cebra
Centre of Excellence for
Biosecurity Risk Analysis

Structured expert judgement elicitation for the attribution of food commodity and pathogen pairs

Anca Hanea, CEBRA, The University of Melbourne

Snezana Smiljanic, FSANZ

Ben Daughtry, FSANZ

1 Introduction

It was estimated by the Australian National University that foodborne disease costs Australia AUD 2.44 billion each year in their report "The annual cost of foodborne illness in Australia" . This process also involved the creation of a Cost of Illness Model for the Authority, which will hopefully be used to attach monetary costs to the outcomes of this project.

The attribution of foodborne illness costs to food commodity groups is currently only partially understood for Australia. Such information, if available, could be used to prioritise resources for research, monitoring and surveillance activities as well as standards development.

Given the challenges associated with attribution of illness to specific food types we propose to use an expert elicitation approach to systematically combine the expert judgements from across Australia.

Expert elicitation is an approach to obtaining probabilistic estimates from experts about unknown quantities or parameters. Structured expert judgement (SEJ) elicitation protocols are a collection of clear steps which have proven useful when working with experts' estimates of uncertain quantities. To the best of our knowledge there is no official definition of structured expert judgement that is unanimously adopted by the community, but Cooke (1991) provides a working definition for it. The qualifier "structured" means that expert judgement is treated as scientific data of a new type. The method formulates four necessary (but not sufficient) conditions/principles, of the scientific method as applied to expert judgement. Any scientific method for expert judgement must satisfy: *Accountability / Traceability:* all data, including experts' names, affiliations and assessments, and all analyses and software tools are available to open peer review. This guarantees that the results are reproducible by competent reviewers without violating confidentiality as experts' names are never linked to individual assessments in any open documentation. *Empirical control:* In addition to the variables of interest, experts quantify their uncertainty with respect to calibration or seed variables, that is, variables from their field for which true values are retrievable. Experts' performance as uncertainty assessors is measured based on the calibration variables. *Neutrality:* The method for evaluating and combining expert assessments should encourage experts to state their true opinion. *Fairness:* Experts are not pre-judged prior to empirical control.

An expert elicitation process using Cooke's method (also called the Classical Model - CM) was carried out in 2020, to support the FSANZ Proposal P1052 Primary Production and Processing Requirements for Horticulture (Berries, Leafy Vegetables and Melons) to estimate the attribution of foodborne illness to specific horticultural foods for four pathogens: non-typhoidal *Salmonella* species, shiga-toxin *Escherichia coli*, *Listeria monocytogenes* and norovirus. This project extends the initial elicitation since the scope has been increased to include ten pathogens.

The combination of pathogen specific foodborne illness cost estimates and expert elicitation attribution proportions will be used to provide initial estimates for costings to food commodities.

This report summarises the expert elicitation process used to complement the previous elicited estimates.

2 SEJ protocol

Maybe the most known structured protocols for eliciting expert judgements are: the Delphi protocol (whose variant, the IDEA protocol (Hemming et al., 2018) has recently gained popularity), the SHELF protocol (O'Hagan et al., 2006) and the CM (Cooke, 1991). Some of the main steps of formal protocols are common to the above mentioned structured protocols (e.g., all three referenced protocols use multiple experts and use aggregated estimates to represent the expert groups' opinion), but not all. In the current project we used the IDEA protocol in combination with the CM, which is the IDEA protocol with empirical control (i.e., calibration variables).

Briefly, the IDEA protocol involves four stages: Investigate, Discuss, Estimate, and Aggregate, as follows.

Investigate In this first phase, experts work individually using whatever resources are available to them to formulate point and interval (low—high) estimates of the unknown quantities. This step is akin to crowd-sourcing the estimation process. It should be preceded by a conference call or meeting with all experts to identify ambiguities, agree on meanings, and affirm the elicitation approach and schedule.

Expert-elicited estimates are then anonymised and summarised by a facilitation team, to be used as spurs for discussion in phase 2.

Discuss The second phase is best carried out as a face to face or remote workshop. Summaries of the anonymised phase-1 responses are presented to the group, with invitations to discuss the distribution of results — why might there be disagreement between experts, what might be reasons for particularly low or high values, etc.

Estimate The experts are then invited to privately update their first-phase point and interval estimates. They may choose not to do so. They are not obliged to disclose either way.

Aggregate The estimates from phase 2 are then mathematically transformed within expert (i.e., often distributions are fitted on the point and interval estimates) and then across expert. The mathematical aggregation is sometimes called linear pooling and consists of a linear combination of experts' distributions with equal or differential weights.

Hemming et al. (2017) provides a comprehensive description of the IDEA process.

The calibration variables used within the CM framework of empirical control allow for the calculation of differential, performance based weights to be used in the linear pooling of experts' distributions.

The answers to the calibration questions form a calibration dataset used to gauge the *quality* of expert judgements. This quality is scored in terms of how calibrated and informative expert judgements are. The calibration and informativeness scores are then combined to form differential (performance) weights. Each expert's judgements can be weighted differently in a linear combination of all judgements which will then serve as the group (aggregated) judgement. An equally weighted combination can be also calculated and compared with the performance based combination using the same proxies for *quality*.

We use the calibration and informativeness measures defined by Cooke (1991). Calibration is measured as the p-value at which the hypothesis that the expert is well calibrated would be falsely rejected. This score ranges from 0 to 1. Higher scores are better, since a low value (near 0) means that it is very unlikely that the discrepancy between an expert's probability judgements and the observed outcomes arose by chance. Informativeness is measured as the Kullback-Leibler divergence Kullback and Leibler (1951) with respect to the uniform distribution (which is what one would assume in absence of experimental or expert elicited data). Informativeness scores are larger than or equal to 0, with higher scores being better. For precise definitions and detailed analysis of these scores we refer the reader to Hanea and Nane (2020).

A good probability assessor (or a good aggregated judgement) is one whose assessments capture the true values consistently in the long run (well calibrated), with distributions that are as narrow as possible (informative). Informativeness is gauged by 'how far apart the percentiles are' relative to an appropriate background (e.g., the uniform distribution). Measuring calibration requires an analysis of the true values relative to the experts' assessments. When the lower bound, best estimate and upper bound are operationalised as the 5%, 50%, and the 95% percentiles, an expert is considered well calibrated if, in the long run, 5% of their answers fall below the fifth percentile, 90% of the answers fall between the fifth percentile and the ninety-fifth percentile, and 5% of their answers fall above the ninety-fifth percentile. In gauging overall performance, calibration is more important than informativeness. Non-informative but calibrated assessments are useful, highly informative but not calibrated assessments are not.

The performance of the experts on the calibration questions is taken as indicative for the performance on the target questions. Therefore, the calibration questions must capture the type of knowledge needed in order to answer the target questions. The more calibration variables the better, but ten has proven to be sufficient, without adding too much to the elicitation burden (e.g.,Colson and Cooke (2017)).

3 Details of the present elicitation

The authors of this report (AH, SS, and BD) organised a series of meetings to aid the parametrisation needed for the foodborne illness cost estimates modelling. The parametrisation needs were translated into questions for experts.

The elicitation scope, and the target and calibration questions were developed with the expert input of Professor Katie Glass (ANU), Dr Angus McLure (ANU), Dr Sandy Hoffmann (USDA) and Food Safety & Microbiology team at FSANZ.

The calibration questions and the target questions were organised in two separate excel spreadsheets. The target questions covered the attribution of foodborne illness to specific horticultural foods for eight pathogens: non-typhoidal *Salmonella* species, *Campylobacter* species, *Listeria monocytogenes*, *Toxoplasmosis gondii*, STEC, *Yersinia* species, *Vibrio* species, and *Bacillus cereus*. The separate food groups were: beef, lamb, pork, poultry, eggs, dairy (milk and cream, fresh uncured cheese, brined cheese, soft-ripened cheese, firm-ripened cheese), finfish, crustaceans, molluscs, fruit, grains and seeds, nuts, vegetables (fungi, leafy vegetables and herbs, root vegetables, sprouts, vine-stalk), other.

The expertise needed to answer such questions was identified, summarised, and used by the FSANZ representatives for them to personally approach and recruit experts. We aimed to gather a diverse group of experts from relevant domains: i.e., scientists, industry representatives and government representatives.

3.1 Process

The recruited experts were contacted and welcomed by the project team in February 2023. Experts were sent a Plain Language Statement and a Consent form. They were informed that they are expected to answer quantitative questions (formulated and distributed by the project team) independently.

Prior to the sending and answering of the questions, a remote inception meeting was organised on April 3, 2023. During the meeting the project team reiterated the scope and steps of the elicitation. Experts worked through (and provided feedback on) the food commodity/pathogen matrix, a practice elicitation question and the calibration questions. The first round of estimates for 12 calibration questions was elicited during this inception meeting in a controlled environment.

SS and BD incorporated the feedback from experts on the food commodity/pathogen matrix, finalised the target questions and a background document and sent them to the expert group in April 14, 2023.

Experts were initially given until May 4, 2023 to answer the questions. Delays were nevertheless inevitable. The quantitative and qualitative judgements were collected via email by the project team. Each expert was given an unique ID and the de-identified answers were collated (per target and calibration question) with the intention to be use as feedback and the starting point for a facilitated discussion.

The facilitated discussion took place during two 3-hour remote workshops, on June, 6 and 7. These meetings were recorded. Ten of the thirteen experts were present on each of the two days (marked in bold in Table 1) with eleven experts being present at least one of the days.

During the workshop, questions were discussed one by one or in groups, question formulations were tightened and clarified. The slides, the feedback plots, everyone's (de-identified) rationales and the recordings were sent to the two experts who did not participate in one of the two days.

3.2 Experts

Thirteen experts (see Table 1) were sent an invitation email to outline the scope of the elicitation, the expectations and the timeline. Even though all answered positively and signed the consent form, only eleven were able to participate.

Table 1: Experts in the Elicitation

Name	Expertise
Mark Turner	Food microbiology
Alison Turnbull	Fish health, biosecurity and seafood safety
Thea King	Food safety
Kate Astridge	Food scientist
Karen Ferres	Food safety and regulation
Robin Sherlock	Food allergen risk scientist
Mark Chan	Risk manager
Stewart Quinn	Environmental health manager
Henry Tan	Health science and environmental microbiology
Allison McNamara	Food policy manager
Stacey Kane	Zoonotic disease epidemiology
Tom Ross	Food Microbiology
Helen Withers	Food Microbiology

SEJ practices usually advise against monetary incentives, but to our knowledge not enough evidence exists to support definitive claims. Monetary incentives have been found both beneficial and detrimental (Lebreton et al., 2018) and more useful when participants do not have an intrinsic motivation (e.g., when students are used instead of professionals, Bojke L (2021)). The experts participating in the current elicitation were not offered monetary incentives.

We aimed for a diverse array of individuals, with diversity covering domain knowledge and demographics (i.e. age, experience, gender). These are proxies for cognitive diversity. The current advise for the optimum number of experts is between four and ten (Hanea et al., 2021).

Even though 13 experts participated in the inception meeting and completed the calibration questions, only 11 (marked in bold in Table 1) completed the first round of the target questions. Ten experts (marked in italic in Table 1) completed both rounds of the elicitation and the results presented in this report are based on aggregating the judgements of these ten experts.

3.3 Facilitators

Most SEJ elicitations featuring extensive discussion need facilitators. Ideally, two types of facilitators are present during the discussion, i.e., a normative and a substantive one. A normative facilitator is one experienced in subjective probabilities and SEJ research, and a substantive facilitator is one experienced in the application area. AH acted as the normative facilitator and BD acted as the substantive facilitator.

3.4 Materials

A Plain Language Statement and a Consent Form were prepared prior to first contacting the experts.

3.4.1 Plain Language Statement

A project statement written in plain language was developed and distributed to the experts in the first communications.

This is a brief document, containing key information on the voluntary aspect of the elicitation, what will be asked of the experts, how the data will be used and stored, how anonymity will be protected, and who should receive enquires and/or complains about the process.

3.4.2 Consent

A consent form allows the experts to sign their acknowledgement that they have read and understood the purpose of the elicitation and they will willingly take part.

3.4.3 Background Document

Often previous data, or data from an adjacent field may inform experts' estimates. Current trends, modelling and simulation exercises may inform the responses. Collecting all available info pertinent to the target questions is considered good practice. SS and BD developed the background document for this elicitation.

3.4.4 Questions

All questions asked are about quantities surrounded by uncertainty and measured on a continuous scale. To acknowledge the uncertainty, we elicit point estimates together with upper and lower bounds for all quantities in question. When answering a question we ask for a lower bound first, then for an upper bound, and only at the end we ask for a best estimate. Answering the questions in this order is meant to avoid anchoring on the best (central) estimate, and in this way reduce overconfidence (the temptation to give intervals which are too narrow). We can represent the uncertainty using a probability distribution, and we will consider the bounds to be the 5% and the 95% percentiles of a distributions, and the best estimate to be its median.

In this elicitation we asked 12 calibration questions and 72 target questions corresponding to pairs of pathogens and food commodities.

An example of questions' formulation and presentation is shown in Figs. 1 and 2.

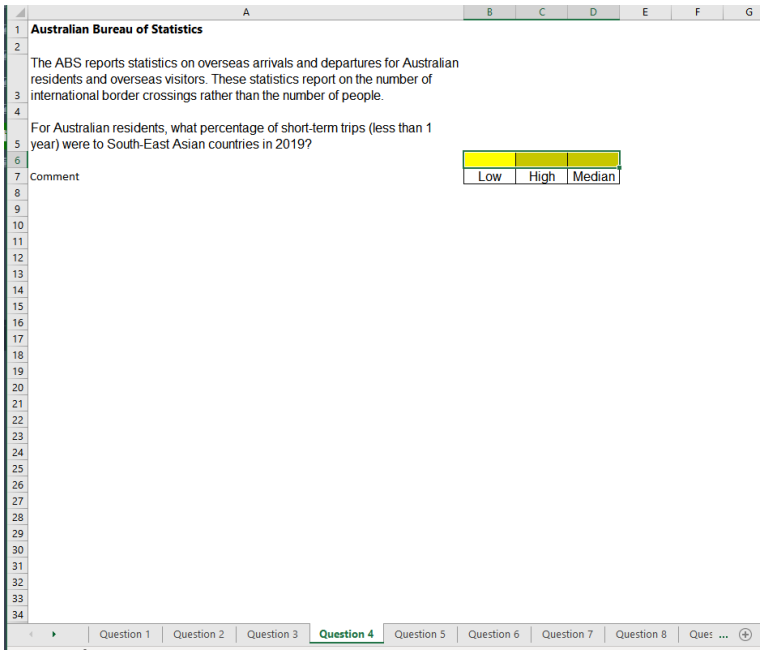
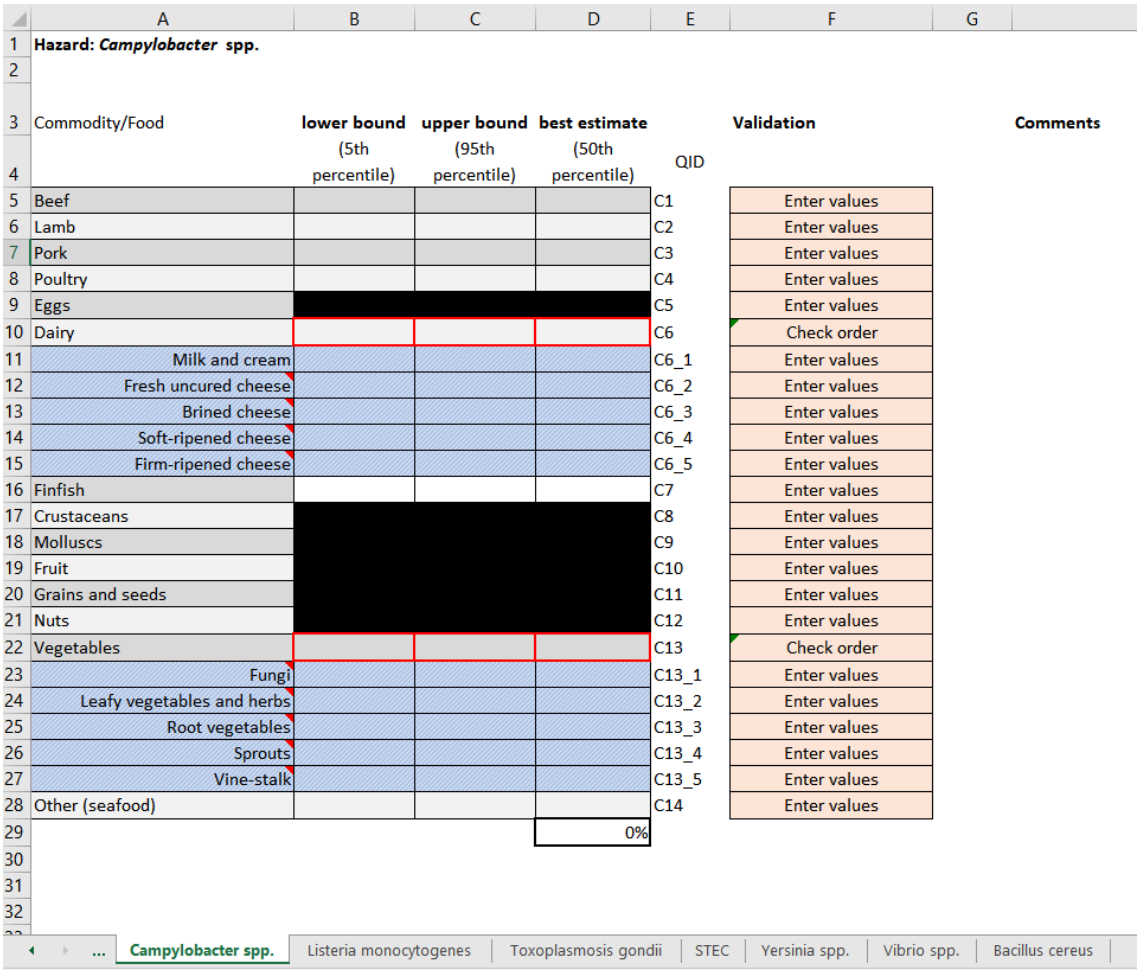


Figure 1: Example calibration question.

3.5 Elicited data

3.5.1 Round 1 estimates

Each expert answered the questions independently and sent the project team the filled-in excel spreadsheet. AH anonymised each spreadsheet by giving each expert an unique ID and compiled separate files per question. Each question was given an



	A	B	C	D	E	F	G
1	**Hazard: *Campylobacter* spp.**						
2							
3	Commodity/Food	**lower bound** (5th percentile)	**upper bound** (95th percentile)	**best estimate** (50th percentile)	QID	**Validation**	**Comments**
5	Beef				C1	Enter values	
6	Lamb				C2	Enter values	
7	Pork				C3	Enter values	
8	Poultry				C4	Enter values	
9	Eggs				C5	Enter values	
10	Dairy				C6	Check order	
11	Milk and cream				C6_1	Enter values	
12	Fresh uncured cheese				C6_2	Enter values	
13	Brined cheese				C6_3	Enter values	
14	Soft-ripened cheese				C6_4	Enter values	
15	Firm-ripened cheese				C6_5	Enter values	
16	Finfish				C7	Enter values	
17	Crustaceans				C8	Enter values	
18	Molluscs				C9	Enter values	
19	Fruit				C10	Enter values	
20	Grains and seeds				C11	Enter values	
21	Nuts				C12	Enter values	
22	Vegetables				C13	Check order	
23	Fungi				C13_1	Enter values	
24	Leafy vegetables and herbs				C13_2	Enter values	
25	Root vegetables				C13_3	Enter values	
26	Sprouts				C13_4	Enter values	
27	Vine-stalk				C13_5	Enter values	
28	Other (seafood)				C14	Enter values	
29				0%			

Figure 2: Example target questions.

unique ID as well (see column E from Fig. 2), for convenience. Fig. 3 shows the de-identified answers for the *Campylobacter* - Lamb pair.

	A	B	C	D	E	F
1	Exp ID	Q ID	Low	Median	High	Comment
2	Expert1	C2	5	6	25	eaten undercooked; higher prevelance
3	Expert2	C2	0	3	5	
4	Expert3	C2	0	3	5	
5	Expert6	C2	1	3	5	
6	Expert8	C2	2	5	10	
7	Expert9	C2	1	2	5	Based on NZ and ANU SAS
8	Expert10	C2	1	6	20	
9	Expert11	C2	0.1	1	1.5	
10	Expert12	C2	0	0	0	Not seen as FBI as foods generally cooked
11	Expert13	C2	1	5	7	
12	Expert14	C2	0.1	0.5	2	

Figure 3: Anonymised group estimates for the *Campylobacter* - Lamb pair.

3.5.2 Discussion and second round

Using the separate files per question, AH produced separate figures per question. These figures were used to guide discussion and the rationales behind the different estimates. Using the same example as in the previous section, Fig. 4 shows the spread of judgements about the *Campylobacter* - Lamb pair.

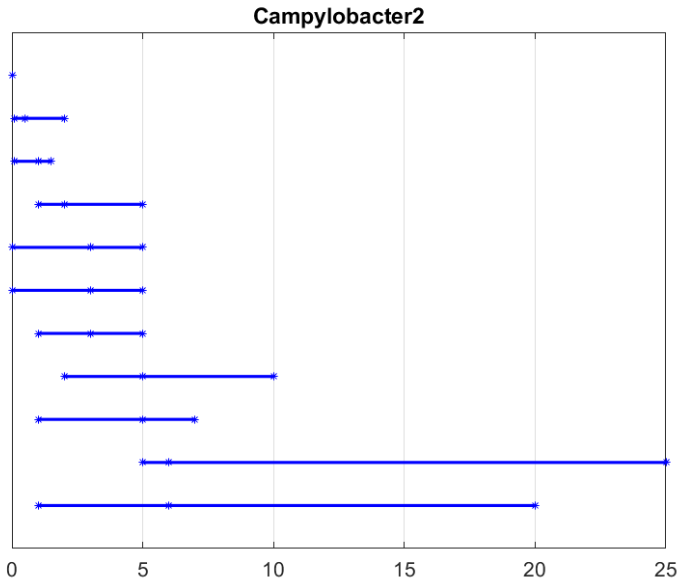


Figure 4: Anonymised group estimates for the pair *Campylobacter* - Lamb pair.

Each horizontal line corresponds to one expert. All medians are less than 10, with lower bounds reaching 0 and upper bounds reaching 25. Differences in the uncertainty around the medians can be also observed. All these differences were discussed and justified and after discussion experts reconsidered their estimates.

3.5.3 Aggregation

For each question, we considered the corresponding random variable, whose distribution is estimated per expert, using their second round estimates. An aggregated distribution per question is obtained by combining their distributions linearly with equal or differential weights.

Throughout the modelling, the best estimates correspond to the median of the subjective (personal) distribution assigned to the variable in question, the lower bounds corresponds to the 5% percentiles, and the upper bounds correspond to the 95% percentiles. Note that having the 5% and the 95% percentiles is not the same as having the physical bounds of the variable (hence, the support of the distribution). A $K\%$ overshoot method can be used, where the minimum (maximum) value is determined as the lowest (highest) percentile value $\pm K\%$ of the distance between the lowest and highest elicited values. The value of K is usually set to 0.1. The extension is symmetrical for simplicity. The physical bounds 0 and 100 are used when the calculated support extends beyond them.

A non parametric, minimum informative distribution is fitted on the three percentiles. This distribution is constructed by interpolating between expert's percentiles such that mass is assigned uniformly within the inter-percentile ranges (i.e., the four intervals determined by the assessed percentiles and the physical bounds). This creates a distribution that is piece-wise linear on the inter-percentile ranges, ensures that the three elicited percentiles correspond exactly to the elicited ones and does not add any extra information to what was elicited (e.g., no smoothness or symmetry).

The probability distributions constructed as above are averaged when equal weighting is used and the same three percentiles can be extracted to be used in further analysis. Fig. 5 shows the distributions for the C2 question constructed as above and equally weighted to obtain the linear pool.

For this elicitation we used the calibration questions to calculate calibration and informativeness scores. The table below summarises these scores for the experts and for two aggregations: the equally weighted aggregation (the average distribution) and the performance based weighted aggregation.

The performance based aggregation obtained a better combined score than the equally weighted aggregation and as a result, the distributions obtained with this aggregation scheme will be used in future calculations.

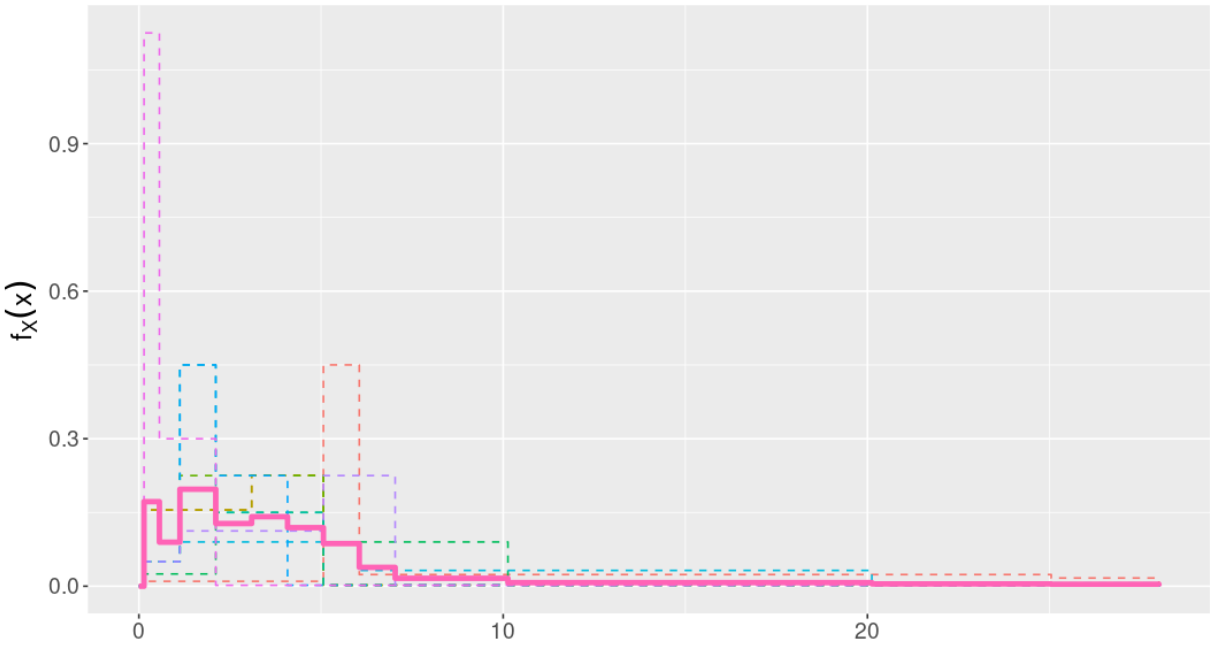


Figure 5: The non-parametric experts' fitted distributions of C2 (dashed) and their average (solid), the linear pool, which represents one version of the group's distribution for C2.

Table 2: Performance scores of experts and aggregations

Expert ID	Calibration	Informativeness
Expert 1	0.04576	0.5047
Expert 2	0.3131	0.7421
Expert 3	0.00628	0.9384
Expert 4	0.006589	1.281
Expert 6	0.06362	0.7568
Expert 8	0.1543	0.7806
Expert 9	0.001065	0.8475
Expert 10	0.0001036	0.9277
Expert 12	0.0001036	0.7767
Expert 13	0.6378	0.5701
Equal weights	0.3697	0.1766
Performance-based weights	0.3697	0.2729

References

Bojke L, Soares M, C. K. C. A. F. A. J. C. e. a. (2021). Developing a reference protocol for structured expert elicitation in health-care decision-making: a mixed-methods study. *Health Technol Assess*, 25(37).

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

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

Hanea, A., Hemming, V., and Nane, G. (2021). Uncertainty quantification with experts: Present status and research needs. *Risk Analysis*, n/a(n/a).

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

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

Hemming, V., Burgman, M. A., Hanea, A., McBride, M., and Wintle, B. (2017). A practical guide to structured expert elicitation using the IDEA protocol. *Methods in Ecology and Evolution*, 9(1):169–180.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86.

Lebreton, M., Slieker, M., Nooitgedacht, J., A.E., G., D., D., van Holst, R., and Luigjes, J. (2018). Two sides of the same coin: Monetary incentives concurrently improve and bias confidence judgments. *Sci Adv.*, 30;4(5).

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