D.8). This questionnaire intended to find out about the nature of the expert's job, and experience with making judgements. Usually this step is necessary to select the experts which will be part of the EP. However, due to the low number of experts available, in this case it was performed at a later stage with the scope to best fit the Delphi protocol.

The questionnaire was compiled and sent back only by two of the EP members, indicating a scarce understanding of the purpose of the questionnaire by the experts.

One month after sending the first message, the EG sent the EP the background information (Appendix D.9).

D.5.3. Documentation of the elicitation process

D.5.3.1. Time line

The time line was agreed by the EG at the beginning of the activity (2 August 2013) as follows:

Step	**Estimation of time needed** Duration	Calendar
Training web session		
1. develop session	1-2 weeks	5-12 August 2013
2. execution (2x)	1 day	7-8 August 2013
First round questionnaire		
1. survey development	1-2 weeks	5-12 August 2013
2. First message to the experts	1 day	7-8 August 2013
3. pilot of survey	1 week	?
4. training	1 day	August 2013
5. send out survey	1 day	August 2013
6. survey out with expert participants	2-4 weeks	August-September 2013
7. send out participant reminder for survey	1 day	September 2013
8. closure of survey and data collation	1 day	September 2013
9. data analysis	1 week	September 2013
Second round questionnaire		
1. survey development	1 day	September-October 2013
2. send out survey	1 day	October 2013
3. survey out with expert participants	2-4 weeks	October 2013
4. send out participant reminder for survey	1 day	October 2013
5. closure of survey and data collation	1 day	October 2013
6. data analysis	1 week	October 2013
…		
After last round		
1. Combine data analysis of all rounds	1 week	November 2013
2. Compile EKE report		November 2013

D.5.3.2. Participant list

Name[1]	Expertise	Institution	Contact details
Expert 5	apple snails, aquatic ecology, invertebrate biology, wetland science		
Expert 2	aquatic ecology, golden apple snail, invasive species, limnology, marine biology, wetland science		
Expert 4	aquatic ecology, island apple snail		
Expert 3	aquatic ecology, malacology		
Expert 1	apple snails, aquatic ecology, invertebrate biology, wetland science		

[1]= answers of the experts are anonymised in the Technical Documentation, for names see D.4.3.

D.5.3.3. Documentation of Delphi study

D.5.3.3.1 First round questionnaire

Structure

The first round of Delphi questionnaire was circulated one month later, in the form of an Excel file (Appendix D.10). In this first step only the optimal temperatures for survival were requested: this would have allowed the experts to focus their replies on one single aspect while gaining experience with the Delphi method. The Excel file sent to the experts was composed by:

D.10.1 Introduction
D.10.2 Short summary of definitions
D.10.3 Training session
D.10.4 1st Delphi round

The questionnaire was created in such a way that the expert had to go through the questions and to provide answers in a predefined order: upper bound, lower bound, median, upper quartile and lower quartile. This decision seemed necessary since the precise wording of the questions and order in which they are asked can contribute greatly to the quality of the resulting judgements.

Responses and analysis

Three out of five experts filled in and sent back to the questionnaire. The missing answers were motivated by time constrains and difficulties in estimating temperatures in the form of Celsius degrees instead of Fahrenheit degrees.

Each response was stored separately[40].

An analysis of the answer was conducted and included in the second round document provided to the EP and is fullt reported in Appendix D.11, in order to allow the reader to understand the reasoning behind the process.

[40] Although in a full technical report this would require to present all the collected answers separately, they are not included in this version in order to make this report more user-friendly and not to duplicate contents, as the same answers are summarized in Appendix D.4.

D.5.3.3.2 Second round questionnaire

Structure

The second round of Delphi questionnaire was circulated 5 months after the first round, again as an Excel file attached to an email message (Appendix D.11), but with few changes from previous

- ✓ summary tables with the answers received from the experts in the previous round with comments from the elicitor were provided, to help the experts in adjusting their judgements in the second step (Appendix D.11.1).
- ✓ questions on minimal and maximal temperatures were added to questions on optimal temperatures (the only already present in the first round).
- ✓ the two questions asking the experts to order probabilities (originally 4 and 5) were removed, as considered not so relevant for the information they were providing
- ✓ a new version of the questionnaire in degree Fahrenheit was made available, in order to answer the specific need of one of the experts, not used to think temperatures in degree Celsius

Responses and analysis

The responses where in part before and in part after the conclusion of the mandate, their analysis has therefore not been included in this report, having the whole process taken more than 6 months more than the hestimated deadline.

D.5.4. Documentation on the data analysis

The documentation and analysis collected along the process are already part of the previous steps in a Delphi protocol, the final results will be reported as soon as available, including interpretation of each expert's rationales modification along the process, changes or consistency of judgements along the process, etc.

D.5.5. Anonymised version of expert rationales for their judgements

The answers received were combined and anonymised by giving each expert a random number from 1 to 5. The same number was then maintained along the exercise, i.e. "expert 2" corresponds always to the same person.

D.5.6. Results for use in risk assessment

The results obtained during the elaboration of this case study could not be used for risk assessment purpose as the timeframe needed to conduct the exercise was not in line with the Pomacea WG needs and deadlines. We expect, however, that final results could be used for adding a sensitivity/uncertainty analysis of the population dynamic model for Pomacea, at a later stage.

D.6. Evaluation of the case study experience and lessons learned

The full process took very long time, for three main reasons

- identification of the right method to test with the Pomacea WG: due to the scope of this specific experience as a case study, the initial phase of presentation and selection of methods to the Pomacea WG required longer than should be needed in real conditions

- creation of the template for the questionnaire and updating of it for each round
- difficulty in motivating the experts: the ones already members of the PLH Pomacea WG could not easily understand the separation of the activities between the two groups and therefore purposes of the EKE exercise, while the ones not participating to the PLH Pomacea WG activities were skeptic about the complex approach proposed for obtaining temperature ranges via email.

Some lessons learned

- The participation of an EKE ambassador in the initial phase (to define feasibility/timelines/resources/method with the specific WG) and of the EKE specialist on the selected method later on (to present the full protocol to the WG) is crucial.
- EKE activity has to be carefully estimated in terms of time and people involvement and integrated in the activity plan of the WG
- Full involvement of a WG member responsible for the model in the EKE process is necessary
- Include some more experts in the EP than needed, as part of them could not conclude the exercise
- The experts have to be convinced of the importance of completing each step of the protocol(including the expertise questionnaire), and of the importance and relevance of the method applied

Future expectations

- clarification on the level of involvement expected for the elicited experts with inclusion/exclusion criteria (do they have to have a DoI approved? Do they need to be outside the EFSA WG?)
- checklist of compulsory aspects for an EKE in EFSA (only quantitative, only for N amount of factors...)
- availability of remote interactive trainings, that will allow to reach the experts at their best convenience and listen/read the training material at all times needed (particularly helpful for non-English mother tongue people)
- motivate and commit experts by favouring their active participation via webconference
- consider involving the experts in a written procedure only after a physical meeting, where the process is explained and they are properly motivated.

The effective timeline

Step	Estimation of time needed Duration	Calendar	Effective time needed
Training web session			
develop session	1-2 weeks	5-12 August 2013	
execution (2x)	1 day	7-8 August 2013	
First round questionnaire			
survey development	1-2 weeks	5-12 August 2013	
preliminary email contact	1 day	7-8 August 2013	
pilot of survey	1 week	?	
training via web/email	1 day	August 2013	
send out survey	1 day	August 2013	
survey out with expert participants	2-4 weeks	August-September 2013	
send out participant reminder for survey	1 day	September 2013	
closure of survey and data collation	1 day	September 2013	
data analysis	1 week	September 2013	
Second round questionnaire			
survey development	1 day	September-October 2013	
send out survey	1 day	October 2013	
survey out with expert participants	2-4 weeks	October 2013	
send out participant reminder for survey	1 day	October 2013	
closure of survey and data collation	1 day	October 2013	
data analysis	1 week	October 2013	
…			
After last round			
Combine data analysis of all rounds	1 week	November 2013	
Compile EKE report		November 2013	

Level of participation from each expert

Name	questionnaire	Step 1	Step 2	Step 3
Expert 1		X	X	
Expert 2	X	X	X	
Expert 3		X	X	
Expert 4				
Expert 5	X			

D.7. APPENDIX 1: FIRST MESSAGE TO THE EXPERTS

On the 13th of August 2013 the following email message and attachment were circulated to the members of the Elicitation Group.

Dear Dr. XXX,

You are probably aware of the work conducted by the European Food Safety Authority (EFSA) on Pomacea sp. presence in Europe. Last year the EFSA Panel on Plant Health published:

- “Statement on the identity of apple snails”
  (http://www.efsa.europa.eu/en/efsajournal/pub/2645.htm)

- “Scientific Opinion on the evaluation of the pest risk analysis on Pomacea insularum, the island apple snail, prepared by the Spanish Ministry of Environment and Rural and Marine Affairs”
  (http://www.efsa.europa.eu/en/efsajournal/pub/2552.htm)

In this context, a working group is currently preparing an environmental risk assessment for the species. Considering your expertise in this field of work, we would like to invite you to take part in a “knowledge elicitation exercise” in order to define some ecological parameters relevant to the Pomacea species.

The problem and the path to a solution

Our working group of experts is studying the potential consequences of the Pomacea genus for the European environment by applying the methodology proposed in the new “Guidance on the environmental risk assessment of plant pests” of the EFSA Panel on Plant Health (PLH) available at http://www.efsa.europa.eu/it/efsajournal/pub/2460.htm .

In order to do so, our working group is developing a population dynamics model for Pomacea spp. In this model several ecological parameters are considered and one appears particularly hard to define: the water temperature for the survival of individuals at different stages of their life (egg, young, adult). We would like to elicit your expertise to define this parameter.

The elicitation procedure

The elicitation procedure will first require reading some written background documentation on the topic followed by instructions on how to participate in the Delphi survey. A Delphi survey can be seen as a series of questionnaires including feedback provided by other participants. After the first round, the subsequent round will repeat the questions but will include as well the responses from other participants, providing you with the possibility to revise your answers in the view of the other results. When no new arguments are obtained from the participants, the procedure will be ended, usually after two to three rounds. Every round will take about one month, wherein you have two weeks for answering the Delphi questionnaire, after which we will develop the subsequent questionnaire.

In preparation of the elicitation procedure we have attached a short questionnaire on some characteristics of your work. We will use the answers to fit the documentation and the Delphi survey as much as possible to your expertise in order to reduce the necessary workload for you. The first Delphi round is planned for September, the final report should be ready by end of this year. In the report the participating experts will be acknowledged without disclosing their individual answers.

Constitution of the stearing group

The elicitation will be conducted by Meike Wentholt as Delphi moderator, Trond Rafoss as representative of the EFSA Working Group on Environmental risk assessment for Pomacea spp., and Sara Tramontini as administrative support from EFSA.

This exercise may provide an opportunity for you to experience a new methodology in the field of information gathering in the context of a pest risk assessment procedure. As well, your participation will be very valuable for EFSA in support to the risk assessment procedure.

We would be very grateful if you could confirm your availability to participate in this elicitation process as soon as possible, by responding to everybody (not only to me).

Looking forward to your positive feedback concerning your availability and interest in participation in our elicitation, I take the opportunity to send you my best regards.

Sara Tramontini

http: //www.efsa.europa.eu

European Food Safety Authority

Save the planet: please don't print this message

D.8. APPENDIX 2: QUESTIONNAIRE ON JOB DESCRIPTION AND CHARACTERISTICS OF WORK

This questionnaire is intended to find out about the nature of your job, and the type of judgements that you make while performing it. These answers will be used to adjust the documentation and Delphi survey on

Temperatures of Survival of *Pomacea* Species

to your needs. In particular, we are interested in whether or not your job requires you to make probabilistic judgements, and how you make such judgements. In addition, we are interested to find out what sort of aids you use in making judgements, whether you received any relevant training, and whether you receive feedback about the quality of your judgements.

Part A: General description of your job

1. What is the title of your job?

---

2. How would you describe your area of expertise?

Part B: The judgements you make (not necessarily related to topic of the expert elicitation)

3. Describe the most important judgements on quantitative parameters that you make on a regular basis in your job.

---

4. 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 statistics, databases or models, etc.)?
Please tick ONE box.

I always use my judgement alone
I mostly use my own judgement
I use partly my own judgement, and partly other sources
I mostly use other sources
I always use other sources alone (not personal judgement)

5. If you do use other information sources, please describe them below.

Part C: Models and feedback

6. In making your work judgements, do you receive any feedback about their accuracy? Please tick ONE box.

Always
Often
Sometimes
Rarely
Never

7. If you receive some feedback, what form does this take?

---

8. Do you make use of a formal model for making your work judgements?
Please tick a box that best represents your opinion.

	1	2	3	4	5	6	7	
Never								Always

9. Did you receive any training to make judgements? If so please describe below.

Part D: Judgments of risk and uncertainty

10. Do you ever make any of the following types of judgements at work (numerically, verbally, or by some other means)? Please tick and fill in as many as are relevant.

    I estimate the likelihood/probability of…________________________________
    I estimate the chances of…____
    I estimate confidence in…____

11. How often, on average, are you called upon to make judgements of risk or uncertainty? Please tick ONE box.

    At least once a day
    At least once a week
    At least once a month
    Less than once a month

12. When you make judgments of risk or uncertainty, what forms do they take? Please tick as many boxes as are relevant.

    Numerical estimates (e.g. 0.5, 50 %, 1 in 2)
    Verbal estimates (e.g. likely, infrequent)
    Comparative (e.g. 'the risk is similar to another risk')

13. If you do make numerical judgements, what forms do these take? Please tick as many boxes as are relevant.

    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 below

---

14. Please give an example of the type of judgement of risk or uncertainty you typically make (if you do make such judgements).

---

15. Did you receive any training to make judgements of risk and uncertainty? If so please describe below.

---

16. When you have to make judgements of risk and uncertainty do you rely on your judgement alone or do you also use other information sources (such as manuals of statistics, computer databases or programs, etc.)? Please tick ONE box.

    I always use my judgement alone
    I mostly use my own judgement
    I use partly my own judgement, and partly other sources
    I mostly use other sources
    I always use other sources alone (not personal judgement)

17. If you do use other information sources, please describe them below.

Thank you for your time and effort.

D.9. APPENDIX 3: SECOND MESSAGE TO THE EXPERTS

On the 10th of September 2013 the following document was provided via email to the members of the Elicitation Group.

EXPERT KNOWLEDGE ELICITATION

Background Information

for Elicitation on

Temperatures for Development and Survival of *Pomacea* species

Steering group on "Case study: Pomacea"[41]

CONTACTS

Elicitor:	Meike Wentholt (EFSA working group on Expert Knowledge Elicitation)
Trainer:	Tony O'Hagan (EFSA working group on Expert Knowledge Elicitation)
	The elicitor and trainer are experienced in the elicitation of expert knowledge.
Plant Health:	Trond Rafoss (EFSA panel on Plant Health)
Administration:	Sara Tramontini, Sybren Vos (EFSA Plant Health Unit) Olaf Mosbach-Schulz (EFSA Scientific Assessment Support)
Email:	sas.expert-elicitation@efsa.europa.eu
Address:	European Food Safety Authority (EFSA), Scientific Assessment Support Via Carlo Magno 1/A, IT-43126 Parma, Italy

[41] The Steering Group is part of the EFSA working group on "Guidelines for Expert Knowledge Elicitation (ExpertKEli) in food and feed safety risk assessment."

LETTER FROM THE ELICITOR

Dear Expert,

We have already been in communication with you regarding your help in conducting a Delphi survey on Temperatures for Development and Survival of *Pomacea* Species. We wish now to formally invite you to take part in a knowledge elicitation exercise — further details of the nature of this exercise are given below.

The elicitation procedure

This elicitation will be conducted using the Delphi method, the first Delphi round is scheduled to start on 19th September 2013. In advance of the survey some preparation is requested, please read the following information:

Summary of the risk assessment problem.

For background information on the topic of the Delphi survey, a summary report of literature on *Pomacea* species is attached in chapter 1.

Elicitation guidance document.

In addition, an elicitation guidance document is attached in chapter 2, which gives you a brief training on the tasks that experts would be asked to perform.

If you have any questions regarding the elicitation, the summary or the guidance don't hesitate to ask us. Please reply always to our functional mailbox below. The responsible team member will contact you with more explanations.

All questions received before **Thursday, the 12th September 2013**, will be answered before the first Delphi round starts.

Please contact the Steering Group at the following e-mail address:
sas.expert-elicitation@efsa.europa.eu

Kind regards on behalf of the Steering Group,

Meike Wentholt
(Elicitor)

TABLE OF CONTENTS

Contacts 218
Letter from the Elicitor 219
Table of contents 220
1. Summary of the Risk assessment Problem 221
Background on the biology of the apple snail 221
1.1. Taxonomy 221
1.1.1. Current status of the taxonomy of *P. canaliculata* and *P. insularum* 221
1.1.2. Taxonomic peculiarities of apple snails found in the Ebro Delta 222
1.2. Life history 222
1.2.1. Definitions 222
1.2.2. Development for *P. canaliculata* (unless otherwise indicated) 223
1.2.3. Survival 224
2. Elicitation: Guidance for Experts 229
2.1. Uncertainty and Expert Judgement 229
2.2. Is Elicitation Scientific? 229
2.3. An Overview of the Process 230
2.4. Upper and Lower Bounds 230
2.5. The Median 231
2.6. The Quartiles 232
2.7. What If You Have No Knowledge? 233
2.8. Order Probabilities 233
2.9. Summary 234

1. Summary of the Risk assessment Problem

Note from the Steering group:

The following chapter is an extract from the current draft of the EFSA Panel on Plant Health Scientific Opinion on "Assessment of the potential establishment and spread of the apple snail in the EU using modelling approaches", which will be published in EFSA Journal after finalization and adoption in 2014. The draft version is under preparation. The EFSA Panel on Plant Health restricted the use of this extract to use only for the elicitation exercise – do not cite – do not circulate.

If you need access to the references stated in the summary please inform us.

Background on the biology of the apple snail

The current scientific studies on the systematics and taxonomy of the genus *Pomacea*, acknowledge that there is still a high degree of uncertainty with considerable dynamics regarding the differentiation of species and their identification, which was earlier described in the Panel's statement on the identity of apple snails (EFSA 2012a). The Panel therefore aims to make these dynamics as transparent as possible by summarizing the most recent literature. General aspects on the biology of the genus *Pomacea* were provided in the Panel's evaluation of the Spanish PRA (EFSA 2012b). For the purpose of this opinion, which focuses on *P. canaliculata* and *P. maculata* the taxonomy, life history and spread are presented in greater detail than in the previous opinion. Available information about *P. maculata* is much more limited than about *P. canaliculata*.

1.1. Taxonomy

1.1.1. Current status of the taxonomy of *P. canaliculata* and *P. insularum*

Hayes et al. (2012) synonymize *Pomacea insularum* and *Pomacea gigas* (*Ampullaria insularum* d'Orbigny, 1835 and *Ampullaria gigas* Spix, 1827) with *Pomacea maculata* Perry, 1810 and clearly discriminate the latter from *Pomacea canaliculata*. They designate neotypes for *P. maculata*, *P. canaliculata* and *A. gigas*, as well as a lectotype for *A. insularum*.

The conchological differentiation of *Pomacea maculata* and *P. canaliculata* is difficult, since differences in shell morphology are only most obvious in recently hatched juveniles. Shells of adults differ primarily in the angulation of the whorl shoulder and pigmentation of the inner apertural lip. When present, this pigmentation is a distinctive feature of *P. maculata*. However, sometimes it is lacking or too faint (e.g. in juvenile *P. maculata*) which can result in misidentification (Martín personal communication), leading to wrong information on their biology, spread and impact. Hayes et al. (2012) therefore made another attempt to clarify the taxonomy, describing their morphological and genetic distinctiveness, and re-evaluating their biogeographic ranges. Their results show that the two species differ most clearly genetically, with no shared haplotypes and a mean genetic distance of 0.135 at cytochrome c oxidase subunit I (COI).

Some other features described by Hayes et al. (2012) to distinguish the two species are the number of eggs per clutch, which is higher in *P. maculata*, and the individual eggs are smaller. Barnes et al. (2008) found that the average number of eggs in a *P. insularum* egg clutch was 2064 – almost ten times higher than the average number of eggs in a *P. canaliculata* egg clutch (Teo, 2004; Martín and Estebenet, 2002). The smaller egg size (determined by egg weight) was confirmed by Matsukura et al. (2013). *P. canaliculata* hatchlings are nearly twice as large as those of *P. maculata*. They also differ in reproductive anatomy. *P. canaliculata* has two distinctive glandular tissues in the apical penial sheath gland, and *P. maculata* has a basal sheath gland instead of a medial sheath gland.

Matsukura et al. (2013) found that *P. maculata* and *P. canaliculata* hybridize both in Southeast Asia and in the lower Río de la Plata basin, the area identified as the origin of multiple introductions of *P. canaliculata* and *P. maculata* (Hayes et al., 2008), although some introductions from Brasil were also likely for the latter. Similar origins were suggested for the continental United States of America by Rawlings et al., (2007). Matsukura et al. (2013) determined two well-supported clades (Clade C and Clade M), in which both species were represented, and some specimen had both Clade C and Clade M EF1a sequences, suggesting genetic exchange between the two clades. A mating experiment between *P. canaliculata* from Clade C and *P. maculata* from Clade M produced viable F1 progeny under laboratory conditions. The genetic exchange was also inferred in some populations collected from Argentina, hinting at hybridization in the native range.

Andree and Lopez (2013) investigated the feasibility of DNA extraction from empty shells to overcome the difficulties of differentiating the species by shell morphology and its plasticity and to improve genetic analysis from field samples. Their method was successful, so that now a distinction of species by empty shells is possible.

1.1.2. Taxonomic peculiarities of apple snails found in the Ebro Delta

In 2009, López et al. genetically identified 9 specimen with high variability in shell colour and shape, some of them consistent with the morphology of *P. canaliculata*, only one haplotype ("O" haplotype), which was clearly *P. insularum*. This haplotype is typical for cultivated apple snails. Due to the fact that morphological and reproductive features are similar to *P. canaliculata* but that genetic identification (COI) cleary points at *P. insularum*, López (Generalitat de Catalunya) and Andree (IRTA; personal communication) suggested that hybridization with *P. canaliculata* or a more variable phenotypic variation being believed until now to be *P. insularum*, could explain the apparent disagreement between genotype and phenotype. This hypothesis is further supported by the fact that the mean number of eggs per clutch found in the Ebro Delta (unpublished data?) is intermediate between the mean numbers of eggs for the two species published in the literature.

1.2. Life history

1.2.1. Definitions

The Egg stage starts when eggs are produced.

The Juvenile stage starts when eggs are hatched.

The Adult stage starts at 1st reproductive event.

The threshold T° for cumulating of degrees corresponds to the threshold for the total sum to complete the development from one step of the life cycle to the next one.

In the table below you can find the definitions to be considered by the experts for the Temperature estimates:

Stage	Optimal T°	Min T°	Max T°	Threshold T° for cumulating of degrees
Egg	The temperature at which embryonic development time from egg spawn until hatching is the shortest	The lowest temperature for embryonic development	The highest temperature at which embryonic development can occur	Equals the minimum T°, used to set the threshold when calculating day-degree sum for complete embryonic development from egg spawn until egg hatch.
Juvenile	The temperature at which juvenile development time from egg hatching till the first reproductive event is shortest	The lowest temperature for juvenile development	The highest temperature at which juvenile development can occur	Equals the minimum T°, used to set the threshold when calculating day-degree sum to complete the juvenile stage from egg hatch till the first reproductive event
Adult	The temperature at which the development rate is the highest in terms of growth starting from the first reproduction event	The lowest temperature for adult development	The highest temperature for adult development/activity	

1.2.2. Development for *P. canaliculata* (unless otherwise indicated)

A lot of recent information from mainland China is available, however this must be considered with caution regarding the identity of the snails, as the taxonomic situation is complicated and unresolved there. For example, Lv et al. (2013) stated: Phylogenetic analyses indicate that *P. canaliculata*, *P. insularum* and two cryptic groups, discovered by the present and previous studies, coexist in the mainland of P.R. China. The mosaic distribution and the high diversity found in the collection sites suggests multiple and secondary introductions […]".

1.2.2.1. Eggs

The minimum temperature for the development of eggs of *P. canaliculata* in mainland China is 14.2°C according to Huang et al. (2010) and Liu Yan-Bin et al. (2011). Seuffert et al. (2012) found ranges from 15.7 to 16.2 °C in *P. canaliculata* from Argentina.

Liu et al. (2011) mentioned the following egg production quantities: 30°C (4200 eggs)> 25°C (2350 eggs)> 35°C (680 eggs)> 20°C (540 eggs). No eggs were produced at 15 and 40°C. This indicates that the optimum temperature for spawning *P. canaliculata* is about 30°C. In contrast to this, Seuffert and Martín (2013) found that snails under a constant temperature of 35°C died without laying eggs, and those reared at 30°C laid many egg masses but their hatchability was quite low. Probably the tolerance of Chinese snails from Zhejiang (30° 16′ N) to high temperature is a little higher than that of snails from Southern Pampas (38° S).

The accumulation of degree days until hatch is 137.41°C.days according to Liu Yan-Bin et al. (2011) and 152.16°C days according to Huang et al. (2010). Seuffert and Martín (2012) report 120.7 to 133.8 °C.days.

Egg development time is 20.7 days on average (Liu et al. 2012). The longest egg development time found by Liu et al. (2012) was 62 days at temperatures around 16°C and the shortest was 5 days at temperatures around 32°C. In the native range (Argentina) egg development took between 13 and 24 days under mean daily temperatures between 14.9 and 25.5°C, respectively (Pizani et al., 2005).

1.2.2.2. Juveniles

Juveniles emerge from eggs and change into reproductive adults, whereby the size and age at maturity are highly variable according to food availability. Also, they are different for both sexes. Tamburi & Martín (2009) reared *P. canaliculata* snails from hatching in a wide gradient of food availability, between 20 and 100% of the *ad libitum* ingestion rate, to investigate this effect. They found that males matured at an age of around 13 weeks irrespective of food availability but their size was highly dependent of food availability: 16 mm at 20% and 29 mm at 100%. For females it was necessary to reach sizes of at least 32 mm to reproduce. These sizes were attained at very different ages according to food availability: 15 weeks at 100% and 50 weeks at 20% of the *ad libitum* ingestion rate.

The optimal temperature for juvenile development according to Liu Yan-Bin et al. (2011) is 30°C. Growth rates increase with temperature from 15 to 25°C but the growth rates at 30 and 35, even though a little higher, were not significantly different from those at 25°C.

Seuffert and Martín (2013) found that at 15 and 20°C there was no mortality but growth rates were very low. In contrast, at 25, 30 and 35°C snails grew faster but survival decreased with increasing temperature. After 10 weeks, the mean shell lengths at temperatures of 30 and 35°C were not significantly different from those at 25°C.

1.2.2.3. Adults

Liu et al., 2011 considered that the optimum temperature for growth, development and reproduction of *P. canaliculata* is about 30°C according to their studies. Gettys et al. (2008) indicate 20-30°C for adult development of *P. insularum*.

Seuffert and Martín (2013) found that at a constant temperature of 35°C *P. canaliculata* females will not lay eggs (unpub. results from a long term study, whose preliminary findings were published recently).

One to three generations may develop per year, depending on the temperature, especially in winter. The average number of days for females to reach sexual maturity in the south of Hunan Province, where winter temperatures are above 9°C, was 59.3 for the first, 45.4 for the second, and 213.0 for the third generation being the longest since they reach sexual maturity the year following the winter (Liu et al., 2012).

1.2.3. Survival

1.2.3.1. Eggs and juveniles

Yingying et al. (2008) mentioned survival rates for juveniles of 97% at 30°C, of 87% at 33°C, and of 47% at 36°C. At 39 and 42°C juvenile mortality was 100%.

The survival rate of juveniles at 15°C was highest (100%), declined to 63% at 12°C, while only 7% of juveniles survived at 9°C and no snails survived at 6°C. The shortest survival time was 2 days and the longest was 7 days (average 4.10±0.24 days). LT50 at 6 and 9°C was 4 and 24 days respectively. Young snails survived 1 – 6 days (average survival time: 2.57±0.32 days (X±SE) at 42°C; and 1-13 days at 39°C (average survival time 6.27±0.45 days). Median lethal time (LT50) at 36, 39 and 42°C was 21, 6 and 2 days respectively.

This results in the following ranking of survival of juveniles: 15 > 30 > 33 > 12 > 36 > 9 > 6 = 39 = 42.

Liu Yan-Bin et al. (2011) ranked survival rates of juvenile *Pomacea canaliculata* from high to low reared at different temperatures as: 20°C > 15°C > 35°C > 30°C > 25°C > 40°C. Seuffert and Martín (2013) have found a 100% survival of juveniles at 15°C and 20°C after ten weeks and all snails at these two temperatures were still alive after 18 months (Martín, personal communication).

The above data are more or less similar to what Yingying et al. (2008) found, but there are also some obvious differences, e.g. 100% survival at 15°C found by Yingying et al. (2008) and Seuffert and Martín (2013) and less than 100% by Liu Yan-Bin et al. (2011).

1.2.3.2. Adults

Howells et al. 2006 stated that members of the *P. canaliculata* complex often have a greater tolerance for low temperatures than e.g. *Marisa cornuarietis*, *Pomacea paludosa*, and *P. bridgesii*. Oya et al. (1987) and Mochida (1991) reported that *P. canaliculata* could survive for 15-20 days at 0°C, for 2 days at –3°C, and for 6 hours at –6°C. Furthermore, Oya et al. (1987) and Syobu et al. (2001) found that younger and smaller snails tolerated cold better than larger snails. However, most studies in Japan show an opposite trend or an optimum size for tolerance, for instance: "Juveniles of intermediate size were more tolerant of cold temperature than very small juveniles and adults" (Wada and Matsukura, 2007). Field data and laboratory experiments in Southern Japan over seven consecutive years showed that snails larger than 6.0mm exhibited higher cold tolerance than small snails (Syobu et al. 2001). Yingying et al. (2008) found that exposure of *P. canaliculata* to 6°C for 7 days caused 100% mortality with a median lethal time (LT50) of 4 days, whereas field tests by Yu et al (2002) showed that their mortality was 100% when exposed to 1~2°C for only 1 day.

Under laboratory conditions, the supercooling point of cold-acclimated and non-acclimated snails was not found to differ significantly and was about - 7°C. Snails did not survive freezing and even died under more moderately low temperatures approaching 0°C (Matsukura et al., 2009). It was found that only one snail out of 80 (1.3%) without cold-acclimation survived a 0°C cold treatment for 5 days, whereas 98.8% of snails with cold acclimation survived under moist conditions and 93.8% under aquatic conditions at these 0°C. In a desiccation-tolerance test, the survival rate after 4 weeks of the start of the experiment of non cold-acclimated snails (71.3%) was significantly lower compared to two groups of cold acclimated snails (approx. 90%). The difference in survivorship was even higher after 8 weeks (Wada and Matsukura, 2011).

Estebenet and Cazzaniga (1992) recorded a maximal longevity under laboratory conditions of 49.5 months at room temperatures fluctuating between 9 to 29°C (Martín, personal communication). Under a constant temperature of 25ºC, maximal longevity was 13.5 months.

Under field conditions, *P. canaliculata* has been found in Paso de las Piedras reservoir (38°04'S – 59°18'W), in the South of Buenos Aires province, Argentina, as the most southerly location (Martín et al., 2001), with a mean annual air temperature of 14ºC and thermal amplitudes of up to 18ºC (Estebenet and Martín 2002). In the northern hemisphere, *P. canaliculata* could not survive in mountain areas of Guangdong during the winter (He et al., 2012).

According to Liu et al. (2011), survival rates of adult *Pomacea canaliculata* reared at different temperatures could be ranked as 20°C =15°C > 25°C > 35°C > 30°C > 40°C. When *Pomacea canaliculata* was reared at temperatures below 40°C, survival rates of both juveniles and adults were lower than 15% after 10 days, and lower than 5% after 15 days, and adults died faster than juveniles; the survival rate of adult snails was also lower than that of the young snails at 30~35°C. Therefore, it can be considered that the high-temperature tolerance of young snails was slightly stronger than in adult snails. When *Pomacea canaliculata* was reared at temperatures of 15~35°C, survival rates among both young and adult snails rose to 70% or above, and the survival rate of *Pomacea canaliculata* reared at 15~20°C for 30 days was up to 90% or above, but activity was weak, food intake was low, and growth and development was slow, while those reared at 25~35°C developed normally.

Mochida (1991) found a high mortality at water temperatures above 32°C, however, this could not be confirmed by Heiler et al. (2008). In fact, the effect of temperature on the mortality of this “amphibious” snail depends strongly on water conditions (fouling and oxygen). In aquaria with food but without artificial aeration, survival times were significantly reduced when access to breathable air was blocked by underwater barriers: at 35°C the mean survival time was less than 2 days (Seuffert and Martín, 2010).

The snails can survive for 15-20 days at 0°C, 2 days at -3°C, but only 6 hours at -6°C. In Okinawa, Japan, it has been confirmed that the snails can survive 234 days without water (Mochida, 1991).Yusa et al. (2006) even found a longer survival time: without watering, five large snails (out of 30 individuals; 17%) survived up to 11 months (approx. 330 days), but no snails survived longer under such dry conditions. Under moist conditions with watering 2–3 times per month, one medium-sized (out of 30 individuals; 3%) and two large snails (out of 50; 4%) survived the entire experimental period of 29 months (approx. 870 days).

The temperature limits for this invasive species have been studied previously by Mochida (1991) who determined a high mortality at water temperatures above 32°C and a life span of only 15-20 days at temperatures of 0°C. Lee and Oh (2006) described the temperature limits as 2°C and 38°C.

1.3 References

Andree KB and López MA, 2013. Species identification from archived snail shells via genetic analysis: a method for DNA extraction from empty shells, Molluscan Research, 33:1, 1-5

Barnes MA, Fordham RK and Burks RL, 2008. Fecundity of the exotic apple snail, *Pomacea insularum*. Journal of the North American Benthological Society 27, 738–745.

EFSA Panel on Plant Health (PLH), 2012a. Statement on the identity of apple snails. EFSA Journal 2012;10(4):2645. [278 pp.] doi:10.2903/j.efsa.2012.2645. Available online: www.efsa.europa.eu/efsajournal

EFSA Panel on Plant Health (PLH), 2012b. Scientific Opinion on the evaluation of the pest risk analysis on *Pomacea insularum*, the island apple snail, prepared by the Spanish Ministry of Environment and Rural and Marine Affairs. EFSA Journal 2012;10(1):2552. [278 pp.] doi:10.2903/j.efsa.2012.2552. Available online: www.efsa.europa.eu/efsajournal

Estebenet AL and Cazzaniga NJ, 1992. Growth and demography of *Pomacea canaliculata* (Gastropoda: Ampullariidae) under laboratory conditions. Malacological Review 25, 1–12.

Estebenet AL and Martín PR, 2002. Workshop: “Biology of Ampullariidae” – Minireview – *Pomacea canaliculata* (Gastropoda: Ampullariidae): life-history traits and their plasticity. Biocell 26, 83–89.

Hayes KA, Cowie RH, Thiengo SC and Strong EE, 2012 Comparing apples with apples: clarifying the identities of two highly invasive Neotropical Ampullariidae (Caenogastropoda). Zoological Journal of the Linnean Society, 2012, 166, 723–753

Hayes KA, Joshi R, Thiengo SC and Cowie RH, 2008. Out of South America: multiple origins of non native apple snails in Asia. Diversity and Distributions 14, 701–712.

He, M., J. Zhang, et al. (2012). "Potential spatio-temporal distribution of golden apple snail and the harm risk on paddy fields in Guangdong based on GIS technology." Journal of South China Agricultural University 33(3): 288-295.

Heiler, K., P. v. Oheimb, et al. (2008). "Studies on the temperature dependence of activity and on the diurnal activity rhythm of the invasive Pomacea canaliculata (Gastropoda: Ampullariidae)." Mollusca 26(1): 73-81.

Howells R, Burlakova LE, Karatayev AY, Marfurt RK and Burks RL, 2006. Native and introduced Ampullaridae in North America: History, status, and ecology. In In Joshi RC & Sebastian LS:

GLOBAL ADVANCES IN ECOLOGY AND MANAGEMENT OF GOLDEN APPLE SNAILS. Philippine Rice Research Institute, Science city of Munoz. 588p

Huang, P., Z. Lin, et al. (2010). "Biological characteristics of the egg stage of Pomacea canaliculata." Journal of Fujian Agriculture and Forestry University (Natural Science Edition) 39(1): 25-29.

Lee, T.-G. and K. Oh (2006). "Golden apple snails in Korea." In Joshi RC & Sebastian LS: GLOBAL ADVANCES IN ECOLOGY AND MANAGEMENT OF GOLDEN APPLE SNAILS. Philippine Rice Research Institute, Science city of Munoz. 588p

Liu, J., Y.-J. He, et al. (2012). "Characteristics of Pomacea canaliculata reproduction under natural conditions." Ying yong sheng tai xue bao = The journal of applied ecology / Zhongguo sheng tai xue xue hui, Zhongguo ke xue yuan Shenyang ying yong sheng tai yan jiu suo zhu ban 23(2): 559-565.

LIU, Yan-bin; HAN, Wei; XIAN, Zhen-hua. 2011. Effect of different temperatures on growth, development and feeding of *Pomacea canaliculata*. Journal of Southern Agriculture

López, M.A., Altaba, C.R., Andree, K.B. & López, V. (2009) First invasion of the apple snail Pomacea insularum in Europe. Tentacle 18, 26–28.

Lv, S., Zhang, Y., Liu, H. X., Hu, L., Liu, Q., Wei, F. R., ... & Utzinger, J. (2013). Phylogenetic evidence for multiple and secondary introductions of invasive snails: *Pomacea* species in the People's Republic of China. Diversity and Distributions, 19(2), 147-156.

Martín PR and Estebenet AL, 2002. Interpopulation variation in life-history traits of *Pomacea canaliculata* (Gastropoda: Ampullariidae) in southwestern Buenos Aires Province, Argentina. Malacologia, 44, 153–163.

Martín PR, Estebenet AL and Cazzaniga NJ, 2001. Factors affecting the distributionn of *Pomacea canaliculata* (Gastropoda: Ampullariidae) along its southernmost natural limit. Malacologia, 2001, 43(1-2):13-23.

Matsukura K, Okuda M, Cazzaniga NJ, Wada T, 2013. Genetic exchange between two freshwater apple snails, Pomacea canaliculata and Pomacea maculata invading East and Southeast Asia. Biol Invasions (2013) 15:2039–2048

Matsukura, K., H. Tsumuki, et al. (2009). "Physiological response to low temperature in the freshwater apple snail, Pomacea canaliculata (Gastropoda: Ampullariidae)." Journal of Experimental Biology 212(16): 2558-2563.

Mochida, O. 1991. Spread of freshwater *Pomacea* snails (Mollusca: Pilidae) from Argentina to Asia. Micronesia 3(supplement):51-62.

Oya S, Hirai Y and Miyahara Y, 1987. Overwintering of the apple snail, *Pomacea canaliculata* Lamarck, in north Kyushu, Japan. Japanese Journal of Applied Entomology and Zoology 31, 206–212.

Pizani, N. V., A. L. Estebenet, and P. R. Martín, 2005. "Effects of submersion and aerial exposure on clutches and hatchlings of *Pomacea canaliculata* (Gastropoda: Ampullariidae)." American Malacological Bulletin 20.1-2 (2005): 55-63.

Rawlings TA, Hayes KA, Cowie RH and Collins TM, 2007. The identity, distribution, and impacts of non-native apple snails in the continental United States. BMC Evolutionary Biology 7, 97.

Seuffert, M. E., & Martín, P. R. (2010). Dependence on aerial respiration and its influence on microdistribution in the invasive freshwater snail *Pomacea canaliculata* (Caenogastropoda, Ampullariidae). Biological Invasions, 12(6), 1695-1708.

Seuffert, M. E., & Martín, P. R. (2012). A lentic dweller in lotic habitats: the behavior of the invasive South American apple snail *Pomacea canaliculata* in flowing water. Aquatic Ecology, 46(1), 129-142.

Seuffert, M. E., & Martín, P. R. (2013). Juvenile growth and survival of the apple snail *Pomacea canaliculata* (Caenogastropoda: Ampullariidae) reared at different constant temperatures. SpringerPlus, 2(1), 312.)

Seuffert, M. E., Saveanu, L., & Martín, P. R. (2012). Threshold Temperatures and Degree-Day Estimates for Embryonic Development of the Invasive Apple Snail *Pomacea canaliculata* (Caenogastropoda: Ampullariidae). Malacologia, 55(2), 209-217.

Spanish Ministry of Environment and Rural and Marine Affairs, 2011. Pest Risk Analysis on the introduction of *Pomacea insularum* (d'Orbigny, 1835) into the EU. April 2011. Prepared by the Spanish Ministry of Environment and Rural and Marine Affairs.

Syobu, S., H. Mikuriya, et al. (2001). "Estimating the overwintering mortality of the apple snail, Pomacea canaliculata (Lamarck) (Gastropoda : Ampullariidae) in a paddy field of southern Japan using temperature data." Japanese Journal of Applied Entomology and Zoology 45(4): 203-207.

Tamburi, N. E. & P. R. Martín, 2009a, Reaction norms of size and age at maturity of Pomacea canaliculata (Gastropoda: Ampullariidae) under a gradient of food deprivation. Journal of Molluscan Studies, 75: 19–26.

Tamburi, N. E. & P. R. Martín, 2009b, Feeding rates and food conversion efficiencies in the apple snail Pomacea canaliculata (Caenogastropoda: Ampullariidae). Malacologia, 51:221–232

Teo SS, 2004. Biology of the golden apple snail, *Pomacea canaliculata* (Lamarck, 1822) with emphasis on responses to certain environmental conditions in Sabah, Malaysia. Molluscan Research, 2004, 24, 139-148.

Wada and Matsukura, 2011. Linkage of cold hardiness with desiccation tolerance in the invasive freshwater apple snail, *Pomacea canaliculata* (Caenogastropoda: Ampullariidae). J. Mollus. Stud. (2011) 77 (2): 149-153. doi: 10.1093/mollus/eyq049 First published online: February 18, 2011.

Wada, T. and K. Matsukura (2007). "Seasonal changes in cold hardiness of the invasive freshwater apple snail, Pomacea canaliculata (Lamarck) (Gastropoda : Ampullariidae)." Malacologia 49(2): 383-392.

Yingying P; Xiaoping D and Shengzhang Y, 2008. Effects of temperature stress on development, feeding and survival of the apple snail, *Pomacea canaliculata* (Lamarck). Acta Phytophylacica Sinica, 2008, vol. 3, p. 009.

Yu XP, Li XF, Lu ZX, et al. 2002. Occurrence of golden apple snail, *Pomacea canaliculata* (Lamarck) in paddy fields and its management in China. // Joshi RC, Wada T. Proceeding of the special working group on the golden apple snail (Pomacea spp.) at the 7th Congress on Medical and Applied Malacology. Philippines, 2002)

Yusa, Y., Wada, T., & Takahashi, S. (2006). Effects of dormant duration, body size, self-burial and water condition on the long-term survival of the apple snail, Pomacea canaliculata (Gastropoda: Ampullariidae). Applied entomology and zoology, 41(4), 627-632.

2. Elicitation: Guidance for Experts

Tony O'Hagan

This document presents guidance and instructions to experts participating in the European Food Safety Authority's elicitation for survival temperatures of *Pomocea* species. This document provides general guidance on the elicitation process – a separate document will be sent to you with the specific questions for the *Pomacea* elicitation.

Please read this guidance carefully and study the worked example in order to understand what is required from you. If you have any questions for clarification, please email them to EFSA before attempting to complete the elicitation questionnaire.

2.1. Uncertainty and Expert Judgement

You will be asked to provide your knowledge about a number of uncertain parameters (various survival temperatures for *Pomacea sp.*). It is important first of all to concentrate on the fact that these parameters are uncertain. Nobody knows their true values. We are asking you because, as an expert in the field, you may be expected to have less uncertainty than EFSA's own staff, but we do not expect you to know the true values.

The parameters play an important role in a population dynamics model that is part of EFSA's current risk assessment exercise for *Pomacea*, and one purpose of the elicitation can be seen as to obtain best estimates for these parameters for use in the model. However, uncertainty is a key component of risk, and EFSA wishes also to assess the uncertainty surrounding its analyses. Therefore, you will not be asked simply to estimate these parameters. Instead you will be asked to give a range of values for each parameter and to indicate which values are more plausible within that range.

We are asking you to make your own personal judgments based on your own knowledge and expertise. You will surely give different answers from the other experts participating in this exercise. There are no absolute right or wrong answers and no rewards or penalties – the right answers *for you* are those which honestly express *your* best professional judgements. By combining the differing judgements of several experts in the field, EFSA will gain an overview of the scientific knowledge in the field, and for this purpose it is essential that we obtain a good representation of each expert's opinions regarding these parameters.

2.2. Is Elicitation Scientific?

Some experts are uncomfortable with the idea of giving their personal judgements; they feel that this process is somehow unscientific. First bear in mind that whenever there is adequate scientific evidence to provide good statistical estimates of parameters, with statistically-sound measures of the uncertainty in those estimates, then of course these would be used in any risk model – elicitation would indeed be unscientific in such cases. Unfortunately, we frequently require parameters for which the evidence is not so strong. The evidence may be conflicting or there may be questions regarding its quality. Often the available evidence is only indirectly relevant to the parameter of interest; it concerns a different (but similar) parameter or its relevance is conditional on assumptions which may not hold in practice.

When evidence is weak, but we need to make decisions, we cannot wait for better data to arrive. In problems of this type (and practical risk assessments routinely fall into this category), elicitation of expert knowledge is the scientific solution. Expert judgements are of course imperfect, but methods of elicitation have been extensively studied and found to give valuable information in a wide range of contexts. The methods we will be using have been designed to minimise the biases and distortions that can arise in expert judgement and accord with EFSA's own guidance on best practice in Expert Knowledge Elicitation.

2.3. An Overview of the Process

For each parameter, we will ask you to state five numbers. First, you will give an upper bound (U) and a lower bound (L). Next you will give an estimate known as the median (M). Finally, you will give two numbers called the upper and lower quartiles (Q3 and Q1) which provide a range around the median that you judge to be the most likely values for the parameter. These are only rough descriptions of the five numbers, to give you an idea of the process. It is *most important* that you read carefully the more precise descriptions later in this guidance.

> Example: The following example will be used to illustrate each step of a five-number elicitation: an Expert is asked for her judgements about the distance (in kilometres) between the airports of Paris Charles de Gaulle and Chicago O'Hare. This distance will be referred to as D. The Expert is a frequent flyer whose knowledge of distances between airports comes from the "air miles" that she receives on her frequent flyer account. However, she has not flown between these particular two airports before. The Expert will give five numbers to complete this table:

Parameter	**L**	**Q1**	**M**	**Q3**	**U**
D					

You will also be asked to provide some text describing your reasons for the judgements you have given. This is your opportunity to highlight any specific data or experience you have that has led you to choose the numbers you have given.

The elicitation is planned to be conducted in three rounds. In round one, you will provide your judgements about six uncertain parameters. You will then be given feedback on your judgements and on the explanations that the other experts have given for their judgements. In round two, you will be asked to provide judgements about the full set of eighteen uncertain parameters. This will be an opportunity to revise your round one judgements if you wish to, in the light of information from the other experts, and to give new judgements about the other twelve parameters. In round three, you will again receive feedback about what the other experts have said, and may revise your round two judgements about all of the parameters.

2.4. Upper and Lower Bounds

The process for each parameter begins with specifying two numbers called the upper bound, U, and the lower bound, L. These are not intended to be absolute or theoretical bounds. For instance, the absolute lower bound for the example distance D is zero, but this is not what is required for the elicitation lower bound L.

The values you give for U and L should be limits within which you are *almost certain* that the parameter will lie. Whilst it might be *theoretically possible* for the true value of the parameter to be above U or below L, you would be *extremely surprised* if it did so.

We call the range from L to U the *plausible range* for the parameter. Remember that this is the plausible range *in your opinion*. U and L are judgements that reflect one aspect of your knowledge about the parameter, and the difference between them provides an overall indication of the strength of your knowledge. If your U and L are very far apart, this tells EFSA that your information is weak, while if they are very close together it indicates that your knowledge is strong.

You should not specify values too close together so as to imply more information than you really have. The following thought experiment may help: imagine that a future experimental study claimed to show that the true value of the parameter was larger than U (or smaller than L) – you should be extremely surprised, to the extent that you would believe the experiment must have been flawed in design or execution. If you could think of a plausible reason why that experimental value might be correct then your bounds are not wide enough.

On the other hand, your bounds should not be too wide. Your plausible range should not include values that you regard as really implausible.

Example: Our Expert has not flown this route before but has flown from London to New York and knows from her frequent flyer account that this distance is about 5500 km. She is very confident that the distance between Paris and Chicago must be at least as much. She would regard it as implausible that D should be less than 5500. However, she is aware that flight paths follow great circle lines and distances on the usual map projections may not be a good guide. She feels it is not implausible that D should in fact be close to 5500 and so sets L = 5500. She feels that the sum of the distances from Paris to London and from New York to Chicago must be no more than 2500 km, and the flight from Paris to Chicago must take a route that is no longer than flying via London and New York. So her upper bound of plausibility is 8000. Putting these figures in the table, she now has

Parameter	**L**	**Q1**	**M**	**Q3**	**U**
D	5500				8000

2.5. The Median

The next step is to specify your median value, M. You can think of this as an estimate of the parameter, but it is an estimate in a specific sense. Your judgement should be that it is equally likely that the parameter would be above M or below M.

The following thought experiment may help with this task: imagine that you are asked to predict whether the true value will be above M or below M, that the true value will soon be determined (by and enormous experiment) and you will be given a reward if you predicted correctly – you should not feel that one prediction ("above M" or "below M") would be a better choice than the other. If you genuinely think the true value is as likely to be above M as to be below M, then both predictions give you equal chances of receiving the reward.

Experts are often tempted to place M mid-way between L and U (particularly when they feel they have little knowledge about the parameter), but this may not be a good choice. Thinking of M as an estimate of the parameter, there is often asymmetry between the magnitudes of positive and negative errors. For instance, suppose that for some uncertain parameter an expert has specified L = 50 and U = 200. The mid-point between these two bounds is 125 but the expert may feel that M = 100 is a better choice because it says that M may plausibly over- or under-estimate the true value by no more than a factor of 2. The message of examples such as this is that it is necessary to think about M – trying to apply a formula, even such an obvious one as that M should be in the middle of the plausible range, can lead to poor judgements.

Example: Having said that she is confident that the sum of distances from Paris to London and from New York to Chicago would be less than 2500km, the Expert estimates that this sum would be about 1500km. So her estimate of D should be around 7000km less the amount by which the direct route is shorter than flying via London and New York. She sets M = 6500. As a check, she asks herself whether she feels it equally likely that D is less than 6500km or more than 6500km, and confirms that she would have no preference for one prediction over the other. The table now looks like this:

Parameter	**L**	**Q1**	**M**	**Q3**	**U**
D	5500		6500		8000

2.6. The Quartiles

The final step is to specify the upper and lower quartiles. The lower quartile, Q1, is a value between L and M, such that you feel the parameter is equally likely to be below Q1 as it is to be between Q1 and M. Notice that this divides the range from L to M in the same way as M divides the range from L to U, i.e. into two sections that you judge to be equally likely.

You can use the same thought experiment to specify this value: if you had to predict whether the true value of the parameter would be less than Q1 or between Q1 and M, you would not have a preference – they are equally likely to be right.

Note that there is asymmetry here, even clearer and stronger than when specifying M. Your bound L is at the limit of plausibility, and so values just above L are "almost implausible". You will almost always regard such values as relatively unlikely. In particular they will generally be considered much less likely than values close to (but just below) M. If you were to set Q1 mid-way between L and M, then you would obviously prefer to predict that the true value is between Q1 and M than below Q1. Except in unusual circumstances where the above argument does not hold, you should generally place Q1 *above* that mid-point.

The upper quartile, Q3, is specified analogously as a value between M and U such that you feel the parameter is equally likely to be above Q3 as it is to be between M and Q3. And a similar argument suggests than it general you should set Q3 *below* the mid-point between M and U.

The range from Q1 to Q3, referred to as the quartile range, is more informative that that from L to U. Between Q1 and Q3 lie the values that you, the expert, feel are most likely for the parameter.

Experts generally find Q1 and Q3 the most difficult values to give. Two more thought experiments may help. First, you should feel that the parameter is equally likely to lie between Q1 and Q3 as it is to lie outside that range. So whereas the range from L to U covers all plausible values, so the parameter should not lie outside that plausible range, the quartile range is such that you judge the parameter is equally likely to lie outside it as inside. If asked to predict whether the true value would lie inside or outside the quartile range you should not have a clear preference.

Finally, the five numbers L, Q1, M, Q3 and U divide the plausible range into four sections[42], and all four sections should feel equally likely. If you were asked to predict in which of these four sections the true value would lie, you should not feel any preference between them.

[42] This is why Q1 and Q3 are called quartiles. Technically, M is the second quartile, Q2.

Example: The Expert first considers Q1. In terms of the additional distance for D versus the 5500km from London to New York, she thinks it is more likely that this distance is greater than 500km than that it is between 500 and 1000km, so Q1 should be higher than 6000. Her choice is Q1 = 6200. When thinking about Q3, she thinks that when M = 6500 is viewed as an estimate of the true value of D errors of under-estimation may be on the whole larger than errors of over-estimation, so she sets Q3 = 6900. As a check, she asks herself whether she thinks that D is equally likely to be between 6200 and 6900 as to be outside this range, and confirms that she would have no preference for one prediction over the other. Her elicitation for D is now complete:

Parameter	L	Q1	M	Q3	U
D	5500	6200	6500	6900	8000

2.7. What If You Have No Knowledge?

Another concern that is often expressed by experts is a feeling that when information is very weak they cannot give any meaningful judgements. The preceding discussion makes it clear that we are interested in judgements that honestly reflect both the expert's knowledge and his or her uncertainty. When information is weak we must expect this to show in your five numbers. In particular, the plausible range will be wide. But even when information and expertise are very weak it will not be infinitely wide. The upper and lower bounds will not be the theoretical, physical limits of the parameter because in practice it is simply implausible that any parameter will reach both of those limits. The plausible range should not be unnecessarily wide. Nor should it be narrower than is justified by your knowledge. Where evidence is weak we rely on your expertise, based on your years of experience, and ask you simply to make your best personal judgements.

2.8. Order Probabilities

In addition to the five-number elicitations for the individual parameters, you will be asked to make some comparisons between parameters. These take the form of asking for your probability that parameter X is greater than parameter Y. If your estimates (medians) for the two parameters are the same, then you can simply enter "50%" for this probability, because you would have no preference for predicting X to be larger than Y or Y to be larger than X.

If you have given a larger median for X, then you clearly would feel that X is more likely to be larger than Y and should give a probability larger than 50%. The value you choose between 50% and 100% reflects the strength of your conviction that X would be larger than Y, with 100% representing certainty.[43]

Conversely, if you have given X a smaller median than Y then you would feel that X is less likely to be larger than Y and should give a probability less than 50%. The value you choose between 50% and 0% reflects the strength of your conviction that X would *not* be larger than Y, with 0% representing certainty.

Example: An Expert has provided the following two five-number elicitations for the average lifespans, in days, of the adult male and female of a species of moth.

Parameter	L	Q1	M	Q3	U
Male	4	7	9	12	17
Female	3	5	6	8	12

[43] We are using probability here in a sense that may be unfamiliar. People are generally taught that probability measures how often something will occur in a long sequence of repetitions, like coin tosses. That clearly does not apply to parameters that can only ever have one value and so cannot be repeated. We are using here a definition of probability that statisticians often employ, which is simply as a measure of strength of belief.

She has given lower values for the female because the males of most moth species tend to live longer than the females. She is now asked for an order probability: “What is the probability that the male average lifespan is longer than the female?” Although the females are thought to live longer on average in some moth species, this is unusual and she feels that the probability of the male living longer in this species should be high. She judges the order probability in this case to be 90%.

2.9. Summary

In summary, your task is as follows.

- For each parameter, specify the five numbers L, Q1, M, Q3 and U by following carefully the above guidance.
- In particular, always follow the sequence above – first specify L and U, then specify M and finally specify Q1 and Q3.[44]
- Always finish one parameter before moving to the next.
- You will find that the process becomes easier with practice, but please do not start to fill in values mechanically – each value for each parameter should be a genuine expression of your knowledge about that parameter.
- Provide concise explanations of your five-number elicitations. You do not need to explain each number in detail, but should justify particularly your central value M, referring to any specific evidence in support of your judgement. Try also to explain your degree of uncertainty as reflected in the plausible range L to U or the quartile range Q1 to Q3.
- For each requested comparison, provide a carefully considered order probability.
- If you have any concerns or questions of interpretation regarding these instructions, please contact EFSA.
- And thank you very much for participating in this important exercise.

[44] There are good reasons for this, based on research about how experts perform in such tasks.

D.10. APPENDIX 4: FIRST DELPHI ROUND

On the 16th of October 2013 an Excel file was provided via email to the members of the Elicitation Group for the first Delphi round. It was composed by four sheets whose screenshots are available here below.

D.10.1. Sheet 1: "Introduction"

Delphi round 1

Introduction

Dear Participant,

Welcome to the electronic Delphi questionnaire on optimal temperatures for development of Pomacea species!
Many thanks for your willingness to participate in our research.
The purpose of this survey is to elicit your knowledge to determine optimal temperatures for development of Pomacea species.

Before answering the questions, please read carefully the SummaryDefinitions and SummaryTraining sheets. These will remind you the background information and guidance you need for performing the exercise.

In the following box, please find an example of a question you will have to answer in the Delphi_Round _No1 sheet.
You can enter your judgements into the green fields. Some help functions in the background will guide you through the procedure.
Please enter your answer in the cell that is dark green colored. (The cursor in EXCEL will NOT automatically move to the next field.)
If you encounter some problems, please "Cancel" your entry and try again. You can clear all entries to one question by pushing the "Clear all" button.

Step Parameter	2nd Lower bound	5th Lower quartile	3rd Median	4th Upper quartile	**1st** **Upper bound**
Question:	(L)	(Q1)	(M)	(Q3)	(U)
What is the **optimal (...) temperature** for (...) **development** of ***Pomacea (...)*** species (in °C)? In addition, please provide below the rational for your judgements.					
Clear all — Rationale:					

Please note that your individual answers and comments, the results of the survey, as well as the feedback we will provide in subsequent surveys will be anonymized.

If you have any difficulties with this questionnaire, or if you would like to have some extra information about the survey, please contact the survey team at the following email address: **sas.expert-elicitation@efsa.europa.eu**

mailto:sas.expert-elicitation@efsa.europa.eu

Kind regards on behalf of the Elicitation Group,

Meike Wentholt (Elicitor), Sara Tramontini (EFSA-Plant Health Unit)

c/o European Food Safety Authority (EFSA)
Via Carlo Magno 1/A, IT-43126 Parma, Italy

Please proceed with the next EXCEL sheet: "Summary Definitions".

D.10.2. Sheet 2: "SummaryDefinitions"

Short summary of definitions

Please remind the current status of the taxonomy of *Pomacea canaliculata* and *Pomacea maculata:*

• Hayes et al. (2012) synonymize ***Pomacea insularum*** and ***Pomacea gigas*** (Ampullaria insularum d'Orbigny, 1835 and Ampullaria gigas Spix, 1827) with ***Pomacea maculata*** Perry, 1810 and clearly discriminate the latter from ***Pomacea canaliculata***. They designate neotypes for *P. maculata, P. canaliculata* and *A. gigas*, as well as a lectotype for *A. insularum.*

We use the following definition for the life stages:

• The **Egg stage** starts when eggs are produced.

• The **Juvenile stage** starts when eggs are hatched.

• The **Adult stage** starts at 1st reproductive event.

In the model a threshold for cummulation of degree-days is used:

• The threshold T° for cumulating of degrees corresponds to the threshold for the **total sum to complete the development from one step of the life cycle to the next one**.

In the table below you can find the **definitions** to be considered by you for the **temperature estimates:**
In this first Delphi round we only ask you about the "optimal temperatures" and your uncertainties of these values.

Stage	Optimal T° for development	Min T°	Max T°	Threshold T° for cumulating of degrees
Egg	**The temperature at which embryonic development time from egg spawn until hatching is the shortest**	The lowest temperature for embryonic development	The highest temperature at which embryonic development can occur	Equals the minimum T°, used to set the threshold when calculating day-degree sum for complete embryonic development from egg spawn until egg hatch.
Juvenile	**The temperature at which juvenile development time from egg hatching till the first reproductive event is shortest**	The lowest temperature for juvenile development	The highest temperature at which juvenile development can occur	Equals the minimum T°, used to set the threshold when calculating day-degree sum to complete the juvenile stage from egg hatch till the first reproductive event
Adult	**The temperature at which the development rate is the highest in terms of growth starting from the first reproduction event**	The lowest temperature for adult development	The highest temperature for adult development/activity	

For further information please consult the Background Information, if needed, and proceed with the next EXCEL sheet: "Summary Training".

D.10.3. Sheet 3: "SummaryTraining"

Short summary training session

For each parameter, we will ask you to state five numbers. First, you will give an upper bound (U) and a lower bound (L). Next you will give an estimate known as the median (M). Finally, you will give two numbers called the upper and lower quartiles (Q3 and Q1) which provide a range around the median that you judge to be the most likely values for the parameter. In addition, you will be asked to provide a rationale for your judgements.

Step Parameter	2nd Lower bound	5th Lower quartile	3rd Median	4th Upper quartile	**1st** **Upper bound**
Question:	(L)	(Q1)	(M)	(Q3)	(U)
What is the **optimal (...) temperature** for (...) **development** of ***Pomacea (...)*** species (in °C)? In addition, please provide below the rational for your judgements.					
Rationale:					

Clear all

- For each parameter, specify the five numbers L, Q1, M, Q3 and U by following carefully the elicitation guidance.
- In particular, always follow the sequence shown by the numbering of the columns – first specify U and L, then specify M and finally specify Q3 and Q1.
  (The first or next field is highlighted in dark green. Please move the curser into this field to enter your answer.)
- Always finish one parameter before moving to the next.
- You will find that the process becomes easier with practice, but please do not start to fill in values mechanically – each value for each parameter should be a genuine expression of your knowledge about that parameter.
- Provide concise explanations of your five-number elicitations. You do not need to explain each number in detail, but should justify particularly your central value M, referring to any specific evidence in support of your judgement. Try also to explain your degree of uncertainty as reflected in the plausible range L to U or the quartile range Q1 to Q3.

In addition to the five-number elicitations for the individual parameters, we would like you to make some comparisons between parameters (order probabilities). In that context you will be requested to provide an estimation that parameter X is greater than parameter Y. These take the form of asking for your probability that parameter X is greater than parameter Y.

Question:

What is the probability that the **optimal (...) temperature for development of *Pomacea (...)* (T_opt_xxx)** is greater than the **optimal (...) temperature for eggs development of *Pomacea (...)* (T_opt_eggs)**?

Probability for "T_opt_xxx greater than T_opt_eggs" (in %)		Probability for "T_opt_xxx less than T_opt_eggs"	

• If your estimates (medians) for the two parameters are the same, then you can simply enter "50%" for this probability.

• If you have given a larger median for X, then you clearly would feel that X is more likely to be larger than Y and should give a probability larger than 50%.

• If you have given X a smaller median than Y then you would feel that X is less likely to be larger than Y and should give a probability less than 50%.

Follow carefully the explanation provided in the elicitation guidance document regarding these probability estimations.

Thank you, now progress onto the questionnaire by going to the next EXCEL sheet: "Delphi_Round_No1".

D.10.4. Sheet 4: "Delphi_Round_No1"

Please answer the following questions for the ***P. canaliculata* species** (Please fill the green cells)

	Step Parameter	2nd Lower bound	5th Lower quartile	3rd Median	4th Upper quartile	1st Upper bound
1. Question:		(L)	(Q1)	(M)	(Q3)	(U)
What is the **optimal air temperature** for **eggs development** of ***P. canaliculata*** species (in °C)? In addition, please provide below the rational for your judgements.						
Clear all	Rationale:					

Step Parameter	2nd Lower bound	5th Lower quartile	3rd Median	4th Upper quartile	**1st** **Upper bound**
2. Question:	(L)	(Q1)	(M)	(Q3)	(U)
What is the **optimal water temperature** for **juvenile growth** of ***P. canaliculata*** species (in °C)? In addition, please provide below the rational for your judgements.					
Clear all — Rationale:					

Step Parameter	2nd Lower bound	5th Lower quartile	3rd Median	4th Upper quartile	**1st** **Upper bound**
3. Question:	(L)	(Q1)	(M)	(Q3)	(U)
What is the **optimal water temperature** for **adults activity (feeding, copulation etc.)** of ***P. canaliculata*** species (in °C)? In addition, please provide below the rational for your judgements.					
Clear all — Rationale:					

Please answer the following questions for the **probabilities of the order** (Please fill the green cells)

In addition to the five-number elicitations for the individual parameters, we would like you to make some comparisons between parameters. These take the form of asking for your probability that parameter X is greater than parameter Y.

4. Question:

What is the probability that the **optimal water temperature for activity (feeding, copulation etc.) of *P. canaliculata* adults (T_opt_adults)** **is greater than** the **optimal air temperature for eggs development of *P. canaliculata* (T_opt_eggs)**?

Probability for "T_opt_adults greater than T_opt_eggs" (in %)		Probability for "T_opt_adults less than T_opt_eggs"	

5. Question:

What is the probability that the **optimal water temperature for growth of *P. canaliculata* juveniles (T_opt_juveniles)** **is greater than** the **optimal air temperature for eggs development of *P. canaliculata* (T_opt_eggs)**?

Probability for "T_opt_juveniles greater than T_opt_eggs" (in %)		Probability for "T_opt_juveniles less than T_opt_eggs"	

Please answer the following question for the **differences between the two species (*P. canaliculata, P. insularum(maculata)*)**

6. Question:

Do you think there are differences between the two species (*P. canaliculata, P. insularum(maculata)*)**?**

Please provide your answer and rationale:

Please answer the following questions for the ***P. insularum(maculata)*****species** (Please fill the green cells)

Step Parameter	2nd Lower bound	5th Lower quartile	3rd Median	4th Upper quartile	**1st** **Upper bound**
7. Question:	(L)	(Q1)	(M)	(Q3)	(U)
What is the **optimal air temperature** for **eggs development** of ***P. insularum*** species (in °C)? In addition, please provide below the rational for your judgements.					

Clear all

Rationale:

Step Parameter	2nd Lower bound	5th Lower quartile	3rd Median	4th Upper quartile	1st Upper bound
8. Question:	(L)	(Q1)	(M)	(Q3)	(U)

What is the **optimal water temperature** for **juvenile growth** of ***P. insularum*** species (in °C)?

In addition, please provide below the rational for your judgements.

Clear all

Rationale:

Step Parameter	2nd Lower bound	5th Lower quartile	3rd Median	4th Upper quartile	1st Upper bound
9. Question:	(L)	(Q1)	(M)	(Q3)	(U)

What is the **optimal water temperature** for **adults activity (feeding, copulation etc.)** of ***P. insularum*** species (in °C)?

In addition, please provide below the rational for your judgements.

Clear all

Rationale:

Thank you for your participation!
Please save this document and send it latest **until Thursday, 17th October 2013,** to the survey team:

sas.expert-elicitation@efsa.europa.eu

mailto:sas.expert-elicitation@efsa.europa.eu

All responses will be aggregated and analysed, at which time a new questionnaire will be developed.
You will receive an invitation to participate in about 3 to 4 weeks.

Print-View

D.11. APPENDIX 5: SECOND DELPHI ROUND

On the 5th of March 2014 another Excel file was provided via email to the members of the Elicitation Group for the second Delphi round, with the following accompanying message:

Dear *Pomacea* expert,
quite a long time passed from our last contact, but I hope you remember our exercise on expert knowledge elicitation (EKE) on temperature ranges for Pomacea sp.

The results of the first round of EKE that you conducted with us were of great help to improve the instrument of Delphi to elicit probability distribution via written procedure. Thanks to these advancements, we could conduct in EFSA the first workshop on EKE in support to risk assessment (end of January 2014). It was a great success and we could present our first experience done with you on Pomacea. The attention and feedbacks from participants were very positive and we decided to go ahead with the full EKE procedure on Pomacea case study in order to make further progress on the methodology.

In the meanwhile, an EFSA opinion on the potential for establishment of the apple snail in the EU has been published:http://www.efsa.europa.eu/it/efsajournal/pub/3487.htm

and a second on environmental risk assessment for Pomacea sp. is in preparation.

However, for our group the priority is now the validation of EKE methodology for pest risk assessment and the exercise that you have been invited to conduct represents a crucial opportunity to validate the working protocol of Delphi questionnaires via written procedure, for future EKEs conducted by EFSA.

Some more details on this second Excel file in which you will find the questionnaire:

In sheet “Delphi_T_opt_rounds 1 and 2” before each question we included summary tables with the answers received from you. They are anonymous with comments from the elicitor, in the last column, whenever relevant and should help you in adjusting your judgements in this new step.

In sheet “Delphi_T_min_round” and “Delphi_T_MAX_round 2”, you will find the same questions but referring to minimum and maximum temperatures respectively. The first impression will be therefore that the questionnaire requires more time to be filled in. However, you have to consider that with your first attempt plus the provided elicitor’s feedbacks, you have already gained a certain experience on expert knowledge elicitation. Furthermore, the two questions on order probabilities (originally 4 and 5) have been removed. We are confident that all these aspects together will simplify and accelerate a lot the process. Because of that, we would be grateful to receive also from you an estimate of the amount of time you needed for completing the first round questionnaire and the amount needed for the second: as you can foreseen, this is a very relevant methodological feedback.

The inclusion of a rationale for each answer you provide is crucial for the elicitor to understand what is behind (data, personal knowledge, experience, uncertainty) the numbers you gave. Therefore we kindly ask you to add as much supporting text as possible to each box for rationale.

The fact that you did not participate to the first step does not represent a problem: you will use, as the others, the included replies and feedbacks as a guide and support to your own answers.

Another version of the questionnaire has been provided for those experts who prefer to work with Fahrenheit instead of Celsius.

The deadline for this second step is Wednesday, 12 March 2014. And a third step is planned in a short time, depending on your availability to reply to the questionnaire. This would allow the elicitors to include this exercise in the guidelines on expert knowledge elicitation at EFSA currently under preparation.

I am looking forward for your priceless contribution

With kindest regards on behalf of the whole group

Sara

D.11.1. SUMMARY TABLES

Summary tables with the answers received from the experts in the previous round with comments from the elicitor were included in the Excel file and provided here below in order to give a concrete example of the analysis conducted on the replies.

Step Parameter 1. Question:	Expert	2nd Lower bound (L)	5th Lower quartile (Q1)	3rd Median (M)	4th Upper quartile (Q3)	1st Upper bound (U)	Provided rationale	Comments by elicitor
What is the **optimal air temperature** for **eggs development** of ***P. canaliculata*** species (in °C)?	1	**25.00**	**25.00**	**27.00**	**30.00**	**33.00**	50	The elicitor interpreted the rationale provided by expert 1 as his/her % degree of confidence for that specific rating.
	2	**14.00**	**20.00**	**28.00**	**30.00**	**38.00**	From litterature and my own experience, egg development starts at around 14 C and ends below 40 C. Optimum seems to be around 30 C but also depend on humidity and other factors.	the expert provided the range of temperatures tolerated, not the optimal.
	3	**32.00**	**33.00**	**34.00**	**35.00**	**38.00**	We studied development duration of the eggs of P. canaliculata at fluctuating temperatures in the lab and also compared our results with those of literature. Developmental duration decreases asymptotically with temperature at least up to 32ºC but probably harmful effects appear at constant temperatures of 38 or more.	If harmful effects appear at Temperatures ≥ 38 °C it is difficult to consider 38 °C still in the range of the optimal temperatures.

Step Parameter	Expert	2nd Lower bound	5th Lower quartile	3rd Median	4th Upper quartile	1st Upper bound	Provided rationale	Comments by elicitor
2. Question:		(L)	(Q1)	(M)	(Q3)	(U)		
What is the **optimal water temperature** for **juvenile growth** of ***P. canaliculata*** species (in °C)?	1	**25.00**	**25.00**	**27.00**	**30.00**	**31.00**	80	
	2	**15.00**	**20.00**	**28.00**	**30.00**	**35.00**	From litterature and my own experience, juvenile is slow below 15 degrees and ends around 35 C water temperature where snail mortality increases	the expert provided the range of temperatures tolerated, not the optimal.
	3	**30.00**	**31.00**	**32.00**	**33.00**	**35.00**	We studied recently the growth rate and survival of P. canaliculata between 15 and 35ºC. The maximun growth rate was attained at 35ºC but at that temperature the snails will not reproduce (at least the females will not lays eggs),	the fact that at 35 °C the snails do not reproduce does not seem a relevant information for juvenile stages.

Step / Parameter / 3. Question:	Expert	2nd Lower bound (L)	5th Lower quartile (Q1)	3rd Median (M)	4th Upper quartile (Q3)	1st Upper bound (U)	Provided rationale	Comments by elicitor
What is the **optimal water temperature** for **adults** activity (**feeding, copulation etc.**) of ***P. canaliculata*** species (in °C)?	1	**25.00**	**25.00**	**27.00**	**30.00**	**30.00**	80	Lower quartile equal to the lower bound and upper quartile equal to upper bound: it is not possible for the optimum temperature for adults to be between the lower bound and the lower quartile, and yet there should be a 25% probability in this region. Likewise there should be a 25% probability to be above the upper quartile and below the upper bound, but both are 30.
	2	**15.00**	**20.00**	**28.00**	**30.00**	**35.00**	From litterature and my own experience, adult activity is slow below 15 degrees and ends around 35 C water temperature where snail mortality increases	the expert provided the range of temperatures tolerated, not the optimal.
	3	**23.00**	**26.00**	**27.00**	**28.00**	**30.00**	In many aspects adults perform equally well at 25 and 30 but at the higher temperature they invest more time crawling to go to the surface to breath, they feed less and probably gain less weight (from our own results)	The mentioned range of equal performance (25-30) is a bit smaller than that proposed in the table (23-30).is there any reason?

Step Parameter 5. Question:	Expert	2nd Lower bound (L)	5th Lower quartile (Q1)	3rd Median (M)	4th Upper quartile (Q3)	1st Upper bound (U)	Provided rationale	Comments by elicitor
What is the **optimal air temperature** for **eggs developmen t** of ***P. insularum*** species (in °C)?	1	**25.00**	**27.00**	**28.00**	**29.00**	**33.00**	40	
	2	**14.00**	**20.00**	**28.00**	**30.00**	**38.00**	From litterature and communication with managers in Spain, egg development starts at around 14 C and ends below 40 C. Optimum seems to be around 30 C but also depend on humidity and other factors.	see previous
	3	**32.00**	**34.00**	**35.00**	**37.00**	**38.00**	based on comparison with P. canaliculata, whose distribution extends farther to higher latitudes.	

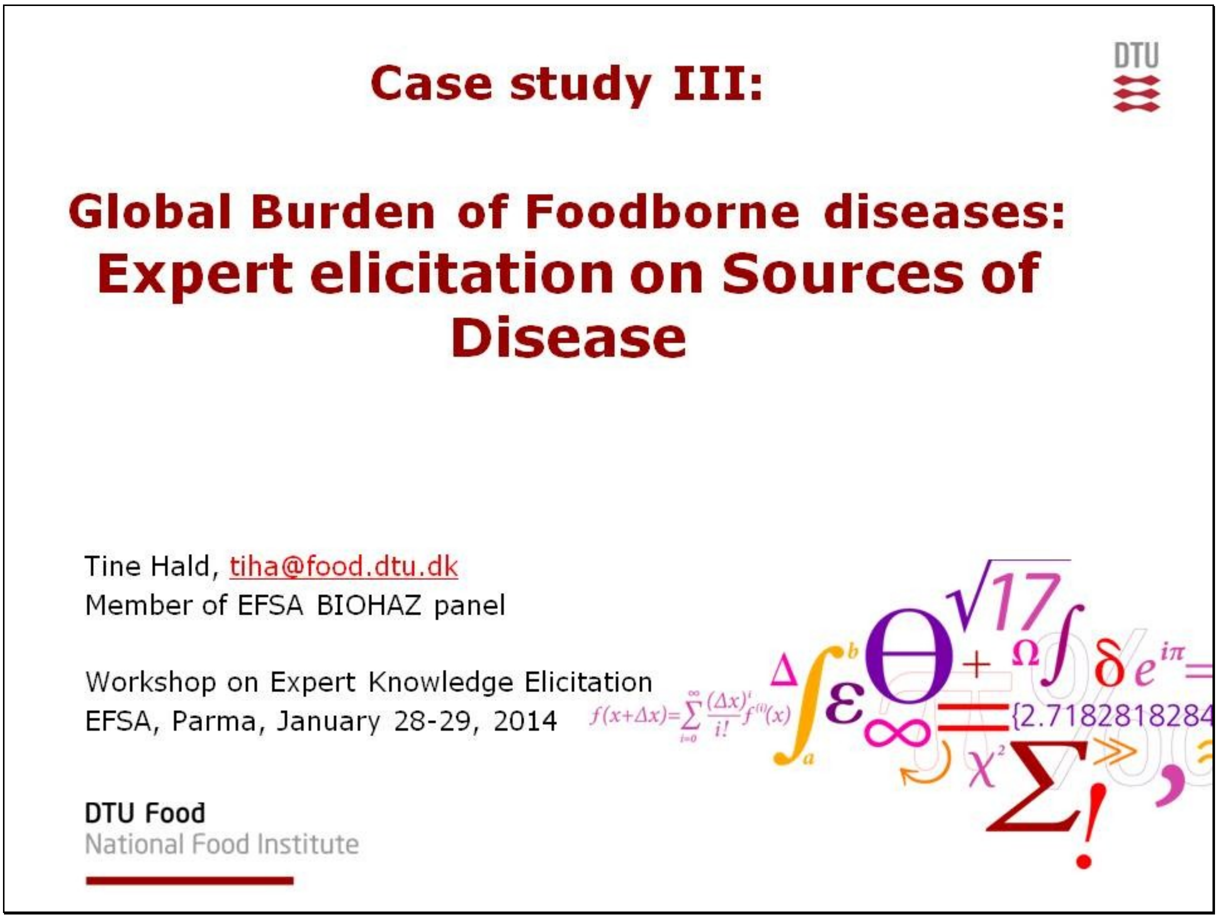
Step Parameter 6. Question:	Expert	2nd Lower bound (L)	5th Lower quartile (Q1)	3rd Median (M)	4th Upper quartile (Q3)	1st Upper bound (U)	Provided rationale	Comments by elicitor
What is the **optimal water temperature** for **juvenile growth** of ***P. insularum*** species (in °C)?	1	**25.00**	**27.00**	**28.00**	**29.00**	**33.00**	40	
	2	**15.00**	**20.00**	**28.00**	**30.00**	**35.00**	From litterature and my own experience, juvenile is slow below 15 degrees and ends around 35 C water temperature where snail mortality increases	see previous
	3	**30.00**	**32.00**	**33.00**	**35.00**	**36.00**	based on comparison with P. canaliculata, whose distribution extends farther to higher latitudes.	

Step Parameter 7. Question:	Expert	2nd Lower bound (L)	5th Lower quartile (Q1)	3rd Median (M)	4th Upper quartile (Q3)	1st Upper bound (U)	Provided rationale	Comments by elicitor
What is the **optimal water temperature** for **adults activity (feeding, copulation etc.)** of ***P. insularum*** species (in °C)?	1	**25.00**	**26.00**	**28.00**	**29.00**	**31.00**	40	
	2	**15.00**	**20.00**	**28.00**	**30.00**	**35.00**	From litterature and my own experience, adult activity is slow below 15 degrees and ends around 35 C water temperature where snail mortality increases	see previous
	3	**23.00**	**27.00**	**28.00**	**29.00**	**31.00**	based on comparison with P. canaliculata, whose distribution extends farther to higher latitudes.	

E. CASE STUDY IN BIOLOGICAL HAZARDS

To cover a real application of expert knowledge elicitation in the field of biological hazards we asked Tine Hald (DTU National Food Institute of Denmark, Lyngby, Denmark) to present the expert elicitation developed for source attribution of global burden of foodborne diseases of the WHO "Foodborne Disease Burden Epidemiology Reference Group (FERG)" on the project workshop held on 28th and 29th January 2014. We got the kind permission to include her slides as additional appendix.

E.1. GLOBAL BURDEN OF FOODBORNE DISEASES: EXPERT ELICITATION ON SOURCES OF DISEASE (BY TINE HALD)



The Foodborne Disease Burden Epidemiology Reference Group (FERG)

- WHO expert group formed in 2007 – hoping to finalise work this year!

Objective

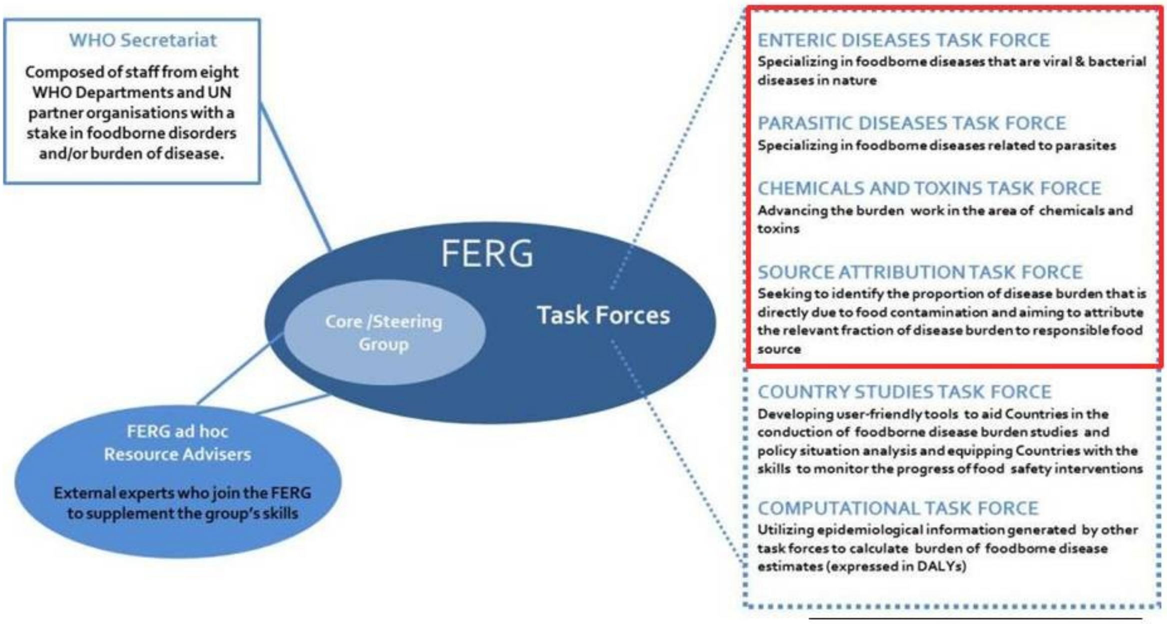
- To estimate the global burden of foodborne diseases:
  - Diseases commonly transmitted through food
  - All causes – pathogens and chemicals
  - Acute and chronic diseases
  - Long-term complications
  - Morbidity, disability and mortality

2

Why hasn't it been done before?

- Many different hazards
- One hazard often has many exposure routes
- Contamination may arise at many different points in the exposure pathway
- Delineating modes of transmission extremely difficult

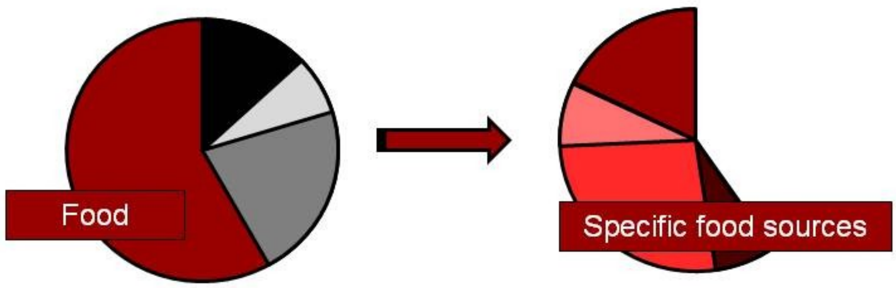
The Foodborne Disease Burden Epidemiology Reference Group (FERG)



4

The primary objectives of the FERG Source Attribution Task Force

- Determine for each hazard (or specific groups of hazards) the proportion of the disease burden that is attributable to food
- Identify – if possible quantify - the responsible reservoirs and/or food commodities leading to illness



5

ToR of the Source Attribution TF

1. To define "foodborne disease" and **"source attribution"** and agree on levels of food categorization and **point of attribution**;
2. To assess all currently available methods attributing causative agents of foodborne diseases to food sources and food commodities;
3. To propose suitable source attribution methods for the causative agents examined by FERG and/or develop new methods;

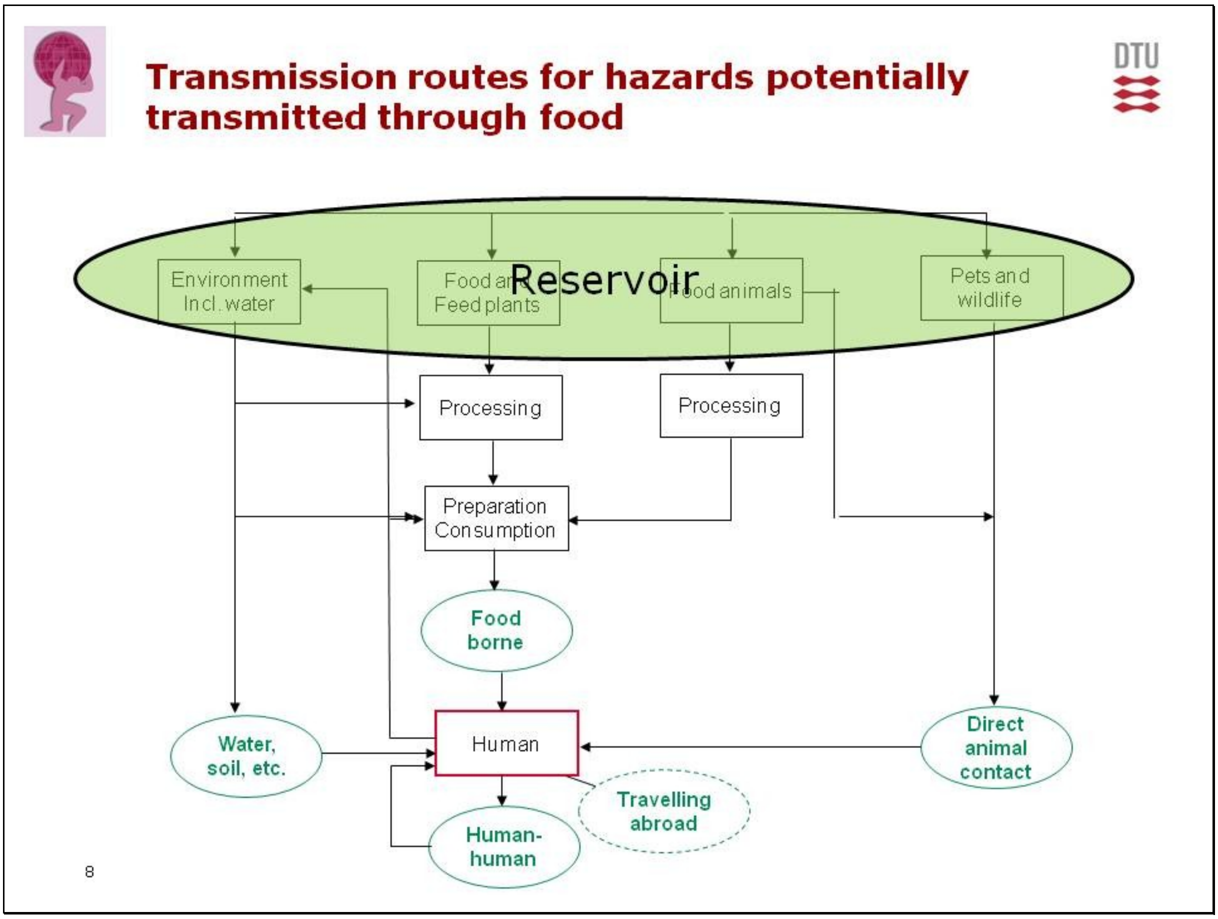
6

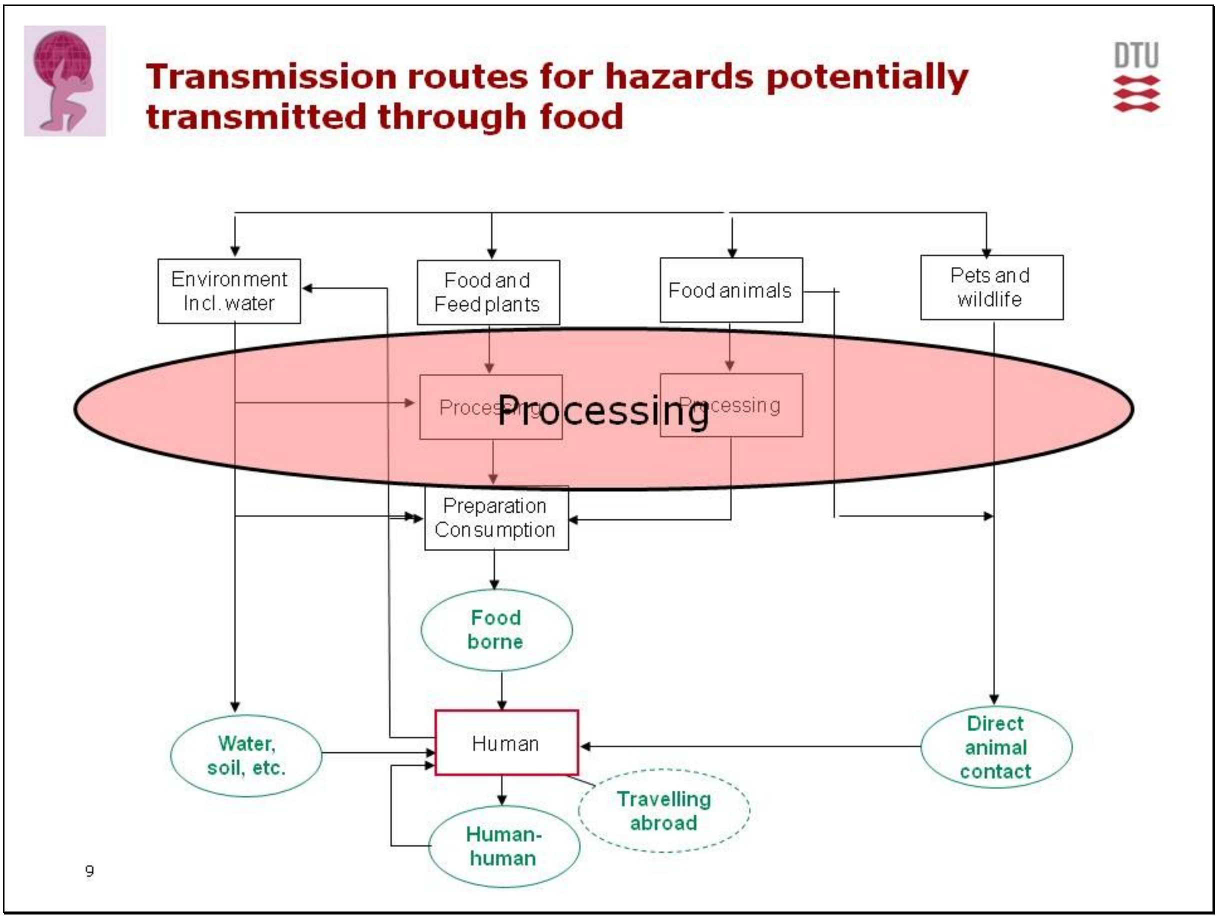
Definition of Source Attribution

- **Source attribution (SA)** is the partitioning of the human disease burden of one or more foodborne diseases to specific sources, where the term 'source' includes animal reservoirs (Pires et al., 2009)

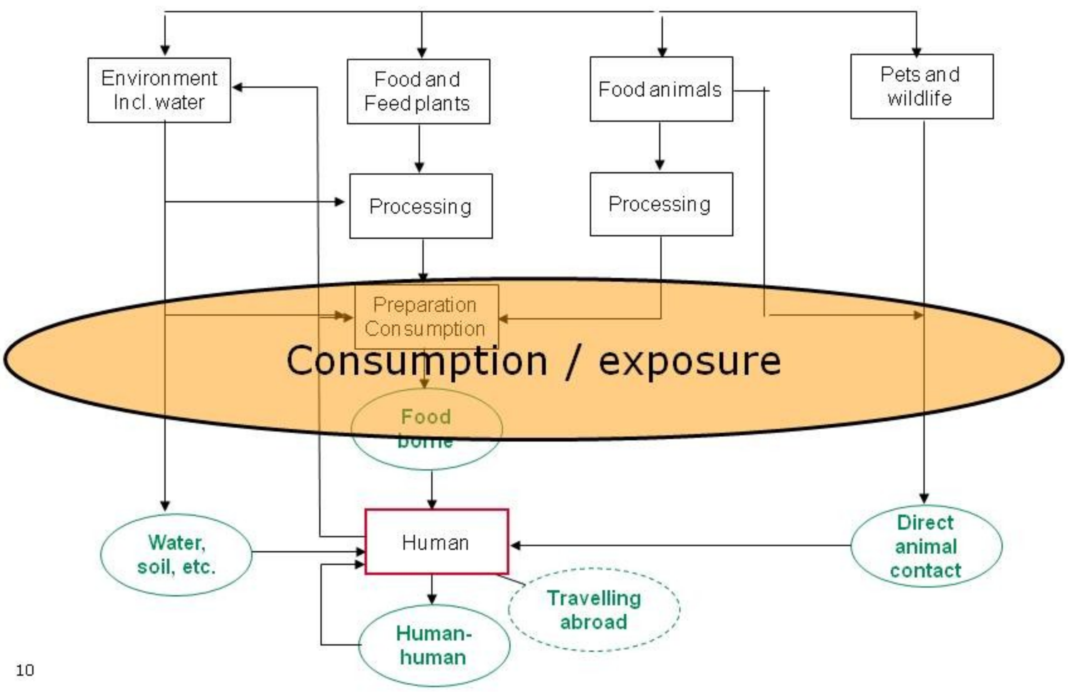
- **Point of attribution:** SA can be conducted at various points along the food distribution chain, including at the reservoir level and at the point of consumption (point-of-exposure)
  - The burden of illness caused by a particular hazard attributed to a specific transmission route may vary, depending on the point-of-attribution chosen

7


Transmission routes for hazards potentially transmitted through food
DTU
Reservoir
Environment Incl. water
Food and Feed plants
Food animals
Pets and wildlife
Processing
Processing
Preparation Consumption
Food borne
Water, soil, etc.
Human
Direct animal contact
Travelling abroad
Human-human
8


Transmission routes for hazards potentially transmitted through food
DTU
Environment Incl. water
Food and Feed plants
Food animals
Pets and wildlife
Processing
Processing
Processing
Preparation Consumption
Food borne
Water, soil, etc.
Human
Direct animal contact
Travelling abroad
Human-human
9

Transmission routes for hazards potentially transmitted through food



10

Commissioning an Expert Elicitation

- For estimating the **overall proportion foodborne,** structured expert elicitation was identified to be the only option for the vast majority (all) of hazards, since
  - Other methods estimate the proportion attributed to specific foods or reservoirs, but not the total
  - The inevitable "unknowns" makes simple summation impossible
- For estimating the **contribution from specific food sources** (or sub-pathways) within the pathway **food**, different methods including outbreak data and expert elicitation will be applied
- Expert elicitation is applied to all hazards that are not (almost) 100% originating from a single food source/reservoir
  - Hazards included were prioritised by the thematic task forces

11

Expert Elicitation on source attribution

DTU

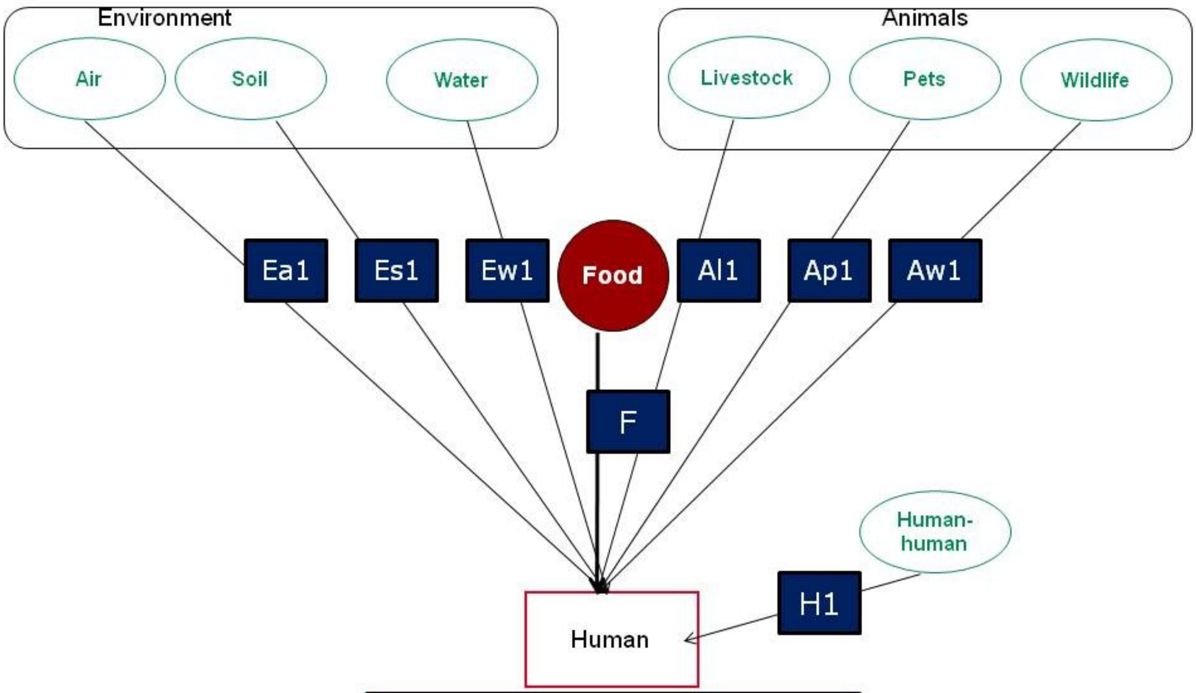
Main Objectives

- For each specified hazard, estimate the proportion at the point of exposure of disease that is transmitted by different major pathways, for each WHO subregion.

12

Transmission routes included in the expert elicitation - point-of-exposure

DTU



$\Sigma E1 + \Sigma A1 + H1 + F \sim 100\%$

13

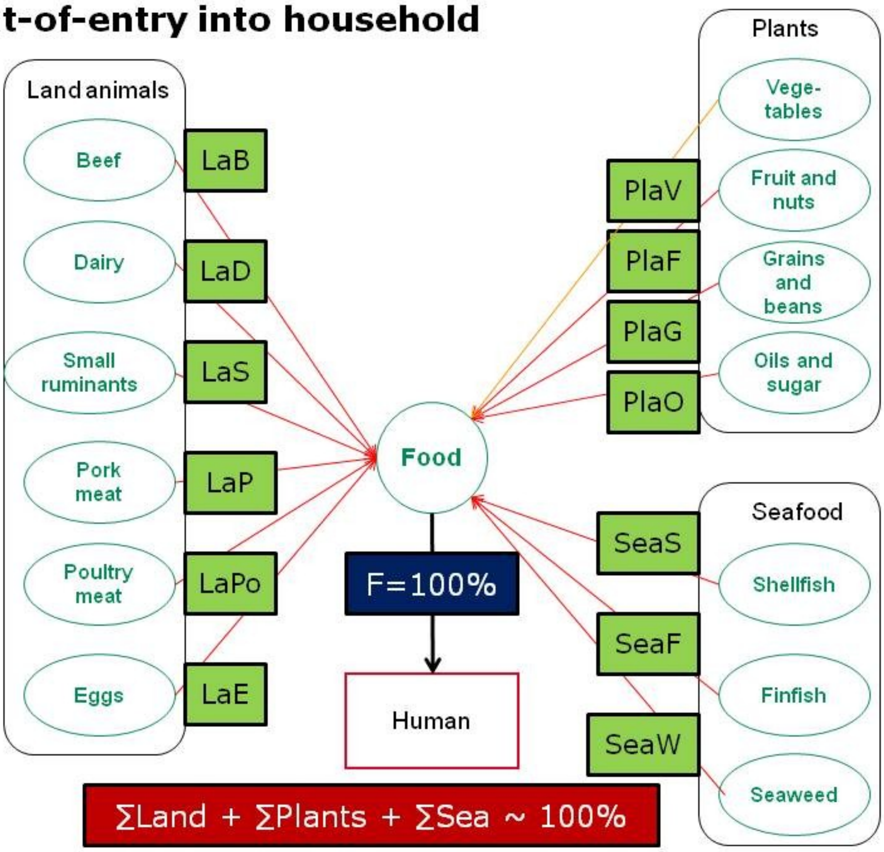
Expert Elicitation on source attribution

Main Objectives

- For each specified hazard, estimate the proportion at the point of exposure of disease that is transmitted by different major pathways, for each WHO subregion.
- Within the pathway food, estimate the proportion of disease that is transmitted by different food categories at point-of-entry into household or food preparation setting, for each WHO subregion.

14

Food sub-pathways included in the expert elicitation - point-of-entry into household

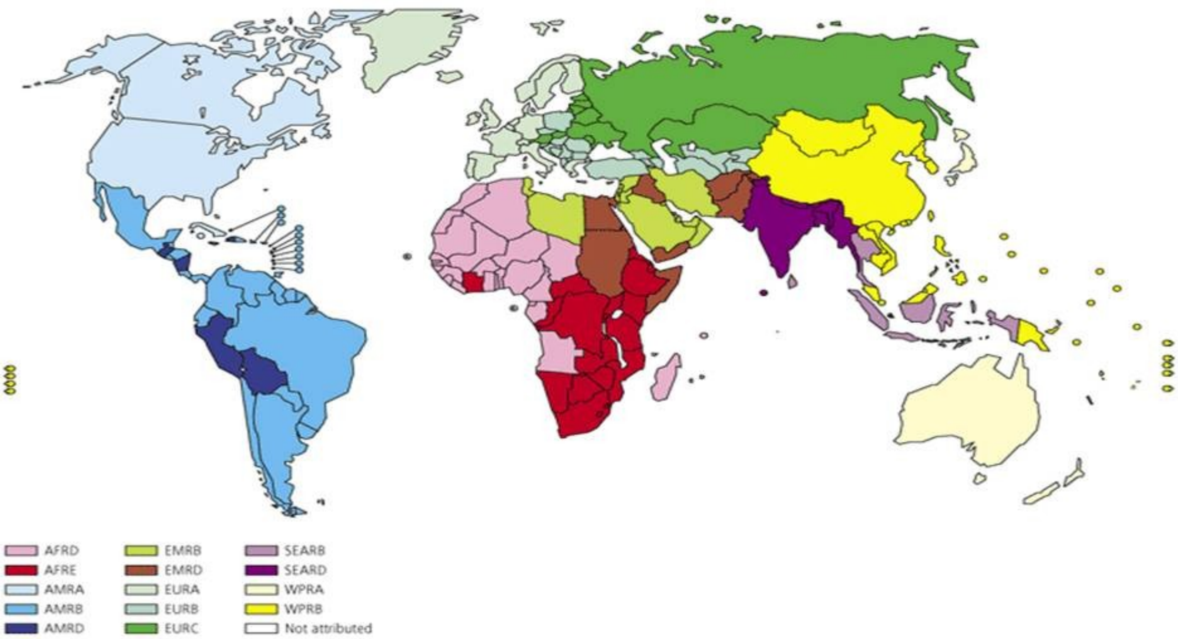


∑Land + ∑Plants + ∑Sea ~ 100%

15

Experts asked to provide estimates for all WHO subregions

Fig. 1. **WHO subregions and mortality strata**



Methodology

- Choice of method
  - Gathering of experts not possible
  - Face-to-face interviews not possible
  - Need for translation (English, French, Russian, Spanish, Chinese)
  - Multidisciplinary panels
  - Performance measurement

→ Cooke's method

17

Methodology

- Set up
  - Definition of panels
  - Identification and enrolment of experts
  - Formulation of calibration questions
  - Formulation of target questions
  - Recruitment and training of facilitators
  - Interviewing experts
    - Calibration questions
    - Training for target questions
  - Experts to return spreadsheets with target question after 2-4 weeks
  - Data cleaning and analysis

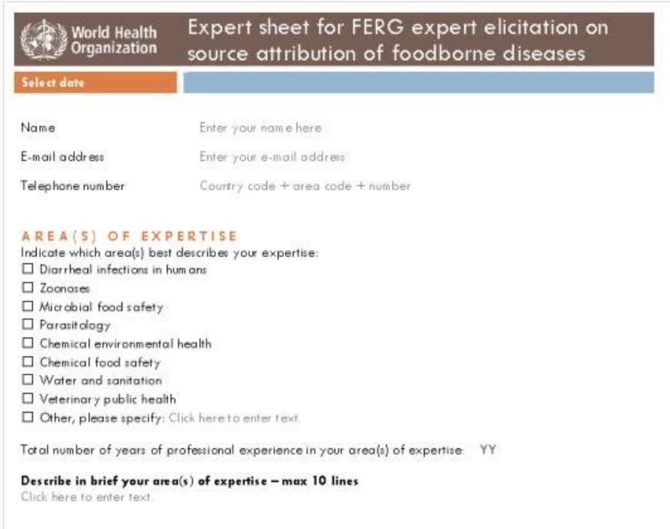
18

Overview of panels and hazards

Panels		Hazards		Pathway	Food
		#			
Heavy metals	Global	3	Lead, Cadmium, inorganic arsenic	Yes	Pb blood, Cd urine conc
Protozoa	Global	3	Cryptosporidium, Giardia, E. histolytica	Yes	Yes
Echinococci	Global	2	E. granulosus, E. multilocularis	Yes	Yes
Toxoplasma	Global	1	T. gondii	Yes	Yes
Ascaris	Global	1	A. lumbricoides	Yes	Yes
Brucella	Global	1	Brucella spp.	Yes	Yes
Hepatitis A	Global	1	Hapatitis A virus - HAV	Yes	No
Enteric pathogens	Regional	3	Salmonella, Campylobacter, STEC	Yes	Yes
		6	Typhoid Salmonella, Vibrio cholerae, Shigella, EPEC, ETEC, Norovirus	Yes	No
Total	**8**	**21**			

Expert panels

- Identification and selection by WHO/FERG
  - Peer identification through network process
  - Self nomination with credentials
  - Invited experts asked to nominate others (snowballing)
  - Review of credentials (CV, expert sheet, DoI) by FERG and WHO


World Health Organization

Expert sheet for FERG expert elicitation on source attribution of foodborne diseases

Select date

Name	Enter your name here
E-mail address	Enter your e-mail address
Telephone number	Country code + area code + number

AREA(S) OF EXPERTISE

Indicate which area(s) best describes your expertise:

- ☐ Diarrheal infections in humans
- ☐ Zoonoses
- ☐ Microbial food safety
- ☐ Parasitology
- ☐ Chemical environmental health
- ☐ Chemical food safety
- ☐ Water and sanitation
- ☐ Veterinary public health
- ☐ Other, please specify: Click here to enter text.

Total number of years of professional experience in your area(s) of expertise: YY

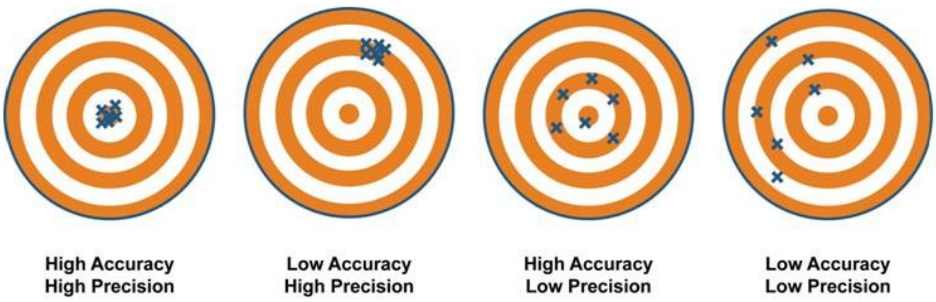
Describe in brief your area(s) of expertise – max 10 lines

Click here to enter text.

20

Calibration questions

- Calibration questions (or seed variables) should resemble the target questions as much as possible.....
- Used to measure the expert's performance
  - Accuracy – good calibration
  - Precision – high information



21

Calibration questions

- Categories of calibration questions for biological hazards
  - Food supply
    - E.g. Among all WHO sub-regions, in 2010 what was the proportion of regional vegetable supply (tonnes) that was imported rather than produced domestically in the WHO sub-region with the highest such percentage?
  - Under 5 years mortality rate
    - E.g. Based on WHO's estimates, think of the country in the WHO African Region that had the largest percentage point decrease from 2000 to 2010 in all-cause under-5 mortality that was due to diarrhea. What was that percentage point *decrease?*
  - Improved water and sanitation
  - Disease surveillance
    - E.g. What will be the rate per 100,000 population of laboratory confirmed human cases of campylobacteriosis in 2012 in all EU member states as reported in EFSA's annual report?
  - Systematic review

22

Calibration questions

- Categories of calibration questions for chemical hazards
  - Lead
    - What did the UNEP Final Review of Scientific Information on Lead report in 2010 as the mean blood lead level for children in Nigeria? Please express your answer as positive micrograms per deciliter (μg/dL)
  - Cadmium
  - Inorganic arsenic
  - Dietary patterns and food supply
    - E.g. Based on this FAO Food Balance Sheet data, in 2009 what was the mean percentage of rice in the national food supply available for human consumption for countries in the WHO South East Asia Region?

	______	______	______
	low	*best*	*high*
Percentile	(5^{th})	(50^{th})	(95^{th})

23

Target questions - pathways

Table 1: Pathogen X Total Exposure

Percent of All Human Cases in a Typical Year

Point of Exposure*	lower credible value (5th percentile)	central value (50th percentile)	upper credible value (95th percentile)
Food	10%	20%	35%
Animal Contact (Domestic and Wild)	0%	5%	12%
Human to Human Contact	5%	10%	28%
Water	50%	65%	80%
Other	0%	2%	5%
		102%	

*Think of the source that was the direct cause of human exposure. We at not asking about how the source was contaminated in this table.

1. To reduce your work, the WHO FERG has identified exposures that its members think are most relevant. If you think points of exposure or foods that are not listed are important in this subregion, please use the category "other foods" or "other".

24

Reducing the pathway-hazard matrices

HAZARD LIST Enterics	PATHWAY FOOD	ANIMAL CONTACT	HUMAN-TO-HUMAN	ENVIRONMENT WATER	SOIL	AIR
Salmonella (non-typhoidal) spp.						
Campylobacter spp.						
Shiga-toxin producing E. coli (STEC)						
Shigella spp.						
Vibrio cholerae						
Brucella spp.						
Norovirus						
Entero-toxigenic E. coli (ETEC)						
Entero-pathogenic E. coli (EPEC)						
Salmonella (typhoid) spp.						
Hepatitis A virus						
Cryptosporidium spp.						
Entamoeba histolytica						
Giardia intestinalis						

25

Target questions – food categories

DTU

Table 2: Pathogen X Foodborne Exposure

Percent of Foodborne Cases in a Typical Ye

Food Consumed**	lower credible value (5th percentile)	central value (50th percentile)	upper credible value (95th percentile)
Beef	10%	15%	25%
Goat, lamb and other small ruminants' meat	3%	5%	18%
Dairy (milk and milk products)	0%	5%	8%
Pork	25%	35%	50%
Poultry Meat	9%	10%	12%
Vegetables (excluding dried legumes)	4%	15%	20%
Fruits	1%	5%	10%
Nuts	0%	1%	2%
Other foods	1%	5%	10%

96%

**Attribute cases of illness to the foods that were already contaminated when they entered the home kitchen or other place of final food preparation. Do not consider cross-contamination in the home kitchen or other place of final food preparation.

26

Materials produced

DTU

- Hazard-pathway matrices
- Hazard-food category matrices
- Invitation letter, Expert Sheet, etc.
- "Line list"
- Elicitation protocol
- Instructions for facilitators
- Instructions for experts
  - Calibration questions
  - Target questions
- Answer sheets for
  - Calibration questions
    - 1 set per panel
    - a summation sheet per facilitator
  - Target questions
    - 1 Excel work book per panel

27

Progress so far

DTU

295 potential experts invited

↓

152 replied positively

↓

117 actually send CV and expert sheet

↓

2 were not cleared by WHO (DoI)

↓

96 enrolled – several serving on more panels

↓

80 interviews conducted

↓

49 target question sheets returned

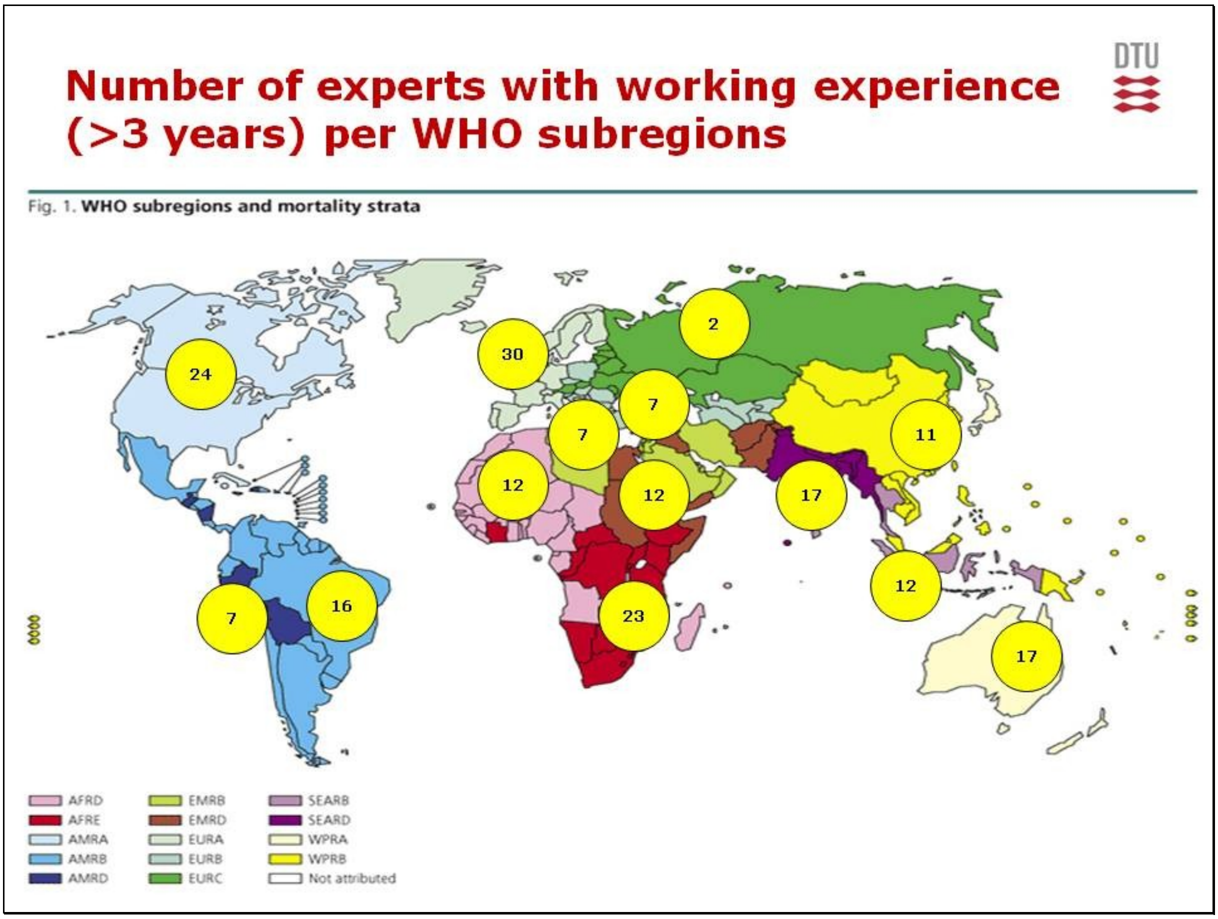
28

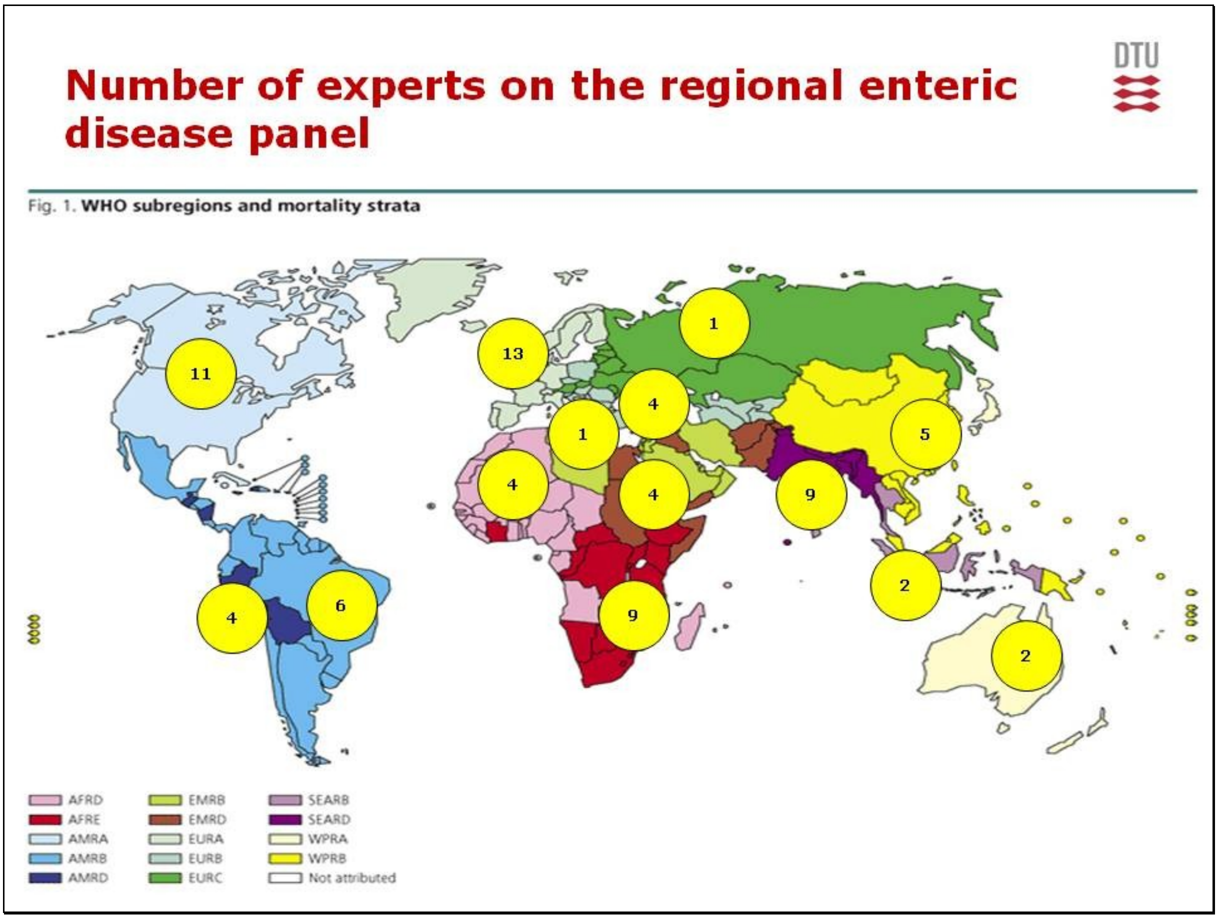
Progress so far

DTU

Panels		Experts		
		#	Interviewed	Returned answers
Heavy metals	Global	11	3	1
Protozoa	Global	10	8	6
Echinococci	Global	9	6	5
Toxoplasma	Global	11	10	6
Ascaris	Global	10	6	5
Brucella	Global	7	5	4
Hepatitis A	Global	8	7	4
Enteric pathogens	Regional	36	35	18
Total	**8**	**102**	**80**	**49**

29


Number of experts with working experience (>3 years) per WHO subregions
DTU
Fig. 1. WHO subregions and mortality strata
24
30
2
7
7
11
12
12
17
7
16
23
12
17
AFRD
AFRE
AMRA
AMRB
AMRD
EMRB
EMRD
EURA
EURB
EURC
SEARB
SEARD
WPRA
WPRB
Not attributed


Number of experts on the regional enteric disease panel
DTU
Fig. 1. WHO subregions and mortality strata
11
13
1
4
1
5
4
4
9
4
6
9
2
2
AFRD
AFRE
AMRA
AMRB
AMRD
EMRB
EMRD
EURA
EURB
EURC
SEARB
SEARD
WPRA
WPRB
Not attributed

Lessons learned/Challenges

DTU

- Agree on point of attribution for food exposures
- Identification and enrolment of expert from low income-high mortality subregions
- Keeping experts committed after enrolment
- Construction of regional panels for the enteric diseases
- Formulation of calibration questions
  - Appropriate for all experts (multidisciplinary panels)
  - Similar to target questions
- High degree of volunteer/in kind contributions
  - Limited time commitment
  - Everything takes longer than expected!
  - Facilitators paid a minor fee per interview
- The vast majority of experts were very positive about the study and agreed to its relevance
  - A few questioned the reliability of the outcome

32

Summary

DTU

- WHO is preparing the 1st global estimates of the burden of foodborne disease
- Attribution to risk factors/sources is an important part of burden of disease studies
- Source attribution of foodborne disease is difficult to do even in high income countries
- Expert elicitation provides transparent means of providing source attribution estimates where other methods and/or data are unavailable

33

Acknowledgements

- Facilitators: Charlene Bayemi, Chloe Liu, Anca Hanea, Taher Jamshidi, Vikram Rao, Marshall Kuypers, Alessandra Lafranconi, Anuj Jindal, Sarvnaz Modaressi, Lohendy Muñoz, Jessica Kratchman, Fjordor Melnikov
- Sandy Hoffmann (USDA), Willy Aspinall (Univeristy of Bristol) and Roger Cooke (RFF)
- FERG member, particularly Arie Havelaar (chair) and the task force chairs: Herman Gibb, Paul Torgerson, Fred Angulo/ Martyn Kirk
- WHO secretariat: Amy Cawthorne, Tim Corrigan, Tanja Küchenmüller, Claudia Stein

AND OF COURSE ALL THE EXPERTS!!

34

GLOSSARY

The following terms have specific meanings in the context of this Guidance, which might be different from the common understanding.

Behavioural aggregation *Behavioural aggregation* means to combine the judgements of several experts via a moderated discussion concluding in a consensus. A structured approach is the → *Sheffield method.*

Background information The *background information* summarize all existing information on the context of the risk assessment and the existing evidence obtained by the → *working group*, which is necessary for the experts to perform the elicitation from a well-informed starting point.

Cooke's method *Cooke's method* uses performance-based linear pooling to combine the judgements of several experts. The experts can be calibrated according to their performance on additional → *seed or calibration variables or questions.*

Delphi method The *Delphi method* uses repeated individual questionnaires to combine the judgements of several experts. In a subsequent round the experts receive the anonymous results of the previous round as feedback and are able to revise their answers. Finally remaining differences are aggregated using equally weighted pooling.

Elicitor The *elicitor* chairs the → *elicitation group* and leads the elicitation.

Elicitation Group The *elicitation group* performs the → *elicitation protocol* and elicits the information from the → *expert panel*. The *elicitation group* is responsible for all contacts with the → *expert panel,* the documentation of the → *elicitation phase*, the → *result report*, and the feedback to the → *experts.*

Elicitation phase The *elicitation phase* is the time of performing the elicitation protocol, starting with the invitation of the →*expert panel*, the information and training of the experts, the elicitation, and ending with the reporting of the results.

Elicitation process The *elicitation process* comprises the whole process of an expert knowledge elicitation, starting with the problem definition (→ *initiation phase*) by the → *working group*, the development of the → *elicitation protocol* (→ *pre-elicitation phase*) by the → *steering group*, the performance of the elicitation (→ *elicitation phase*) by the → *elicitation group*, ending with the technical documentation and evaluation of the results (→ *post-elicitation phase*).

Elicitation protocol The *elicitation protocol* explicates all selections and reasoning in defining the elicitation, esp. the framing of the problem (e.g. the elicitation questions to be asked), the selection of the experts (e.g. short list with experts to be invited) and the elicitation method (steps and timeline of the elicitation to be performed).

Expert An *expert* refers to a knowledgeable, skilled or trained person, e.g. staff of national authorities, consultants or professionals, practitioners from industry, or specialists involved in specific processes

Expert panel The *expert panel* is a group of experts selected for elicitation according to an → *elicitation protocol.*

Expert Knowledge Elicitation *Expert Knowledge Elicitation* is a systematic, documented and reviewable process to retrieve expert judgements from a group of experts (→ *expert panel*) in the form of a → *probability distribution*.

Expert role The expert role is the social role (e.g. job title, job function) or background (e.g. country, region of origin) of an individual → *expert* in the → *expert panel*. The *roles* are defined by the → *steering group*.

Expertise profile The *expertise profile* is a list of desirable characteristics, knowledge and skills of an individual → *expert* to be included in the → *expert panel*. The *profiles* are defined by the → *steering group*.

Facilitator Synonym for → *Elicitor.*

Initiation phase The *initiation phase* is the time to identify and justify the need for an → *expert knowledge elicitation.* The → *working group* is responsible to define the problem, review the existing evidence and conclude on lacking evidence.

Mathematical aggregation *Mathematical aggregation* means to combine the individual judgements of several experts by a mathematical formula. Typical methods are Bayesian aggregation or pooling (e.g. → *Cooke's method*).

Post-elicitation phase The *post-elicitation phase* is the time after the elicitation, when the whole → *elicitation process* will be completed, documented and reviewed. Responsibility of the → *steering group* is to evaluate the compliance with the → *elicitation protocol*. Finally the → *working group* will evaluate the results for use in the risk assessment.

Pre-elicitation phase The *pre-elicitation phase* is the time of developing the → *elicitation protocol,* comprising the framing of the problem, the selection of the experts and the elicitation method, incl. setting the timeline. The → *steering group* is responsible for developing the → *elicitation protocol*.

Probability distribution A *probability distribution* is a thorough description of uncertainty regarding a quantity. It is built up from a series of expert judgments about ranges of the uncertain quantity containing the true value with a particular probability.

Profile matrix The *profile matrix* combines expertise profiles with roles in the intended → *expert panel* to identify required or missing experts.

Result report The *result report* is the summary of the elicitation and the final result for the use in the risk assessment procedure.

Seed variable / question *Seed or calibration variables* are variables from the experts' field whose realisations are (or will be) known to the analysts, but unknown to the experts. → *Cooke's method* assumes that the (future) performance of the experts on the variables of interest can be judged on the basis of their (past) performance on the *seed variables*.

Sheffield method The *Sheffield method* uses the Sheffield Elicitation Framework (SHELF) to structure the moderated group discussion during a face-to-face workshop to reach an appropriate expert consensus.

Steering Group: The *steering group* is proposed by an EFSA → *working group* to develop the → *expert elicitation protocol* during the → *pre-elicitation phase*. The group comprises domain experts (from the → *working group*), experts on elicitation, administrative staff and the → *elicitor,* once appointed.

Working Group: The *working group* or network is established by EFSA, a responsible panel or the Scientific Committee to address a mandate of EFSA. The *working group* is responsible to review the existing evidence and to initiate an → *expert knowledge elicitation*, if needed. For this the *working group* decides on the context (e.g. the risk assessment model, existing evidence → *background information*) and the elicitation problem (e.g. the parameter to be elicited). Finally the *working group* proposes a → *steering group* to develop the → *elicitation protocol*.

ABBREVIATIONS

AHAW	EFSAs Animal Health and Welfare Panel / unit
CDF	Cumulative Distribution Function
CV	Curriculum Vitae
EKE	Expert Knowledge Elicitation
EG	Elicitation Group
EP	Expert panel
EXCALIBUR	EXpert CALIBRation Software
GEM	Generalized Expertise Measure
L	Lower bound
M	Median
PDF	Probability Density Function
PLH	EFSAs Plant Health Panel / unit
Q1	Lower quartile
Q3	Upper quartile
SHELF	SHeffield ELicitation Framework
SG	Steering Group
U	Upper bound
WG	Working Group