Uncertainty quantification with experts: present status and research needs

Anca M. Hanea*[1], Victoria Hemming[2], and Gabriela F. Nane[3]

[1]Centre of Excellence for Biosecurity Risk Analysis, The University of Melbourne, Melbourne, Victoria, Australia

[2] Department of Forest and Conservation Sciences, The University of British Columbia, Vancouver, Canada

[3]Delft Institute of Applied Mathematics, Delft University of Technology, Delft, The Netherlands

Abstract

Expert elicitation is deployed when data are absent or uninformative and critical decisions must be made. In designing an expert elicitation, most practitioners seek to achieve best practice while balancing practical constraints. The choices made influence the required time and effort investment, the quality of the elicited data, experts' engagement, the defensibility of results, and the acceptability of resulting decisions. This piece outlines some of the common choices practitioners encounter when designing and conducting an elicitation. We discuss the evidence supporting these decisions and identify research gaps. This will hopefully allow practitioners to better navigate the literature, and will inspire the expert judgement research community to conduct well powered, replicable experiments that properly address the research gaps identified.

Keywords: Expert elicitation protocols, Uncertainty quantification, CM, IDEA, SHELF.

*Address correspondence to Anca M. Hanea, Center of Excellence for Biosecurity (Cebra), University of Melbourne; anca.hanea@unimelb.edu.au

Pre-print: Hanea, A.M., Hemming, V., Nane, G.F. (2021) "Uncertainty quantification with experts: present status and research needs". *Risk Analysis* http://doi.org/10.1111/risa.13718

1 INTRODUCTION

Risk assessment is often viewed as an objective and scientific basis for making decisions. It involves identifying events (hazards) of interest, and understanding the likelihood and consequences of these events on things that we care about. For example, consider the impact of an invasive species or a novel virus on human lives, threatened species, and economies. The inherent uncertainties and consequences need to be quantified and combined to inform the assessment of risk, and the subsequent risk management decisions.

The process of assessing risk and informing decisions, even after being formalised, is not only informed by experts and data, but it also involves value judgements. Scientific judgement is in itself value-laden, and bias and context are inescapable in data collection and analyses. Hence, facts and values often become entangled, and yet, ideally, form a modelling perspective they should be kept separate, as far as is possible. This distinction is important at a qualitative level, when building conceptual models, and at a quantitative level, when populating conceptual models with numerical parameters. Moreover, at both levels, this distinction will help direct the search for, and use of, appropriate expertise and data. A fairly recent book concerning these matters is Dias et al. (2018). The book is about "the facilitation of the quantitative expression of subjective judgement about matters of fact, interacting with subject experts, or about matters of value, interacting with decision makers or stakeholders." In this piece we are concern solely with the former, namely with expert elicitation for uncertainty quantification of matters of fact.

A vast array of approaches for eliciting expert judgements about uncertainty exist. Some of these approaches are supported by empirical research, some by practicality, while others propose interesting ideas but require more research. However, for anyone seeking to deploy an expert elicitation, the current state of knowledge can be a an unwieldy, confusing and contradictory field to navigate because: 1) the terminology is used inconsistently[1] 2) it almost never directly compares available elicitation protocols[2] and 3) it may be outdated. As a result, practitioners may adopt methods for elicitation applied in their own fields, without understanding that these may have been implemented out of practicality, have not yet been tested, or may even go against best practice.

In an attempt to summarise the current status of research, the editors of Hanea et al. (2021) collected most recent advances on the theory, methods and practice of drawing quantitative judgments from panels of experts to aid risk analysis and decision making. Complementing Hanea et al. (2021), similar recent efforts are made by the authors of Winkler et al. (2019) and McAndrew et al. (2020). The former is an opinion

[1] The authors of McAndrew et al. (2020) discuss this issue at length.
[2] with a couple of exceptions, e.g., EFSA (2014); Williams et al. (2020)

piece, whereas the latter is a review paper, and both are concerned with the different approaches for the combination/aggregation and evaluation of multiple expert judgements.

We aim to synthesise comparative advice and best practices while discussing practical constraints and limitations of elicitation protocols.

1.1 What to expect from this piece

In this paper we discuss uncertainty quantification with experts. This excludes verbal or qualitative descriptions of uncertainty and does not cover qualitative methods for problem formulation, value judgments, or stakeholders' preferences.

We outline instead, some of the common decisions practitioners encounter when designing and conducting an elicitation. These decisions are discussed in the context of a formalised protocol (Section 2) through a comprehensive collection of steps (Section 3). We emphasize the evidence supporting these steps and identify research gaps. These gaps are summarised in Section 4.

This piece will hopefully allow practitioners to better navigate the literature and to discern the importance of each step. Ideally, the expert judgement research community will be inspired to conduct well powered, replicable experiments that properly address the research gaps identified.

1.2 A few words of advice

Expert elicitation is a decision *support* tool. It provides quantitative estimates of likelihoods, unknown variables, and estimates model parameters when data are absent or uninformative. This means it is only one piece of a complicated puzzle. Before engaging in an elicitation, several initial steps must be undertaken. These are: defining the decision problem, specifying objectives, developing conceptual models, and identifying the modelled variables (e.g., Gregory et al., 2012). Our paper assumes the practitioners have undertaken these steps.

Expert judgment *should not* replace empirical data, but should only be used when other forms of data are insufficient or inappropriate, and opportunities to obtain more data are limited[3]. This means that practitioners should 1) have a clear idea of what data would be appropriate to collect, and 2) identify appropriate available data sources *before* engaging in an expert elicitation.

Expert elicitation *requires* methodological rigor. Like any other type of data, expert judgements can be prone to errors, contextual biases or other limitations. As such, the process employed should derive

[3]A hierarchical way to categorise data and evidence is presented for example in Pullin and Knight (2003)

the best possible judgements from experts and subject these judgements to the same level of empirical control and transparency as would be expected from data-driven methods (Cooke, 1991; French, 2012; Drescher and Edwards., 2018). The need to use expert judgement across almost every domain has prompted research on how to do this best, and how to organize this process into a ***structured protocol***. These protocols are a collection of steps required prior to, during, and post an expert elicitation. The steps are informed by research and applications in decision theory, statistics, social sciences, psychology and engineering. However, advice for implementing each step is far from unanimous. This advice may concern, e.g., the definition of expertise, the number of experts needed, the diversity of the expert group, the choice of questions and their format, the extent of feedback provided to experts, how experts interact during the elicitation. Some guidelines have been tested, some are suggested for practical reasons, while others are unsupported.

2 STRUCTURED ELICITATION PROTOCOLS

We claim that improving expert judgements can only be done through structured protocols. Structured elicitation protocols combine various steps and advice into a formalized protocol[4]. In our interpretation, a structured protocol includes the opinions of more than a single resident or eminent scientist or practitioner, avoids unstructured round table discussions, and is designed to answer questions that are hypothetically verifiable. Even though structured protocols have been increasingly adopted in various application areas, informal methods continue to prevail. We classify informal[5], unstructured protocols as simply unscientific, and shall not cover these here.

Throughout the paper, we refer to three established protocols often applied for quantifying uncertainty. These are: the IDEA protocol (Hemming et al., 2018a), the SHELF protocol (O'Hagan et al., 2006) and the Cooke's protocol (Cooke, 1991), also known as the Classical Model (CM). They were used in more than a hundred commissioned applications, using hundreds of experts who answered hundreds of questions (e.g., Colson and Cooke, 2017; Hemming et al., 2018b; O'Hagan, 2019). The protocols have several steps or advice in common, and several steps where they differ in their approach or advice. For example, all three protocols use multiple experts and strive to obtain aggregated estimates which represent expert groups' judgements. In contrast, aggregated estimates are obtained mathematically in CM and IDEA and using behavioural aggregation in SHELF[6]. SHELF and IDEA are similar in that they provide the opportunity of

[4]To the best of our knowledge there is no definition of structured protocol unanimously adopted by the community. Hanea et al. (2018) provides a working definition that may seem a bit restrictive. Here we address elements of that definition, emphasising their importance.

[5]We classify informal methods as those that that don't aim to meet these three elements.

[6]In later descriptions of SHELF, if consensus is not achieved an average estimate of the diverging opinions is used instead.

Table 1: The roles of the actors from the elicitation team.

Actors	***Roles***
Decision Maker	Identifies the need of the analysis, contacts analyst(s).
Analyst(s)	Designs the elicitation, contacts facilitator(s), analyses results.
Facilitator(s)	Facilitates discussion, manages experts and their interaction.
Expert(s)	Helps formulating the questions for elicitation, guides the elicitation format.

extensive discussion between experts. To the best of our knowledge, these (and other) protocols have not been compared in a well powered experiment.

Practitioners may seek to apply a specific protocol or even to design their own, such that it best fits their purpose. For this, the improvements obtained by individual steps need to be understood.

3 COMMON STEPS IN STRUCTURED PROTOCOLS

This section explores the generic steps of formal protocols and reasons behind their deployment. We focus on differences and similarities between the three above mentioned structured protocols because they represent a large array of decisions that practitioners will encounter.

3.1 Actors of an expert elicitation

3.1.1 *The elicitation team*

An elicitation team includes a decision-maker (problem owner), an analyst, a facilitator and at least one domain expert (whose role does not include answering the elicitation questions). Their roles are outlined in Table 1. These roles can be undertaken by a single person; however, it is better if the analyst and facilitator are neutral to the decision outcome.

Prior to the elicitation it is the responsibility of the elicitation team to ensure that the experts: 1) understand the importance of the study and of their own input, 2) are sufficiently motivated to provide their assessments, 3) are informed about the elicitation protocol and format, and 4) receive information of how their assessments are further processed and aggregated (e.g., Hemming et al., 2018a; EFSA, 2014).

All protocols generally agree on this broad structure of the elicitation team. A key difference is that for CM, the facilitator must only have a good understanding of the aggregation method, while for IDEA and SHELF the facilitator must be able to facilitate group discussions as well. Intuitively, a more experienced facilitator should be able to handle group interactions better, but we are not aware of any study investigating

the influence of the facilitators on group dynamics in this context.

3.1.2 *The experts*

All protocols agree that more than one expert should be convened. Several rationales are given for this. Firstly, any one person holds an incomplete representation of "the world" based on their own experience. Including more people with diverse experiences increases the knowledge about the problem and therefore should provide a better representation of the uncertainty to be quantified. Aggregates of a diverse group are known to be more accurate and better calibrated than a single well credentialed individual (e.g., Sniezek and Henry, 1989; Hemming et al., 2020b). Finally, in processes that include deliberation between experts, the inclusion of others can help cross-examine reasoning and evidence.

The definition of expertise remains a question of research and debate. Providing judgements under uncertainty requires that individuals not only possess domain knowledge but that they can adapt and communicate that knowledge. While credentials, peer-recommendations, and experts' status may seem important, research has found that there is no correlation between peer-recommendation and credentials, and the quality of judgements about uncertain facts (e.g., Mellers et al., 2015; Burgman et al., 2011). These metrics can lead to pre-judgement and exclusion of potentially knowledgeable individuals, and bias in the selection of experts (e.g., French, 2011; Shanteau and Pounds., 2003). If the aggregation is to best represent the true uncertainty then an emphasis should be placed on recruiting a diverse array of individuals, with diversity covering domain knowledge and demographics (i.e. age, experience, gender). These are proxies for cognitive diversity (e.g., Keeney and von Winterfeldt, 1991; Page, 2008).

Studies examining the optimum number of experts required have mostly focused on investigating the changing performance of the group (when averaging judgements) as the group size increases. Studies based on simulations and experiments suggest that a number between 5-12 is sufficient for obtaining a relatively robust equal weighted aggregation (e.g., Clemen and Winkler, 1999; Vercammen et al., 2019). Group sizes below five are sensitive to outliers, while above 12, the input of additional individual adds very little to the existing information. For methods like IDEA and SHELF which include extensive discussion between experts, larger groups will be difficult to manage. CM advises on a minimum of 5, but it theoretically has no upper limit. However, for practical reasons, the current advise for all protocols is anything between 4 and 10, (e.g., O'Hagan, 2019; Hanea and Nane, 2020). Recruiting a few additional experts is recommended to ensure sufficient numbers for the final aggregation, even in case of dropouts.

The discussion above however does not advise on how to identify experts a-priori, neither does it prescribe

a clear size and composition (in terms of diversity) of a group that can be considered optimal. These choices remain subjective and often driven by resources.

3.2 Questions for uncertainty quantification

After a conceptual model allowed the analysts to identify the data gaps and the required expert input, and the elicitation team and expert group are assembled, the actual questions for experts need to be formulated. Questions should be clearly formulated, so that all experts have the same understanding of the questions and context. There are two ways in which uncertainty is acknowledged and modelled through expert elicited data. One is to ask the expert group about point estimates of (theoretically) measurable variables and take the variability between experts' answers as a measure of uncertainty. Another is to ask a panel of experts about their subjective distribution associated with a (theoretically) measurable variable. The latter approach, even though more demanding, is recommended as it captures much more information and a better description of uncertainty.

Quantifying uncertainty may take two forms: eliciting probabilities, or eliciting values of a continuous variable corresponding to different percentiles. Eliciting probabilities can take several forms depending on the variables involved in the model. These may correspond to: 1) probabilities of event occurrences; 2) probabilities of various states of discrete variables; 3) conditional probabilities for combinations of discrete states of variables, e.g., conditional probability tables in Bayesian networks[7].; or in the case of a continuous variable, 4) the probabilities of fixed values of a variable[8].

What we appeal to, when asking for probabilities, are the different interpretations of probabilities available to the experts; such probabilities can be either thought of as subjective degrees of belief, or as relative frequencies. An extra complication associated with eliciting probabilities is the quantification of the imprecision in such estimates within a probabilistic framework. To account for such imprecision, upper and lower bounds are asked for, in addition to a best estimate for a probability. When the probabilities can be interpreted as relative frequencies, the bounds can be interpreted as percentiles of the experts' subjective probability distribution. However, when the relative frequency interpretation is not appropriate (i.e., when eliciting the probability of unique events) the bounds may be criticised for lacking operational definitions[9].

[7]These discrete variables can be genuinely discrete, or discretized variables that are theoretically measured on a continuous scale. Sometimes, the variables modelled and quantified are defined using constructed scales, e.g. airplane pilot fatigue. These scales are often vague and subjective and hence open to criticism.

[8]Some argue that answering questions about probabilities is more difficult than answering questions about quantities because a "probability doesn't exist". Hence, when continuous variables are modelled and their distribution needs to be elicited, we recommend eliciting values of continuous variables corresponding to a finite number of different percentiles, typically three.

[9]The imprecise probability framework may be invoked, but we class it as an alternative to probability theory and do not discuss it here.

However, the IDEA and SHELF practitioners argue that the main reason to elicit bounds in such cases is to improve thinking about the best estimates (O'Hagan et al., 2006; Hanea et al., 2018).

We suggest to *always* ask for bounds. If the probabilities in question can be represented as relative frequencies, we suggest formulating the questions in terms of relative frequencies, treating these as continuous variables and ask for percentiles of the experts' subjective distribution. If the probabilities needed for the model correspond to unique events, asking for bounds[10] provokes counterfactual thinking and helps with the estimation of the best estimate. However, we then recommend using only the best estimate in further probabilistic analysis. The conjecture that asking about bounds improves the estimation of the best estimate (in the case of unique events' probabilities) is a sensible but insufficiently tested intuition.

A summary of what can be elicited is presented in Table 2. The approach that modellers end up choosing depends on the context, the needs of the chosen probabilistic model where the elicited estimates will be embedded in, the resources available, and the familiarity of the experts with probabilistic concepts. Even though more than one option can be chosen, it is unusual for elicitations to take various forms for purely theoretical investigations. To our knowledge there are no experiments designed solely to compare and contrast these alternatives.

Table 2: Elicited estimates when quantifying uncertainty

Variable type	Estimate elicited	Treatment of uncertainty	Interpretation of probability
Discrete	—probability of event occurrences	point estimate / point estimate & bounds	subjective[11] / frequentist
	— (conditional) probabilities of discrete states	point estimate / point estimate & bounds	subjective / frequentist
Continuous	—probability of a given value	point estimate	subjective
	—percentiles of a distribution	credible interval	subjective distribution of the expert

There is little known about how many questions can be asked in a day of an elicitation. Trade-offs between the number of questions and 1) the quality of elicited judgements, and 2) the retention of experts, are speculated, with more questions leading to declines in both. Advice by Hemming et al. (2018a) suggests 20 questions in a day is reasonable ask. Others have reported being able to ask many more, but to our knowledge no study attempted to quantify the effects of the number of questions on performance.

[10] We note that asking for bounds in this setting is against the CM recommendation.

[11] The probability itself is a quantification of uncertainty. Second order uncertainty may be expressed through bounds. These bounds can be operationalized as percentiles of the experts' subjective distribution on an unknown relative frequency. If the probability is given for an one off event, then the bounds cannot be interpreted within the classical probability theory framework.

3.3 Training experts and facilitators

Reasoning under uncertainty is far from being a trivial endeavor. Expert training is typically recommended in the guidelines for structured protocols (e.g., O'Hagan et al., 2006; Cooke and Goossens, 2008; Gosling, 2018). Most protocols recommend that the experts should be trained in the use of the elicitation method, and in assessing uncertainty (a.k.a. probability training). Training is thought to: 1) reduce experts' apprehension, 2) increase the understanding of the process by which expert judgements are collected and aggregated, 3) motivate the experts, 4) identify common biases and the extent to which experts are predisposed to these, and 5) provide the facilitators and team with guidelines for working with the expert group (Hodge et al., 2001; Revie et al., 2011; Keeney and von Winterfeldt, 1991).

All this guidance has emerged from good practice rather than formal studies. The effectiveness for improving judgements, and relieving apprehension is yet to be rigorously quantified.

In training experts, practice questions are recommended, since they allow experts to quantify their uncertainty in the format employed by the elicitation method. Due to the scarcity of domain questions, some practitioners opt for domain neutral questions (e.g.,from sports, weather). Furthermore, they argue that domain neutral questions would permit experts to focus on uncertainty quantification rather than on domain specific matters. For SHELF and IDEA, a domain practice question may come with the disadvantage of a prolonged discussion. However, this approach is recommended by some practitioners, , e.g., Hartford and Baecher (2004), who claim that professionally engaging experts with domain specific practice questions is beneficial. We are not aware of formal studies that investigate the benefits of either approach. It seems prudent to include both type of questions, if possible.

Proving feedback on the training question is also thought to be good practice. However, aside from some studies on undergraduate students (Stone and Opel, 2000; Subbotin, 1996), no formal study has proven the effectiveness of training and feedback for structured protocols.

The effectiveness of the role and influence of the facilitator for any of the structured protocols is yet to be rigorously quantified. Nonetheless, facilitation is a skill, and for those approaching elicitation for the first time it can be a daunting process. Some advice for facilitation for the CM, IDEA, and SHELF protocols can be found (Bonano et al., 2020; Hemming et al., 2018a; Hart et al., 2016). These documents followed from training courses developed for EFSA[12] or within a COST action[13].

[12]European Food Safety Agency
[13]https://expertsinuncertainty.net/

3.4 Interactions

Once the experts are trained and ready to provide assessments, would it be beneficial for the experts to interact? If they do, should they interact before or after answering questions and how much interaction is beneficial?

All protocols agree that it is crucially important that experts provide their initial judgements (together with rationales) privately and independently. Retaining this independence of their initial assessment helps to mitigate against group think and deference to dominating personalities. The interaction between experts is kept to the minimum, prior to answering the questions, with the scope of discovering and correcting most of the misunderstandings and ambiguities.

The experts' initial judgments can be used to formulate an aggregated result, and this is what the CM protocol proposes. Experts do not get feedback on what others have thought, and are not given the chance to change their mind based on a debate of reasons provided by their peers.

The (dis)advantages of feedback and discussion and the way these may improve or degrade the quality of judgements have been extensively discussed in Hemming et al. (2020a) and the references therein. One disadvantage of discussion is the introduced dependence between experts' assessments. However some consider that the benefits of feedback are greater than the disadvantage of introducing dependence. The evidence for such trade-offs is little in the context of eliciting uncertainty with groups of experts (Hanea et al., 2016; Wilson and Farrow, 2018), but is informed by research discussed in Hemming et al. (2020a).

The discussion phase helps identify and debate the sources of evidence that underpin the judgements. This additional evidence, in the form of rationales (also asked for, but not discussed in CM) is critically important for decision-makers. It can also provide an early indication that experts have developed alternative interpretations of the questions, allowing the elicitation team to refine the questions before the elicitation concludes. These qualitative advantages are often strong motivators for including a discussion phase.

Unless a subsequent round of individual estimates is recorded, it is hard to know if discussion and feedback improved individual performance. The IDEA protocol deploys two rounds of elicitation. The experts are asked to provide a private individual estimate, and estimates are mathematically aggregated. The SHELF method differs by asking the experts to create a consensus distribution, after feedback and discussion, hence the final results obtained by SHELF correspond to the group only. A few IDEA elicited data sets suggest that experts tend to strongly anchor on their initial judgements and only adjust if they hear good reasons to do so (e.g., Hemming et al., 2018a; Hanea et al., 2018). In addition, experiments also show that this discussion can improve individual judgements, and usually improves group judgements (e.g., Hemming et al., 2020a,

2018a; Hanea et al., 2018). A very recent study (Williams et al., 2020) suggests that the group judgement obtained through SHELF in an elicitation with three experts was better than the three individual initial estimates.

Approaches applied in the literature include variants which involve multiple rounds of estimation which continue until consensus is achieved, or responses do not change, they also vary in how experts are allowed to interact with some methods just including feedback, and some enabling discussion and feedback. To our knowledge the critical mass of evidence that any of these variations produces more reliable results is yet to be collected.

The interactions between experts, may happen in a face to face environment, or in a remote environment (i.e. over email, or webinar). The choice is often intuition-based and mainly dictated by practical constraints (like a pandemic) rather than theoretical considerations. To our knowledge, only a handful of studies (e.g., Baker et al., 2014; Grigore et al., 2017) compared these formats and their influence on final estimates. Much more research is needed for strong recommendations.

3.5 Aggregation

We have already touched upon different aggregation methods of experts estimates. Even though, sometimes there are arguments for not aggregating (e.g., Morgan, 2015), but rather presenting the decision maker with a portfolio of estimates (mostly when more divergent "schools of thought" generate very different estimates), most often one aggregated estimate is needed.

The two main ways in which expert judgements are pooled are using *behavioural aggregation*, which involves striving for consensus via discussion, or using *mathematical aggregation* which provides a more explicit and objective approach to aggregation[14]. A weighted linear combination of opinions is one example of such aggregation. Equal weighting is often used because of its simplicity. Evidence also shows that the equal weighting scheme frequently performs quite well relative to more sophisticated aggregation methods (e.g., Clemen and Winkler, 1999), but not always (e.g., Cooke, 2015; Mellers et al., 2014; Hemming et al., 2018a).

We recommend that differential weighting based on anything other than prior expert performance on similar tasks (i.e. quantifying uncertainty) should be avoided. A few examples from literature, where self-ratings, peer-ratings, and citation indices were used to obtain weights (Woudenberg, 1991; Burgman

[14]Mixed protocols combine behavioural and mathematical aggregation methods (Ferrell, 1994). IDEA is a mixed protocol with the twist in the sense that it does not seek consensus, but rather encourages counterfactual thinking and evidence based debate. It allows for a second round of independent estimates which are mathematically aggregated using equal or differential weights (calculated as in CM).

et al., 2011; Cooke et al., 2008) support this recommendation. Arguably the most widely used version of a differential weighting scheme is CM, which uses calibration variables[15] to derive performance based weights[16] for continuous probability disttributions.

Performance-based weights have been shown to lead to aggregated assessments which are, most of the times, more calibrated and informative than assessments resulting from equal weighting of expert assessments. In Hanea and Nane (2020), results of 322 experts from 63 professional studies were used to assess the performance of aggregated assessments obtained using performance, versus equal weights. In Williams et al. (2020), a study with three experts compared the performance of SHELF with that of an equally weighted aggregation and a form of cross validated performance weighted aggregation. To our knowledge this is the first time when SHELF (i.e. the behavioural aggregation) was validated and assessed for performance, so even though these are only very preliminary results, the effort is most laudable. More studies to compare the performance of aggregated assessments are definitely needed.

3.6 Validation and calibration

To allow comparisons of different aggregation methods, studies need to use questions whose answers become available in a reasonable time frame. Unfortunately very few such situations exist, so true out of sample validation studies are very rare. What is possible though are in sample or cross validation studies and these studies are possible when calibration questions are used. The analysis in Colson and Cooke (2017) remains, to date, the largest cross-validation analysis in the field of structured experts judgment.

One of the key considerations however, is the development of *good* calibration questions. There is little known about what makes a good calibration question[17]; however, the main assumption of measuring prior performance on calibration questions is that this measure is a good predictor of future performance on the target questions. Therefore having calibration questions which are representative for the target questions is paramount. Because trust in weighted aggregations depends on the development of good calibration questions, a strong recommendation for analysts is to consult with domain experts when developing calibration questions[18]. The main types of calibration variables are discussed in detail in Cooke and Goossens (2000) and Hanea and Nane (2020) and the references therein.

[15]Calibration variables are domain questions for which the realizations are known, or will become known, within the time frame of the study Aspinall (2010).

[16]It it worth mentioning that performance weighted aggregation techniques, other than the CM, have been developed. While this has become an area of research in itself, few methods have been substantially applied to real expert elicitation studies or adhere to the requirement of proper scoring rules

[17]Attributes of calibration questions are discussed in Hemming et al. (in review).

[18]These experts should be part of the elicitation team. Given their involvement with the calibration questions, their judgements cannot be formally elicited during the elicitation.

Ideally, the analysts should have access to ongoing experiments or relevant data which become available shortly after the elicitation. When the above is not possible, data from recent studies within the subject matter or adjacent subject matters are often the only option. When elicitations involve more sub-domains, the set of calibration questions should cover each of them. This is easier said than done since the boundaries between sub-domains are often blurred and we are yet to learn how well can experts extrapolate their knowledge to answer questions from adjacent domains (see Hemming et al. (in review) and references within).

The calibration questions should be asked in exactly the same format as the target questions, since there is no reason to believe that good performance on a certain type of task is transferable to different tasks. The calibration questions should be triggering the same type of thinking as needed when answering the target questions, that is experts need to be able to make challenging judgements of appropriate, composite uncertainties.

Even when performance weighted aggregation does not outperform equal weighted, having calibration questions provides the only means one can use to choose between different aggregations based on their "quality". This can help to justify and defend the final representation of uncertainty. Moreover, having calibration questions is the only way one will have the option of performing in-sample or cross validation exercises when using mathematical aggregations. However, when using a behavioral aggregation method cross validation is not possible.

Calibration questions are not adopted with ease by the expert judgement community, apart from the proponents of CM. They are time consuming, they may not be representative for the target questions, they place an extra burden on experts, etc. However, without calibration questions validation is nearly impossible.

4 RESEARCH NEEDS AND CONCLUDING REMARKS

Many of the topics discussed in the previous sections pointed out the lack of controlled, highly powered experiments, which may confirm (or infirm) intuitions formed on signals from adjacent research. We summarize the research gaps in Table 3.

The ever changing science of expert elicitation is, relatively speaking, still in its infancy. The need for thorough investigations prior to introducing variations to protocols is maybe being overwritten by the urgency of the embedding models much needed quantification. However, if such behaviour pertains, there will be no guarantees of repeatability across different novel protocols.

Even though many of the constituting steps of the current protocols have passed the test of time and

Table 3: Identified research gaps

Research Themes	***Research Questions***
Facilitators (Section 3.1.1)	What is the influence of the facilitators on the group dynamic?
Experts (Sec. 3.1.2)	How do we determine a-priori who is an expert? How many experts in an optimal group? How diverse is diverse enough?
Questions (Sec. 3.2)	What format is best? Does thinking about bounds provoke counterfactual thinking? How many questions are too many for a day?
Training (Sec. 3.3)	What practice questions are best in experts' training? Does de-biasing work? What facilitators' skills are desirable and how are they measured?
Interactions (Sec. 3.4)	How much interaction leads to too much dependence? Can the quality of interaction be measured? Remote vs face-to-face
Aggregation (Sec. 3.5)	Behavioural vs. Mathematical Equal vs. performance based weighting - is it worth it? When not to aggregate
Calibration questions (Sec. 3.6)	What makes good calibration question? How to cross-validate behavioural aggregation? Should we always have calibration questions for validation purposes?

have been proven beneficial over and over again, such investigations should continue, every time a variation is proposed. In an ideal setting, cost-benefit analyses would be performed to understand how much improvement each incremental step of the elicitation actually derives for the costs involved.

In the meantime, if previously collected expert elicited data can be analysed in a statistical sensible way, this can inform meaningful experimental designs that in turn may answer many of the questions and fill many of the research gaps identified in this paper.

References

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

E. Baker, V. Bosetti, K.E. Jenni, and E.C. Ricci. Facing the experts: Survey mode and expert elicitation. Technical report, Climate Change and Sustainable Development, Fondazione Eni Enrico Mattei, 2014.

P. Bonano, A Colson, and S. French. Chapter 14. developing a training course in structured expert judgement. In A.M Hanea, G.F. Nane, T. Bedford, and S. French, editors, *Expert Judgment in Risk and Decision Analysis*. International Series in Operations Research & Management Science, Springer, Cham, 2020.

M.A. Burgman, M. McBride, R. Ashton, A. Speirs-Bridge, and L. Flander. Expert status and performance. *PLoS ONE*, 6:e22998, 2011. doi: 10.1371/journal.pone.0022998.

R. Clemen and R. Winkler. Combining probability distributions from experts in risk analysis. *Risk Analysis*, 19:187–203, 1999.

A. R. Colson and R. M. Cooke. Cross validation for the classical model of structured expert judgement. *Reliability Engineering and System Safety*, 163:109–120, 2017.

R. M. Cooke. The aggregation of expert judgment: Do good things come to those who weight? *Metron*, 35: 12–15, 2015. doi: 10.1111/risa.12353.

R.M. Cooke. *Experts in uncertainty: Opinion and subjective probability in science.* Environmental Ethics and Science Policy Series. Oxford University Press, 1991.

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

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

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

L.C. Dias, A. Morton, and J. Quigley, editors. *Elicitation: The Science and Art of Structuring Judgement.* International Series in Operations Research & Management Science, Springer, Cham, 2018.

M. Drescher and R. C. Edwards. A systematic review of transparency in the methods of expert knowledge use. *Journal of Applied Ecology*, 2018. doi: 10.1111/1365-2664.13275.

EFSA. Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment. *European Food Safety Authority Journal*, 12(6):Parma, Italy, 2014.

W. Ferrell. Discrete subjective probabilities and decision analysis: elicitation, calibration and combination. In *Subjective probability*, volume Eds. Wright,G. and Ayton, P. Cambridge Press, New York, 1994.

S. French. Aggregating expert judgement. *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales.Serie A. Matematicas*, 105:181–206, 2011. doi: 10.1111/j.1755-263X.2011.00165.x.

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

J.P. Gosling. Shelf: the sheffield elicitation framework. In *Elicitation*, pages 61–93. Springer, 2018.

R. Gregory, L. Failing, M. Harstone, G. Long, T. McDaniels, and D. Ohlson. *Structured decision making: a practical guide to environmental management choices*. John Wiley & Sons., 2012.

B. Grigore, J. Peters, C. Hyde, and K. Stein. EXPLICIT: A feasibility study of remote expert elicitation in health technology assessment. *BMC Medical Informatics and Decision Making*, 17:1–10, 2017.

A. Hanea, M. McBride, M. Burgman, and B. Wintle. Classical meets modern in the idea protocol for structured expert judgement. *Journal of Risk Research*, 2016. doi: 10.1080/13669877.2016.1215346.

A.M Hanea and G.F. Nane. An in-depth perspective on the classical model. In A.M Hanea, G.F. Nane, T. Bedford, and S. French, editors, *Expert Judgment in Risk and Decision Analysis*. International Series in Operations Research & Management Science, Springer, Cham, 2020.

A.M. Hanea, M. McBride, M.A. Burgman, and B. Wintle. The value of performance weights and discussion in aggregated expert judgments. *Risk Analysis*, 38 (9):1781–1794, 2018.

A.M Hanea, G.F. Nane, T. Bedford, and S. French, editors. *Expert Judgment in Risk and Decision Analysis*. International Series in Operations Research & Management Science, Springer, Cham, 2021.

Andy Hart, Anthony O'Hagan, John Quigley, and Fergus Bolger. Training course on steering an expert knowledge elicitation. *EFSA Supporting Publications*, 13(5):1009E, 2016.

D.N.D. Hartford and G.B. Baecher. *Risk and uncertainty in dam safety*. Thomas Telford, 2004.

V. Hemming, M.A. Burgman, A.M. Hanea, M.F. McBride, and B.C. Wintle. A practical guide to structured expert elicitation using the idea protocol. *Methods in Ecology and Evolution*, 9:169–180, 2018a.

V. Hemming, T. Walshe, A.M. Hanea, F. Fidler, and M.A. Burgman. Eliciting improved quantitative judgements using the idea protocol: A case study in natural resource management. *PLOS ONE*, 13(6):1–34, 06 2018b. doi: 10.1371/journal.pone.0198468. URL `https://doi.org/10.1371/journal.pone.0198468`.

V. Hemming, N. Armstrong, M.A. Burgman, and A.M. Hanea. Improving expert forecasts in reliability. application and evidence for structured elicitation protocols. *Quality and Reliability Engineering International*, 36:623–641, 2020a.

V. Hemming, A.M. Hanea, T. Walshe, and A.M. Burgman. Weighting and aggregating expert ecological judgements. *Ecological Applications*, 2020b. doi: https://doi.org/10.1002/eap.2075.

V. Hemming, A.M. Hanea, T. Walshe, and M.A. Burgman. What's a good question? the effect of irrelevant questions on performance weighted aggregation. *Risk Analysis*, this issue, in review.

R. Hodge, M. Evans, and J. et al. Marshall. Eliciting engineering knowledge about reliability during design-lessons learnt from implementation. *Quality and Reliability Engineering International*, 17(3):169–179, 2001.

R.L. Keeney and D. von Winterfeldt. Eliciting probabilities from experts in complex technical problems. *IEEE Transactions on Engineering Management*, 38:191–201, 1991.

N. McAndrew, T.and Wattanachit, G.C. Gibson, and N.G. Reich. Aggregating predictions from experts: A review of statistical methods, experiments, and applications. *WIREs Computational Statistics*, n/a(n/a): e1514, 2020. doi: https://doi.org/10.1002/wics.1514. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.1514`.

B. Mellers, E. Stone, T. Murray, A. Minster, N. Rohrbaugh, M. Bishop, and E. Chen. Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science*, 10(3):267–281, 2015.

B.A. Mellers, L. Ungar, J. Baron, J. Ramos, B. Gurcay, K. Fincher, and P.E. Tetlock. Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, 25(4), 2014.

M. G. Morgan. Our Knowledge of the World is Often Not Simple: Policymakers Should Not Duck that Fact, But Should Deal with It. *Risk Analysis*, 35:19–20, 2015. doi: 10.1111/risa.12306.

A. O'Hagan, C.E. Buck, A. Daneshkhah, J.R. Eiser, P.H. Garthwaite, D.J. Jenkinson, J.E. Oakley, and T. Rakow. *Uncertain judgements: Eliciting experts' probabilities*. Wiley, London, 2006.

A. O'Hagan. Expert knowledge elicitation: subjective but scientific. *The American Statistician*, 73(sup1): 69–81, 2019.

S.E. Page. *The difference: How the power of diversity creates better groups, firms, schools, and societies.* Princeton University Press, 2008.

A. S. Pullin and T. M. Knight. Support for decision making in conservation practice: an evidence-based approach. *Journal for Nature Conservation*, 11:83–90, 2003.

M. Revie, T. Bedford, and L. Walls. Supporting reliability decisions during defense procurement using a bayes linear methodology. *IEEE Transactions on Engineering Management*, 58(4):662–673, 2011.

D. J. Weiss R. P. Thomas Shanteau, J. and J. Pounds. How can you tell if someone is an expert? performance-based assessment of expertise. In S.L. Schneider and J. Shanteau, editors, *Emerging perspectives on judgment and decision research*, pages 620–642. International Series in Operations Research & Management Science, United Kingdom, Cambridge, United Kingdom, 2003.

J.A. Sniezek and R.A. Henry. Accuracy and confidence in group judgment. *Organizational behavior and human decision processes*, 43(1):1–28, 1989.

E.R. Stone and R.B. Opel. Training to improve calibration and discrimination: The effects of performance and environmental feedback. *Organizational behavior and human decision processes*, 83(2):282–309, 2000.

V. Subbotin. Outcome feedback effects on under-and overconfident judgments (general knowledge tasks). *Organizational Behavior and Human Decision Processes*, 1996.

A. Vercammen, Y. Ji, and M.A. Burgman. The collective intelligence of random small crowds: A partial replication of kosinski et al.(2012). *Judgment and Decision Making*, 14, 2019.

C.J. Williams, K.J. Wilson, and N. Wilson. A comparison of prior elicitation aggregation using the classical method and shelf. *arXiv preprint arXiv:2001.11365*, 2020.

K.J. Wilson and M. Farrow. Combining judgements from correlated experts. *Towards a general theory of expertise: prospects and limits*, Eds. Dias L and Morton A. and Quigley J.:vol 261. Springer, Cham, 2018.

R.L. Winkler, Y. Grushka-Cockayne, K.C. Lichtendahl, and V.R.R. Jose. Probability forecasts and their combination: A research perspective. *Decision Analysis*, 16(4):239–260, 2019.

F. Woudenberg. An evaluation of Delphi. *Technological Forecasting and Social Change*, 40:131–150, 1991.