Improving expert forecasts in reliability. Application and evidence for structured elicitation protocols.

Running Head: Application and evidence for structured expert elicitation

Victoria Hemming[1,2], Nicholas Armstrong[3], Mark A. Burgman[4,2], Anca M. Hanea[1,2].

[1] The Centre of Excellence for Biosecurity Risk Analysis, The University of Melbourne, Melbourne, Victoria, Australia. Twitter: @CEBRA_UOM

[2] The Centre for Environmental and Economic Research, The University of Melbourne, Melbourne, Victoria, Australia.

3 The Defence Science Technology Group. The Australian Department of Defence. Melbourne, Victoria, Australia. Twitter: @DefenceScience

[4] The Centre for Environmental Policy, Imperial College London, London, United Kingdom @IC_CEP

* Corresponding author: Victoria Hemming (vhemming@unimelb.edu.au, twitter @v_hemming, CEBRA, The University of Melbourne, Parkville, Australia 3010). ORCID ID: 0000-0003-3220-6161.

Words: 9,447.

Abstract

Quantitative expert judgements are used in reliability assessments to inform critically important decisions. Structured elicitation protocols have been advocated to improve expert judgements, yet their application in reliability is challenged by a lack of examples or evidence that they improve judgements.

This paper aims to overcome these barriers. We present a case study where two world-leading protocols, the IDEA protocol and the Classical Model were combined and applied by the Australian Department of Defence for a reliability assessment. We assess the practicality of the methods, and the extent to which they improve judgements.

The average expert was extremely overconfident, with 90% credible intervals containing the true realisation 36% of the time. However, steps contained in the protocols substantially improved judgements. In particular, an equal weighted aggregation of individual judgements, and the inclusion of a discussion phase and revised estimate helped to improve calibration, statistical accuracy and the Classical Model score. Further improvements in precision and information were made via performance weighted aggregation.

This paper provides useful insights into the application of structured elicitation protocols for reliability and the extent to which judgements are improved. The findings raise concerns about existing practices for utilising experts in reliability assessments and suggest greater adoption of structured protocols is warranted. We encourage the reliability community to further develop examples and insights.

Keywords: *Reliability, Expert elicitation, Defence, Procurement, Performance weighting.*

1 Introduction

Defence organisations procure equipment to enhance future capabilities. Despite substantial resource investment, poor procurement decisions are commonplace [1-3]. Such mistakes are costly, and they can leave Defence organisations unable to satisfy future capability challenges [4, 2, 3]. Improved forecasts of failures prior to purchase and operation would improve Defence agency decisions.

While hard data are invaluable, they are often incomplete or generated too late in the procurement process to inform decisions [5, 6]. Quantitative expert judgement is routinely required to support proactive decisions, including which and how many procurements to invest in, and the best means to mitigate potential equipment failures before they arise [6-9].

Structured elicitation protocols are advocated as a means to improve expert judgement in data poor contexts [10-15]. These protocols acknowledge that expert judgement is used as a form of scientific data contributing to important decisions [16]. Like other forms of data, it is affected by the methods used to gather it [17-20]. It is therefore important to apply methods which ensure the final judgements available to inform decisions and assessments are the best possible judgements.

Structured elicitation protocols do this by asking questions with clear operational meanings (for example, something that in principle would be measurable if time and resources would permit), eliciting quantitative judgements and uncertainty, anticipating and mitigating sources of bias, providing opportunities for validation, and transparently documenting the methods and judgements so that they are appraised critically and repeatable[21].

The fields of defence, engineering, and risk assessment were instrumental in the early development and promotion of structured elicitation protocols (for example, methods proposed by Keeney and von Winterfeldt [14], Cooke [22], Morgan et al. [23]). There has been some discussion as to how these protocols could be adapted and applied to reliability assessments [24, 8, 5, 25, 16, 26]. However, recent examples of their application and discussion of their advancement are few and far between [7, 27], suggesting that such protocols are not being widely applied in reliability assessments.

Examples differ in their advice as to 'best practice', and demonstrate structured elicitation protocols can entail a significant time and resource investment [14, 24]. Often, they fail to provide evidence about the extent to which investment in structured elicitation leads to improved judgements compared to simpler approaches (i.e. relying on one expert). This makes motivating and justifying their adoption difficult [9].

For procurement agencies such as defence organisations, accessible, practical and evidence-based examples are critical [9]. Often procurements are highly classified, and the ability to draw on external agencies to assist with an elicitation is limited. The iterative and frequent nature of reliability estimation for procurements often precludes deploying elaborate protocols. Thus, protocols may need to be applied in-house by those who have very little previous exposure to structured elicitation and at a modest budget.

In recent years Australian Department of Defence have sought to improve quantitative expert judgements through the application of structured elicitation protocols; however, their ability to do so was inhibited by a lack of appropriate examples, differences in proposed approaches, and the dearth of evidence to support their application. This predicament led to current research to understand how structured elicitation protocols could realistically be applied within reliability assessments, and the extent to which they improve judgements.

In this paper, we combine two world leading protocols, the IDEA protocol and the Classical Model and apply them to a procurement assessment by the Australian Department of Defence.

We outline key steps of their implementation, and validate the extent to which key steps improve judgements.

1.1 The IDEA protocol

IDEA stands for key steps, "Investigate", "Discuss", "Estimate", and "Aggregate" [28, 29]. Prescriptive advice for the protocol is outlined in Hemming et al. [30]; however, briefly, it involves the following steps:

- A diverse group of experts is recruited to answer questions with probabilistic or quantitative responses.
- The experts are asked to first *Investigate* the questions and to clarify their meanings, and then to provide their private, individual estimates with uncertainty.
- The experts receive feedback on their estimates in relation to other experts.
- With assistance of a facilitator, the experts are encouraged to *Discuss* the results, resolve different interpretations of the questions, cross-examine reasoning and evidence, and then provide a second and final private *Estimate*.
- The individual estimates are then combined using mathematical *Aggregation*, usually an equally weighted aggregation.

Applying these steps has been shown to substantially improved the accuracy of point estimates and the calibration of interval judgements [18, 31, 32]. Such judgements can be obtained from relatively small groups of five to nine experts using face-to-face workshops or remote elicitation [33, 28, 34].

In relation to existing structured elicitation protocols advocated in the engineering literature, IDEA is most similar to the elicitation approach outlined by Keeney and von Winterfeldt [14]. Steps which are shared include recruiting a diversity of experts, allowing the experts to investigate the problem and to provide their private individual estimates (quantitative and probabilistic judgements, with uncertainty), then bringing experts into a discussion, and allowing them to revise their estimates. The final representation of uncertainty is often an equal weighted aggregation of expert estimates.

The inclusion of an initial round of estimation, followed by a discussion, and a revised estimate means IDEA can be likened to a Delphi style elicitation [15, 18]. Delphi is not well-defined and encompasses a large array of approaches [35, 36]. However, key differences to many Delphi approaches are: 1) IDEA uses discussion between experts rather than feedback; 2) it only involves a single round of discussion followed by a single opportunity to revise judgements; and 3) consensus is not required.

1.2 The Classical Model

In the Classical Model [22], experts are asked a set of questions for which the answers can be obtained, to validate judgements (termed *seed* or *calibration* questions). These questions relate to the main questions of the elicitation (termed *target variables* or *questions of interest*). Experts are asked to specify quantiles of a continuous non-parametric distribution (usually 5^{th}, 50^{th}, and 95^{th}) for both calibration and questions of interest. Weights are based on their performance on the calibration questions using an asymptotically proper scoring rule. Those who perform well on the calibration questions are afforded more weight in the final aggregation for the questions of interest. While there is no guarantee that performance weights will outperform equal weights, a recent analysis of the Classical Model suggests it often does [37].

1.3 Motivation for this study

In 2015, the Australian Department of Defence began investigating how structured elicitation protocols could be used to improve expert judgements. At that time the expected attrition rates for a proposed aircraft procurement termed the "Skua" (a pseudonym) were required [39]. As noted by van Gelder et al. [39] the Australian Defence Force (ADF) had never flown these vehicles. Therefore, there were no historical data for their use in an Australian context. To forecast expected reliability, the Defence Science and Technology (DST) group first obtained historical data on Skua attrition rates as used elsewhere to construct a model of attrition due to a range of socio-environmental and technical factors including flight control, weather, organisational and cultural factors. Expert judgements were needed to adjust these data to reflect differences in how the Skua would be used by Australian forces.

As stated by Bedford et al. [6], typically such judgements are adjusted through multiplication factors; however, the methods employed are often not supported by clear and transparent expert judgement protocols or models. Other researchers have supported this contention indicating that in their experience judgements are often made heuristically, or represent the guesstimates of a single individual [3, 27, 9].

The DST Group chose to use the case study of the Skua to explore the application of structured expert elicitation protocols. Their initial aim was to improve the transparency and defensibility of the final judgements. They adopted the IDEA protocol to elicit judgements from experts which could be used to adjust parameters in the model [29, 28, 32, 30].

In their study on the Skua, van Gelder et al. [39] found that the IDEA protocol was practical to apply and improved the transparency and the information obtained to inform decisions. There were, however, aspects of the elicitation that van Gelder et al. [39] believed could be improved. Notably, the protocol used an equal weight aggregation (or the wisdom of the crowd [40]), which has been demonstrated usually to outperform the median-ranked individual (in terms of accuracy of point estimates and calibration of interval judgements), and often outperforms the best-credentialed expert, but assumes that all experts have equally valid knowledge and judgement [41, 33, 31]. van Gelder et al. [39] suggested that the scoring and aggregation rules of the Classical Model may further improve the final aggregations relative to equal weights [22].

The Classical Model has been cited in the reliability literature as a method requiring investigation [6, 3]. However, reservations have been expressed as to whether the method can be practically applied to reliability questions (i.e. it may be difficult to develop good calibration questions [3, 38]).

In 2017, an opportunity arose to revise the estimates from van Gelder et al. [39] regarding the Skua. DST decided to use this opportunity to explore how performance weighting from the Classical Model could be incorporated into the IDEA protocol to further improve judgements. The inclusion of calibration questions provided a unique opportunity to explore how the application of the IDEA protocol and performance weighting from the Classical Model may improve judgements.

The 2017 study forms the basis of this paper. We first clearly demonstrate how the two structured elicitation protocols were combined and applied to forecast reliability for the Skua. We then use the calibration questions to obtain insights into the improvements to the final judgements available to analysts and decision-makers via their application. Specifically:

- The recruitment and equal weighted aggregation of a diverse group of individuals relative to a single expert.
- The application of performance weights relative to equal weights.
- The inclusion of a second round of judgements

Our overarching aim is to help procurement agencies to understand the benefits derived from these protocols, how they might reasonably be applied, and where possible to avoid pitfalls encountered in their application.

2 Ethics

Human subjects research ethics approval was obtained under the DST Group ethics application for low risk projects: AD 02-17.

3 Methods

3.1 Preparation

Preparation for the elicitation began in February 2017, and largely followed advice in Hemming et al. [30]. We summarise key steps below.

3.1.1 The project team

DST personnel developed questions, motivated the experts, and organised the logistics for the workshop. Researchers at the University of Melbourne familiar with the IDEA protocol and the Classical Model provided high-level advice on these aspects until security clearances were granted, after which they were able to help facilitate the workshops and undertake the analysis of the data.

3.1.2 Elicitation format

The IDEA protocol involves an initial round of estimation (Round 1), followed by a feedback and discussion stage, and an opportunity to revise judgements (Round 2). It was decided that judgements would be elicited in Round 1 using remote elicitation. This would involve sending a spreadsheet containing the questions to experts via internal email. The feedback and discussion phase, and the collection of the revised private estimates (Round 2) would take place during a two-day workshop.

3.1.3 Question development

3.1.3.1 Number of questions

The elicitation consisted of two types of questions:

1) Questions of interest (used to update the model but for which answers cannot be obtained); and
2) Calibration questions (questions related to the types of knowledge experts need to have to accurately answer questions of interest).

A total of 32 questions were asked, enough for updating the model for the Skua project (17 questions), developing performance-based aggregations, cross-validating the predictive performance of the aggregations if desired (15 calibration questions), while avoiding fatigue in experts.

3.1.3.2 Determining the questions of interest

An email was sent to Defence personnel to determine if additional concerns had arisen since van Gelder et al. (39). A total of 52 replies were received. DST personnel overseeing the Skua elicitation considered that the existing 17 questions adequately addressed the concerns listed from the exercise.

3.1.3.3 Framing questions of interest

The same general format of the questions of interest which had been devised by van Gelder et al. (39) was used:

"The FAF (Foreign Air Force) and the ADF (Australian Defence Force) both operate the same size fleet of Skuas concurrently over the same life cycle. Suppose that the FAF loses X Skuas over their life cycle due to Y [e.g., aircrew inexperience]. All other factors being equal how many Skuas would the ADF lose due to Y [e.g. aircrew inexperience]".

It was used because the experts seemed to find the format intuitive and simple to use. The classes of *Y* were kept broad, for example difference in communications systems, or differences in the weather.

Each question was followed by the four-step question format [42]:

- Realistically, what is the lowest plausible estimate for the number of losses that Australia could experience as a result of differences in Y? _____
- Realistically what is the highest plausible estimate for the number of losses that Australia could experience as a result of the differences in Y? _______
- Realistically what your best estimate for the number of losses that Australia could experience due to of the differences in Y? _____
- How confident are you that your interval from lowest to highest will capture the realised truth (estimate between 50% and 100%). _____

This question format obtains a best estimate and upper and lower credible estimate (an interval judgement). The fourth question asks experts to estimate how sure they were that the interval they created contained the true value [42], this is used to standardise intervals to 90% credible intervals. This step reduces overconfidence relative to eliciting fixed intervals [42]. Its application across a range of domains, and via remote elicitation (where a facilitator is not present), suggests it is helpful for overcoming the persistent challenge faced in deriving quantitative judgements from experts [43, 18, 34].

As the judgements needed to be converted to continuous probability distributions for the Classical Model, the instructions explained to experts that the best estimate would be interpreted as a median. Experts were allowed to adjust their estimates in Round 2 if the adjustments did not accord with their true beliefs. Their judgements were subsequently used as 5^{th}, 50^{th} and 95^{th} quantiles of a nonparametric probability distribution.

3.1.3.4 Calibration Questions

Cooke et al. [44] state that it is impossible to give an effective procedure for generating meaningful calibration questions, however, they provide some basic advice, including:

- At least 10 calibration questions are required;
- Questions should reflect predictions related to the questions of interest. While questions can be developed based on adjacent knowledge, and / or past events these are believed to be less predictive of good judgement of future events in a domain; and
- Questions should relate to uncertainty, rather than general knowledge events, or established facts in a domain.

Due to the absence of new datasets which were not known to experts all calibration questions except for two related to past events. The remaining two questions related to future weather events and could be considered adjacent predictions.

Calibration questions were developed by staff at DST based on publicly available websites and peer-reviewed literature. They aimed to cover the diverse knowledge captured by the questions of interest including modes of failure attributed to human factors, engineering, and differences in weather patterns.

After security clearances for researchers at the University of Melbourne were obtained the questions could be reviewed. It was noted that the calibration questions requested ratios rather than quantities, while the questions of interest asked for frequencies. Ideally questions of interest and calibration questions should be in the same format. There were also only 10 calibration questions, which was suitable for developing scores and weighting experts but provided little power to distinguish between performance weight and equal weight aggregations in any subsequent cross-validation (which had been another aim of the project). Unfortunately, by the time this was noted the questions had already been sent to experts. To overcome these limitations, an additional five calibration questions were subsequently sent to experts (making a total of 15 calibration questions) in the same format as the questions of interest (frequencies).

3.1.4 Recruitment

The IDEA protocol places an emphasis on recruiting a diversity of individuals because, whereas expertise for judgements under uncertainty cannot be easily predicted *a-priori*, diverse groups with knowledge of the questions often form well-calibrated and accurate judgements [30].

Participants were recruited with diverse specialisations relevant to the procurement, including knowledge related to mechanical, electrical, structural engineering, and human and environmental factors. They were invited from at least two military bases of Australia, both whom had relevant but complimentary experience with the procurement. Diversity was also reflected in recruiting a diverse range of backgrounds and experience levels of participants, from principal engineers to new graduates.

In total, 113 people were invited to participate. Of those, 79 were men, and 34 were women. Of those invited, 21 males and two females agreed to take part in the elicitation.

Three of these experts withdrew in Round 2 due to urgent project commitments, leaving 20 participants. Funding was made available to hold two separate two-day workshops. Participants self-nominated to attend one of these two workshops based on their availability.

3.1.5 Demographic data

Demographic information was collected prior to the elicitation, including self-rating and years of experience relevant to the Skua. All except for one expert provided this information.

3.2 The Elicitation

3.2.1 Introductory meeting

An introductory meeting was held by project staff at DST Group offices. The purpose was to the introduce the intent of the elicitation, and outline assumptions made in the existing model for which the estimates were to inform.

3.2.2 Round 1 elicitation

Round 1 commenced with an email to experts containing a spreadsheet of the questions (17 questions of interest, and 15 calibration questions). The calibration questions were largely interspersed between the questions of interest. The email included instructions for completing the survey. Experts were provided two and half weeks to answer the questions in their own time.

Experts were told that when making estimates they could speak to anyone they liked and consult any literature they felt necessary. However, they were not to speak to one another about the elicitation prior to the discussion phase. They were also not to look up the references for the calibration questions.

3.2.3 Round 1 analysis

Data from Round 1 were cleaned and converted to quantiles of 90% non-parametric distributions, this entailed a linear extrapolation of expert judgements to 90% credible intervals, and minor adjustments to judgements to avoid zeros, the bounds of bounded ranges, and to ensure there was some uncertainty contained between expert's quantiles (Supporting

Information 1: Section 1). The standardised judgements from experts were entered into Excalibur (the program used to develop weights for the Classical Model [(45)]), as two case files, one for each workshop group. Excalibur was then used to generate an equal weight aggregation for each of the calibration questions, and each of the questions of interest. *RMarkdown* was used to create feedback documents for each of the workshops, which combined graphs and tables containing the standardised 90% credible intervals for each question into a word document. These feedback documents were sent to experts two days prior to each workshop.

It was revealed during conversations with the experts prior to the workshop that some participants had misread the instructions and accessed the links provided in the questionnaire to look up the answers to calibration questions (thinking this was additional material that they should consult rather than prohibited links). It was decided to remove these participants from the aggregations for those questions, but still involve them in the discussion round.

It was also apparent during the analysis that asking for ratios for some of the calibration questions had possibly confused some experts. The main confusion arose from whether they were being asked to estimate A (the unknown quantity) relative to B in some questions and B relative to A in other questions. For this reason, and because none of the questions of interest were asked as ratios, it was decided to provide feedback to experts first on their estimates as ratios and then demonstrate how these estimates translated into frequencies, requesting that they provide their Round 2 estimates as frequencies.

The analysis also revealed that two calibration questions when converted to frequencies were asking for the same quantity, therefore, one calibration question was removed (leaving only 14 calibration questions).

3.2.4 Discussion and Round 2 elicitation

At the start of each workshop, a short presentation was provided to motivate the experts. It reiterated the need for the elicitation, evidence underpinning each of the steps of the elicitation, and explained how the estimates may have changed through the analysis and how to interpret the results of the graphs and tables provided.

Experts were informed that those who had looked at the references had been removed from the feedback and aggregation for those particular calibration questions. Those individuals were asked to remain quiet during the discussion stage of those questions to avoid biasing the other experts. If experts felt they had been incorrectly removed, they were asked to alert the facilitator to this, so that this could be corrected in the final analysis.

For each question, the graphical output was displayed on a screen and the facilitator asked experts to consider the full range of opinions expressed, and reasons for these opinions (i.e. consider counterfactuals). To provide a record of the rationales that underpin the uncertainty of the final aggregations, a second facilitator documented the dialogue between experts.

For some questions of interest, it became clear that the question could take on multiple meanings. Additional clarification was sought to eliminate ambiguities before the second round of estimates commenced. For one question, two equally valid interpretations had been made. It was decided to split the question to capture both of these interpretations. Following discussion of each question, the experts were provided an opportunity to revise their estimates (Round 2).

3.2.5 Round 2 analysis

In the Round 2 analysis, data were cleaned as in Round 1, they were then imported into Excalibur. Experts who did not attend the workshops were removed from the analysis (i.e. three experts). Eleven experts who reviewed or knew at least one of the references were also removed from the analysis.

Due to the low number of experts from each workshop that had not looked at least one of the calibration questions (nine participants), the estimates from experts from both workshops were combined for the analysis in Round 2 (EW).

To compare the effect of removing the eleven experts (who looked at the calibration questions) on judgements for the questions of interest, a sixth aggregation was developed taking the equal weight aggregation of all 20 experts who took part in Round 1 and Round 2 (denoted as X.EW).

One question asked about the amount of rainfall for a particular weather station in regional Australia. The resolution to this question was not available at the time of analysis, therefore, an average of nearby weather stations was used instead. This assumption was checked with two of the experts and was deemed to be a reasonable compromise (with little variation between the weather stations).

3.2.6 Assessing judgements

In this study we assess changes in performance according to the performance measures of the Classical Model and the IDEA protocol. We outline the basics of these performance measures here (see Supporting Information 1: Section 2 for more detail).

3.2.6.1 The Classical Model

The Classical Model has two main performance measures which assess the ability of an expert to provide a well-calibrated and informative probability distribution:

- <u>*Statistical accuracy*</u> (often referred to as *calibration* and often denoted by 'C') assesses the ability of experts to answer according to the theoretical multinomial distribution $p = (0.05, 0.45, 0.45, 0.05)$. The actual proportion of realisations within each interquantile range for each expert e (or aggregation) is tallied to create a multinomial distribution for each expert: $s(e)=(Q_1, Q_2, Q_3, Q_4)$. The realised distribution $s(e)$ is then compared to the theoretical distribution p using the Kullback-Leibler (KL) divergence measure and a Chi-square test with three degrees of freedom. Statistical accuracy is the p-value of this test. Higher values indicate an expert's distribution more closely matches the theoretical distribution. A statistical accuracy below 0.05 is often used as a cut-off point at which an expert is considered statistically inaccurate (i.e. Bamber et al. [46], Colson and Cooke [37]).

- *Information* (often referred to as informativeness) under the Classical Model measures the degree to which the distribution supplied is concentrated and to which it deviates from a uniform or log-uniform distribution (which are considered the least informative distributions). It uses the Kullback-Leibler (KL) divergence measure which is scale invariant (Quigley et al. 2018). Information is calculated per question and does not depend on the realisation. The information of an expert is the average information taken across all calibration questions. Higher numbers represent distributions which show greater departure form a uniform or log-uniform distribution.

These two performance measures are combined to form the CM Score:

- CM Score: Statistical accuracy and information are often inversely related to one another [47, 37]. In the Classical Model this trade-off is negotiated by multiplying the two scores to obtain the CM Score.

3.2.6.2 IDEA performance measures

Decision-makers may seek to understand the accuracy of the best estimate, the calibration of interval judgements (i.e. avoiding surprises outside of the 90% credible intervals), and the precision of those interval judgements. This information is not directly provided by the performance measures of the Classical Model, and thus we also assess individual and aggregated judgements based on the performance measures from the IDEA protocol. These performance measures were only used to assess improvements in performance related to the aims of this paper, and not to develop weights or aggregations:

- *ALRE accuracy* assesses the difference between the prediction b (the expert's best estimate) and observed value x. It is measured using the average log ratio error (ALRE) of expert responses. The measure is a relative measure, scale invariant, and emphasises order of magnitude errors rather than linear errors. Smaller scores indicate more accurate responses. For any given question the log ratio scores have a maximum possible range of 0.31 (=$\log_{10}(2)$), which occurs when the true answer coincides with either the group minimum or group maximum.
- *Calibration* is the proportion of intervals provided by the experts containing the realised truth relative to their assigned confidence [48, 49]. For example, if we standardise the intervals of an expert to 90% confidence intervals then we expect that for 100 questions, 90 of the realisations will fall between their 5^{th} and 95^{th} quantiles. If they capture fewer realisations than this, they would be considered overconfident, if they capture more realisations they may be considered underconfident. The measure is an absolute measure and is scale invariant.
- *Precision* (usually termed informativeness, but distinguished here to avoid confusion with information from the Classical Model) in the IDEA protocol relates to the width of the credible intervals. It is a relative score which measures the proportion of variable's range (calculated from the minimum and maximum values of the pool of experts) that expert's credible interval captures. For each question the worst score an expert can get is '1' where their estimate is equivalent to the background range (i.e.

represents 100% of the range). Scores close to zero are better and indicate their interval judgements were more precise, a score of 0.25 for example would indicate on average an expert provided intervals which captured 25% of the background range provided for that question. A zero is only possible if the expert fails to provide any uncertainty for any question.

3.2.7 Weighting and aggregating judgements

The performance measures from the Classical Model are used to develop aggregations. There are five basic ways in which experts may be weighted and aggregated in the Classical Model:

- *Equal Weights (EW)*: The equal weight group aggregation is a linear pool of all expert distributions using the arithmetic mean of their distributions. All experts receive the same weight regardless of how well they perform on calibration questions. It can be calculated without calibration questions.
- *Global Weights (GW):* is calculated based on the combined statistical accuracy and information scores averaged across all calibration questions. Experts who performed better on the calibration questions are afforded more weight on the aggregations for the questions of interest than those who performed poorly.
- *Itemised Weights (IW):* uses the same statistical accuracy scores as Global Weights. However, the weight each expert is awarded will change per question because it considers the informativeness of the expert for each particular question of interest rather than the average calculated on all calibration questions. This often leads to a slightly more informative decision-maker on average than Global Weights.
- *Global Weights Optimised (GWO)* and *Item Weights Optimised (IWO)*: these performance measures are similar to their unoptimised variants described above (i.e. Global Weights and Item Weights). However, several weighted combinations are

created, each corresponding to a combination where the poorly calibrated experts are excluded sequentially, by successively raising the level at which an expert is considered statistically inaccurate from an alpha level $\alpha = 0$ [(47)]. The combination which achieves the highest score is considered to be the optimised combination.

One can create a set of pooled judgements for each question under each weighting scheme. These pooled judgements can then be scored on the calibration questions (in-sample validation). The aggregation of expert judgements which derives the highest CM Score on the calibration questions is usually taken as the preferred weighting when combining expert judgements on the questions of interest. If two aggregations result in the same statistical accuracy, that with a higher information score is preferred [(50)].

3.2.8 Analysis

We used the performance measures and aggregation methods of the Classical Model to develop five differentially weighted pooled aggregations (EW, GW, IW, GWO, IWO) for both Round 1 and Round 2. The aggregations contained the judgements of the nine experts who did not look at the calibration questions.

We then use the five performance measures outlined in Section 3.2.6, and the CM Score to examine the extent to which judgements were improved by:

- The recruitment of a diverse group of individuals relative to a single expert.
- The application of performance weights relative to equal weights.
- The inclusion of a second round of judgements.

The measures of precision, information, and accuracy are relative. Therefore, when assessing the performance of individuals and aggregations, a single file containing the estimates of experts and aggregations for both Round 1 and Round 2 was created. Excalibur was used to obtain the overall scores for statistical accuracy, information and the CM score for individuals and aggregations in Round 1 and Round 2. R was then used to score the performance of individuals and aggregations on calibration, precision and ALRE accuracy in Round 1 and Round 2.

As measures of ALRE and precision are relative, we occasionally discuss improvements in terms of percentage of the background range. For ALRE accuracy this range is 0.31 per question, while the range for precision is 1 per question.

We use boxplots to compare the performance of each individual and each aggregation method across each performance measure. The boxes represent the 25th, 50th and 75th percentiles of the scores of the individual experts for each measure. When 95% confidence intervals are provided they represent a non-parametric confidence interval around the median score (Supporting Information 1: Section 1).

4 Results

An example of the expert judgements for calibration questions and questions of interest is provided in Supporting Information 1: Section 3. These graphs also show that the removal of the 11 individuals does not appear to substantially change the equal weighted aggregation for the questions of interest. Removing these individuals did reduce the gender diversity of the group, with the nine remaining experts all being male. However, the group retained other forms of diversity, including technical specialisations, years of experience, age, and self-rated expertise.

4.1 Assessing improvements in judgement

In this section we focus on understanding how the steps of the IDEA protocol and the Classical Model improved the final judgements available to the decision maker on the 14 calibration questions (Figure 1).

4.1.1 Individual performance versus the group

In Round 1, only one expert was considered statistically accurate at the 0.05 level (Figure 1a). When viewed in terms of calibration, experts were extremely overconfident, with the average expert providing 90% credible intervals that only captured 36% of the realisations (5 out of 14) [i.e. a median score for individuals of 0.36, 95% CI: 0.29, 0.71].

We explored whether experts who did better on the calibration questions could be predicted based on their experience or self-rating (Supporting Information 1: Section 4). We note that our sample is small, and demographic data for the best performing expert in Round 1 was not supplied. However, there appears to be no overall correlation.

In contrast, the Equal Weights (EW) outperformed all individual experts in terms of the statistical accuracy score (a score of 0.53, Figure 1a, translating to a near perfect multinomial distribution Table 1), and calibration score (in which the 90% credible intervals contained the realised truth 93% of the time (a score of 0.93, Figure 1b)). It also outperformed all but one individual in terms of accuracy of the best estimate (an ALRE accuracy score of 0.08 compared to a median score by individuals of 0.10 [95%CI: 0.08, 0.11]) (Figure 1f).

All individuals obtained a higher information score (minimum score by an individual of 1.21 compared to 0.77 for EW, Figure 1c), and were on average twice as precise as EW (median score of 0.31 [95%CI: 0.22, 0.44], compared to EW, 0.63, Figure 1d). However, EW was better calibrated and more statistically accurate than individuals. In the Classical Model the trade-off between information and statistical accuracy is navigated via multiplication of the two performance measures to create the CM Score. The results show that the CM score for EW was much higher than any of the individuals (maximum score of any individual was 0.10, compared to 0.41 for Equal Weights) (Figure 1e), suggesting that overall EW outperformed individuals.

4.1.2 Performance weighting

Performance weighting on calibration questions achieved further improvements to the final judgements beyond that of EW. In Round 1, each of the performance weighted aggregations was able to outperform EW in terms of information (minimum score of 0.84 (Global Weights (GW) and the maximum score of 1.07 (Itemised Optimised Weights (IWO), compared to 0.77 (EW) (Figure 1c)). Performance weighted aggregations had a better precision value, outperforming EW by 8-18% of the background range.

Each of the performance weighted aggregations had a higher statistical accuracy score compared to EW (0.66 compared to 0.53) (Figure 1a). However, as can be seen from Table 1, and Figure 1b, the improvement did not translate to a difference in calibration of the 90% credible intervals (which were already perfectly calibrated), but rather an adjustment in where a realisation fell between the 90% credible intervals relative to the best estimate.

Performance weighting in the Classical Model does not seek to optimise the accuracy of the best estimate and made only minor improvements on this metric relative to EW (improvements of up to 0.01 in ALRE accuracy, or 3% of the ALRE range) (Figure 1f).

If measured according to the CM score, each of the performance weighted aggregations was deemed to be an improvement to that of EW (EW had a score of 0.409, compared to CM scores of ranging from 0.554 (GW)-0.707 (IWO) for performance weights).

4.1.3 Improvements in Round 2

4.1.3.1 Individuals

Round 2 largely improved the judgements of individuals. Eight of the nine experts improved their statistical accuracy (from 1.9e-05 in Round 1 to 6.4e-03 in Round 2 (Figure 1a)). However, six out of the nine experts were still considered statistical inaccurate. The eight experts substantially improved their calibration in Round 2 (a median improvement of 18% (0.18), or 2.5 additional realisations contained between their 90% credible intervals, per expert [min: 0.14 – max 0.29]). The best estimate of seven experts also came closer to the true realisation, with a median improvement in ALRE accuracy scores by 8% [min: 3% – max: 16%] of the possible range for ALRE accuracy (0.31). While the remaining two experts decreased their ALRE accuracy it was by a marginal amount of 1% and 2% of the possible range for the ALRE measure.

The improvement in statistical accuracy made by individuals appeared to come at the expense of information scores (median decrease in information per expert of 0.25 [min:0.11-max:0.50]). However, in terms of the precision of credible intervals there was no consistent difference. Five individuals become more precise (median change of 0.06 (or 6% of the variable range)), and four less precise (median change of 0.03 (or 3% of the variable range)) (Supporting Information 1: Section 5).

In terms of the CM score, there was an improvement of 0.03 on average for eight of the individuals [min: 2.95E-05 and max: 0.18], and a decrease in performance for one individual (by 0.01).

4.1.3.2 Equal weights

Round 2 had little effect on the performance of EW. The EW aggregation still outperformed all experts in terms of statistical accuracy (an increase of 0.13 to a score of 0.66) (Figure 1a). This was attributed to a change in the distribution of realisations between the 90% credible intervals rather than an improvement in calibration, for which it was already perfectly calibrated (Figure 1b and Table 1). There was a minor improvement in terms of ALRE accuracy of the best estimate (score of 0.07, an improvement of 3% of the background range relative to Round 1) (Figure 1f). The EW aggregation decreased in terms of information (by 0.034, Figure 1c), which translated to a small decrease in precision, on average by 4% of the range (Figure 1d). Overall, the CM score for EW improved slightly from 0.41 in Round 1 to 0.49 in Round 2 (Figure 1e).

4.1.3.3 Performance weights

The effect of Round 2 on performance weighting was less clear. In Round 2, IWO and IW achieved the highest (equivalent) CM Scores and both out-performed equal weights and each of the experts in term of the CM Score. However, each of the performance weights performed worse in Round 2 than in Round 1 in relation to the CM Score (decreases between 0.03-0.22). This could be attributed to a decrease in information (decreases between 0.06-0.13). These reductions equated to a reduction in precision by 1 – 5% of the background range. For two of the weights (GW and GWO), the statistical accuracy also decreased in Round 2 (by 0.132) making them lower than EW, however, this had no effect on their calibration (all performance weighted aggregations retained a calibration of 0.93).

4.1.3.4 Information and precision on Questions of Interest

The decreased precision and information of performance weighted aggregations in Round 2 made it difficult to understand which aggregation should be chosen (i.e. Round 1 or Round 2).

While the questions of interest cannot be validated with data, we could explore how the different aggregation approaches compared in terms of precision and information in Round 1 and Round 2 (Supporting Information 1: Section 6).

In Round 2, seven experts increased their precision with the median improvement in precision by 11% [min: 2%, max: 31%]). Equal Weights also improved in precision by 29%. Each of the performance weights also improved in Round 2, however, only two of the performance weights were able to improve upon equal weights (IW and IWO with an improvement in 20% precision relative to equal weights) (Supporting Information 1: Section 6).

Figure 1. Place holder

Table 1. Place holder

5 Discussion

Procurement agencies require expert judgement to inform large, complex and costly decisions. Structured elicitation protocols have been advocated to improve expert judgements [(14, 3)]; however, their adoption is hampered by a lack of evidence of their benefits and practical examples of implementing them. This study aimed to overcome these barriers. We applied two leading protocols [(15)], the IDEA protocol and the Classical Model to a real reliability assessment for the Australian Department of Defence. We assessed whether the additional time and effort entailed in their application leads to improved judgements.

5.1 Do structured elicitation protocols improve judgements?

In reliability, judgements used to inform decisions and assessments are often derived by a single expert and may be made heuristically [8, 3]. Our study revealed the serious risk such practices may pose for reliability assessments. However, key steps of structured elicitation protocols can improve the final judgements.

5.1.1 Equal weighted aggregation vs individual experts

In Round 1, individual experts were highly informative, however, experts provided 90% credible intervals which captured the true realisation on average only 36% of the time, and only one expert was considered statistically accurate at the 0.05 level (obtaining a calibration of 0.78). While our results are based on one case study, they closely match the average overconfidence levels found by Keeney and von Winterfeldt [14] and Mosleh et al. [51] in other engineering applications.

There is a perception that better performing experts can be selected or weighted based on their credentials [6, 52, 3, 22, 38]. We found little evidence to support this contention (Supporting Information 1: Section 4). While our sample size is small, and the best performing expert in Round 1 did not supply demographic data, there appears to be no overall trend in performance associated with years of experience or self-rating. These findings reflect a suite of studies which have failed to find a connection between good judgements under uncertainty and credentials [31, 18, 53].

In our study we examined the improvements in judgement that may be made by taking an equal weighted aggregation of a diverse group of individuals. We found that an equal weighted aggregation of the nine individuals (EW) was able to outperform all individuals in relation to calibration and statistical accuracy in both Round 1 and Round 2 (Figure 1a and 1b). In Round

1, EW outperformed all individuals in terms of the ALRE accuracy of the best estimate and performed as well as the median score attained by an individual for this measure in Round 2 (Figure. 1f).

The EW aggregation gave better calibrated and more statistically accurate estimates, however, this improvement came at the expense of information and precision. Informative and precise estimates are desirable, but decreases in these measures are not necessarily detrimental. If a final, less precise estimate is a truer representation of inherent variation and lack of knowledge, then the outcome is improved. Nonetheless a trade-off exists, making it difficult to understand whether an overall improvement was made. In the Classical Model this trade-off is navigated by multiplying statistical accuracy and information to form the CM Score. In our study, EW outperformed individual experts on the CM Score in both Round 1 and Round 2, suggesting an improvement was made.

The need to recruit more than one expert has been acknowledged as a means to improve judgements [24, 55, 27, 52]. Yet this step appears to be rarely undertaken. It is worth highlighting how easily this step was achieved. We used remote elicitation, whereby an email with a spreadsheet of questions was sent to experts with knowledge of the domain. We then aggregated their judgements via an equal weighted linear pool of distributions (EW). This step is only the first part of the IDEA protocol and should be achievable within the practical and financial constraints of most reliability studies, even if no other steps are taken.

5.1.2 Performance weighted aggregation

van Gelder et al. [39] and others [16, 6, 3] in the reliability literature have suggested that an equal weighted aggregation may be further improved by performance weighting using the Classical Model. We confirmed this speculation and found that performance weighting improved on equal weights largely via improvements in information and precision.

5.1.3 Discussion and revised judgements

Following Round 1, the IDEA protocol involves a subsequent discussion phase and an opportunity for experts to revise their judgement (Round 2). The potential advantages of discussion [14, 8, 24] and whether it improves or degrades the quality of judgements have been debated [10, 56, 32, 18, 57, 58]

In our results, Round 2 estimates following discussion led to a clear improvement to the majority of individual judgements in terms of calibration, statistical accuracy, and ALRE accuracy (Figures 1a, 1b, 1f). The improvements made by individuals helped to further improve EW in terms of statistical accuracy and ALRE accuracy (Figures 1a, 1f). Round 2 estimates were slightly less informative and precise judgements for individuals and EW (a reduction in precision by 4% of the background range, Figure 1d). However, the overall CM Score for these judgements improved (Figure 1c). When viewed in terms of the precision and information scores for the questions of interest, Round 2 improved EW and individuals (Supporting Information 1: Section 6).

The effect of Round 2 on performance weighted aggregation was less clear. In Round 2, each of the performance weighted aggregations improved in terms ALRE accuracy (Figure 1f), and retained the same perfect calibration score of 0.93. However, information (reductions in 0.06-0.13) and precision (reductions by 1-5%) decreased slightly, reducing their overall CM score (0.03-0.22) (Figures 1c, 1d, and 1e). For two of the performance weighted aggregations (GWO and GW) statistical accuracy scores also decreased (by 0.132). While this did not affect their calibration score, it meant they were ranked lower than EW in relation to statistical accuracy and their CM Score (Figure 1a and 1b). When we examined these findings in terms of changes in precision and information on the questions of interest, the inclusion of Round 2 improved the performance weighted aggregations on these measures (median improvement in precision of 0.37 [Min: 0.27, Max 0.47], and information of 0.42 [Min: 0.29, Max 0.79]).

5.1.4 Broader implications of these findings

Overall our study suggests that improvements could be made to the final judgements for reliability assessments simply by deploying key steps of the IDEA protocol. For instance:

1) an equal weighted aggregation (in our case study this was via linear pooling of distributions) will achieve more accurate, better calibrated, more statistically accurate estimates, and a higher overall CM Score than relying on a single individual; and

2) discussion and revised estimation can further improve individual and equal weight aggregated judgements on these measures.

In our case study, no individual outperformed equal weighted estimates on these measures. There also appeared to be no correlation between individual performance and experience or self-rating, suggesting that selecting experts based on these metrics is fraught.

We found that if time and resources are available, then performance weighting via the Classical Model can be used to further improve the judgements derived (beyond EW). In our study this was largely through improvements in precision and information.

5.2 Additional benefits

We can see additional benefits, aside from improvements in performance, which may further justify the application of structured elicitation protocols in reliability. The methods we examined are systematic and transparent. They enable critical appraisal and review of the steps to derive and aggregate the final judgements. This is important for decision-makers and procurement agencies who have to make decisions based on these judgements, applying to expert assessments the same transparent and repeatable methodologies that are expected for other forms of empirical data [11].

The two protocols derive quantitative judgements with uncertainty, which can be directly incorporated into models and decisions. This differs from many approaches in the reliability literature which ask experts to provide qualitative, or fuzzy, statements (i.e. “likely” or “highly likely”), or simply to list the evidence from which an analyst constructs a probability distribution [59, 17]. It can be impossible to interpret what each expert means by their qualitative statements[24], making meaning ambiguous (see Wallsten et al. [60] and Kent [61]).

The discussion and feedback stage of the IDEA protocol was beneficial in terms of documenting reasoning and evidence to support the judgements provided by experts. In this study, over 207 factors related to the procurement (which could lead to improved or worse attrition rates) were identified. To a large extent they justify the judgements and uncertainties of the expert’s judgements. These factors could be used to investigate the risks posed by differences between the ADF and the FAF. Methods to extract and use such factors provide exciting new sources of information for reliability [62-65].

We found that performance weighting could help to improve the final judgements provided to decision-makers. However, this is not guaranteed (as can be seen by the performance of GW and GWO in Round 2). Regardless, we believe calibration questions are advantageous as they provide empirical validation for the final aggregation that is missing from most elicitation exercises. In our study, this was especially important given that emphasis was placed on recruiting a diversity of knowledgeable individuals rather than the most senior or well-credentialed individuals.

5.3 Challenges and future directions

In applying the structured elicitation protocols, we encountered challenges that would have been useful to understand prior to an elicitation. We outline these here and propose solutions for overcoming them.

Anecdotally, we felt the face-to-face workshop led to more engaged discussion and more substantial rationales than obtained by remote elicitations (undertaken in other projects by the facilitators). However, we also felt that it was more challenging to manage personality types in a workshop. We suggest if workshops are undertaken, that they are facilitated to manage personality types and provide opportunities for less confident individuals to voice their beliefs.

While the inclusion of performance weighting from the Classical Model improved judgements, the development of calibration questions entailed significant effort by the DST Group. It also reduced the number of questions of interest which could be asked in a single elicitation. The difficulty in obtaining calibration questions was in part due to the lack of appropriate databases for which experts would not already know the answers. This may not normally be a problem in applications of the Classical Model, as the expert judgements are elicited using interviews and are not permitted to consult sources to inform their judgements.

However, in this study the use of remote elicitation meant we relied on experts avoiding sources for the calibration questions. Telling experts to avoid certain links was also a mistake, and 11 experts had to be removed. Some of these experts also stated that while they did not look at the answers, secondary sources had quoted the answers, precluding them from the study.

Including questions about future events is a possible solution but runs the risk that the calibration questions will not be resolved. This arose for one of our questions, and has been noted previously by others implementing the Classical Model [66].

In addition, the Classical Model performance measures are designed to assess continuous distributions. One of our questions related to count data, and the realisation was equivalent to the upper range for the question. We adjusted experts' estimates prior to them being entered into Excalibur to avoid the limits of bounded distributions. This adjustment meant that this realisation would always fall above the expert's 90% credible intervals. The inclusion of such questions may have altered the statistical accuracy of experts and aggregations. Such questions should be avoided, constraining the types of datasets which can be accessed to develop calibration questions. Alternatively, the robustness of weights and aggregations to these questions could be checked via sensitivity analysis in Excalibur and these questions could be removed prior to aggregation if necessary.

If a decision is sufficiently important to incorporate performance weighting, then it would be worthwhile to improve access to databases from which calibration questions could be developed. If this is not possible, then developing calibration questions will prove challenging. Eggstaff et al. [3] propose an approach to developing weights which takes advantage of the iterative nature of reliability assessments. This solution requires further investigation.

We noted differences between the change in information for the calibration questions compared to the questions of interest between Round 1 and Round 2. Experts became less informative and precise on the calibration questions and more precise on the questions of interest (when scored using the IDEA protocol performance measures). We believe this reflects differences in the question framing between calibration questions and questions of interest. The questions of interest anchored experts on X aircraft being lost and asked for a relative change. Experts often conveyed their estimates as integers, not realising that the increase or decrease in attrition by Y aircraft corresponded to a change of Z% of losses in each direction. When the estimates where extrapolated from their assigned confidence (e.g. 60%) to 90% credible intervals their estimates changed substantially and became uninformative. The effect may not have been as severe for questions which used an absolute frequency format or ratios, and thus experts are less likely to have adjusted their intervals as much in Round 2 for these questions. This reiterates the need for the calibration questions and the questions of interest to be in the same format.

For question wording, we used the format provided in van Gelder et al. [(39)]. Many experts did not have a problem with the format *per-se*, but suggested a better phrasing for aviation may be to ask experts for the attrition rate per 100,000 hours of operation.

We found that the performance measures of the Classical Model were not easy to interpret in relation to a final decision. For example, to achieve the highest statistical accuracy possible on 14 questions, the experts and aggregations would need to provide intervals which contain 12 out of 14 realisations. However, the aggregations in this study actually captured more than this (13 out of 14 realisations), for which they were penalised by reducing their statistical accuracy score by 28.1%. This difference is due to the way in which the Classical Model scores multinomial distributions. This may be counterintuitive for many decision-makers who seek to avoid surprises outside of their 90% credible intervals (i.e. would prefer aggregations that capture 9/10 realisations between the 90% credible intervals over 8/10 realisations), and obtain more precise uncertainty bounds.

We found that understanding can be improved by accompanying scores with their multinomial distributions (see Hemming et al. [(67)] for code), and utilising the performance measures of the IDEA protocol to convey this information (Supporting Information 1: Section 2).

The Classical Model aims to achieve rational consensus, that is, an agreement from the outset as to how a consensus distribution should be achieved [(68, 22)]. A limitation of this study is that key decision-makers and experts were not asked prior to the elicitation which judgement attributes they most wanted to reward, and perhaps did not understand the reward structure of the Classical Model prior to implementation. As no aggregation outperformed all others across all performance measures, this made choosing the best aggregation difficult. We suggest that prior to an elicitation, the key decision-makers and experts agree on the calibration questions and discuss the aspects of good judgement they most wish to reward, as well as the trade-offs they are willing to make in terms of alternative performance measures.

In our study we only elicited a small number of parameters (17 questions of interest and 15 calibration questions). This was sufficient for our case study but may be less than required for most procurement projects. It is possible to elicit many more parameters by recruiting more experts (and expanding the definition of an expert), and/or allowing more time to assess the suite of parameters [(34)]. In any case, decision-makers should take steps to focus an elicitation on the most important / influential parameters to their decision (i.e. via a sensitivity analysis [(69)]).

6 Conclusion

Expert judgement continues to be required in reliability to inform critically important decisions. The need to adopt more rigorous approaches to the collection of expert judgement has long been echoed, but practical and evidence-based examples have been lacking to support their widespread application. Our study was developed in response to this need. It provides an empirically validated example and evidence for the improvements that can be achieved via structured elicitation protocols. In deciding when to adopt the approaches outlined, we distilled the improvements made by each of the steps implemented in the IDEA protocol and the Classical Model, and compared improvements across a range of performance measures. However, we echo the sentiments of Eggstaff et al. [(3)], if the decision is considered important enough to require expert consultation, then it is probably important enough to consult several experts. We suggest this should be undertaken in a structured and empirically validated manner. The results of this study motivate wider consideration and investigation of structured elicitation protocols for improved reliability assessments.

7 Data Availability

Author elects to not share data: Data and code used to generate the results relate to judgements of National security and as such are classified by the Australian Department of Defence. As such they cannot unfortunately be shared.

8 Acknowledgements

The authors would like to acknowledge Ross Antoniou, Geoff Stuart, Andrew Goodwin, Sonya Slater, and Joanna Kappas from the Australian Department of Defence for their assistance and feedback regarding the expert elicitation project. The authors would also like to thank members from CEER and CEBRA at the University of Melbourne who helped to refine the design of the elicitation questions including Dr Janet Carey who helped to facilitate the workshop, and Prof Andrew Robinson. Finally, we would like to thank each of the experts who were involved in the expert elicitation workshops for their willingness to address the requirements and provide their judgements and advice for the project.

9 Declaration of interests

The study was funded by the Australian Department of Defence. VH received funding to draft this publication by the Australian Research Training Program, and the David Lachlan Hay

Memorial Fund. VH and AH are employed by the Australian Centre of Excellence for Biosecurity Risk Analysis and the Centre for Environmental and Economic Research. NA is employed by the Defence Science Technology Group. MB is employed by Centre for Environmental Policy, Imperial College London.

10 Supporting Information

Supporting Information 1: Data analysis and equations to support paper.

11 References

1. Woolner D. Why australia's defence procurement is lacking military precision. The Conversation 2011.
2. Kinnaird M, Early L, Schofield B. Defence procurement review 2003. 2003.
3. Eggstaff JW, Mazzuchi TA, Sarkani S. The development of progress plans using a performance-based expert judgment model to assess technical performance and risk. *Systems Engineering*, 2014; **17** (4):375-91.
4. Department of Defence. Defence white paper. Canberra, Australia.
5. Walls L, Quigley J, Marshall J. Modeling to support reliability enhancement during product development with applications in the u.K. Aerospace industry. *IEEE Transactions on Engineering Management*, 2006; **53** (2):263-74. 10.1109/TEM.2006.872342
6. Bedford T, Quigley J, Walls L. Expert elicitation for reliable system design. *Statistical Science*, 2006:428-50.
7. Revie M, Bedford T, Walls L. Supporting reliability decisions during defense procurement using a bayes linear methodology. *IEEE Transactions on Engineering Management*, 2011; **58** (4):662-73. 10.1109/TEM.2011.2131655
8. Hodge R, Evans M, Marshall J et al. Eliciting engineering knowledge about reliability during design-lessons learnt from implementation. *Quality and Reliability Engineering International*, 2001; **17** (3):169-79.
9. Hokstada P, Øien K, Reinertsen R. Recommendations on the use of expert judgment in safety and reliability engineering studies. Two offshore case studies. *Reliability Engineering & System Safety*, 1998; **61** (1):65-76. https://doi.org/10.1016/S0951-8320(97)00084-7
10. Morgan MG. Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences*, 2014; **111** (20):7176-84.

11. French S. Expert judgment, meta-analysis, and participatory risk analysis. *Decision Analysis*, 2012; **9** (2):119-27. doi:10.1287/deca.1120.0234
12. Aspinall WP. A route to more tractable expert advice. *Nature*, 2010; **463** (7279):294-5.
13. Sutherland WJ, Burgman M. Policy advice: Use experts wisely. *Nature*, 2015; **526** (7573):317-8.
14. Keeney RL, von Winterfeldt D. Eliciting probabilities from experts in complex technical problems. *IEEE Transactions on engineering management*, 1991; **38** (3):191-201.
15. O'Hagan A. Expert knowledge elicitation: Subjective but scientific. *The American Statistician*, 2019; **73** (sup1):69-81.
16. Cooke RM. Special issue on expert judgment. *Reliability Engineering & System Safety*, 2008; **93** (5):655-6. https://doi.org/10.1016/j.ress.2007.03.001
17. Booker JM, McNamara LA. Solving black box computation problems using expert knowledge theory and methods. *Reliability Engineering & System Safety*, 2004; **85** (1-3):331-40.
18. Hemming V, Walshe TV, Hanea AM et al. Eliciting improved quantitative judgements using the idea protocol: A case study in natural resource management. *PLOS ONE*, 2018; **13** (6):e0198468. 10.1371/journal.pone.0198468
19. Tversky A, Kahneman D. Judgment under uncertainty: Heuristics and biases. *science*, 1974; **185** (4157):1124-31.
20. Gigerenzer G. How to make cognitive illusions disappear: Beyond "heuristics and biases". *European review of social psychology*, 1991; **2** (1):83-115.
21. Hanea A, McBride M, Burgman M et al. The value of performance weights and discussion in aggregated expert judgments. *Risk Analysis*, 2018; **38** (6). 10.1111/risa.12992
22. Cooke RM. *Experts in uncertainty: Opinion and subjective probability in science*. K Sharader-Frechette, editor. New York: Oxford University Press; 1991. (K Sharader-Frechette editor). 321
23. Morgan MG, Henrion M. Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis cambridge university press. *New York, NY, United States of America*, 1990.
24. Walls L, Quigley J. Building prior distributions to support bayesian reliability growth modelling using expert judgement. *Reliability engineering & system safety*, 2001; **74** (2):117-28.
25. Sigurdsson J, Walls L, Quigley J. Bayesian belief nets for managing expert judgement and modelling reliability. *Quality and Reliability Engineering International*, 2001; **17** (3):181-90.
26. Forester J, Bley D, Cooper S et al. Expert elicitation approach for performing atheana quantification. *Reliability Engineering & System Safety*, 2004; **83** (2):207-20.
27. Revie M, Bedford T, Walls L. Evaluation of elicitation methods to quantify bayes linear models. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 2010; **224** (4):322-32.
28. Hanea A, McBride M, Burgman M et al. $\mathrm{I}_{\mathrm{nvestigate}}\mathrm{d}_{\mathrm{iscuss}}\mathrm{e}_{\mathrm{stimate}}\mathrm{a}_{\mathrm{ggregate}}$ for structured expert judgement. *International Journal of Forecasting*, 2016; **33** (1):267-9.
29. Burgman MA. *Trusting judgements: How to get the best out of experts*. Cambridge, United Kingdom: Cambridge University Press; 2015. 203 p.
30. Hemming V, Burgman MA, Hanea AM et al. A practical guide to structured expert elicitation using the idea protocol. *Methods in Ecology and Evolution*, 2018; **9**:169-81. 10.1111/2041-210X.12857
31. Burgman MA, McBride M, Ashton R et al. Expert status and performance. *PLoS One*, 2011; **6** (7):1-7.
32. Hanea AM, Burgman M, Hemming V. Idea for uncertainty quantification. In: LC Dias; A Morton; J Quigley, editors *Elicitation: The science and art of structuring judgement*. Cham, Switzerland: Springer International Publishing; 2018; p. 95-117.

33. Hemming V, Walshe T, Burgman M et al. Eliciting improved quantitative judgements using the idea protocol: A case study in natural resource management. *PLoS One*, In Review.
34. McBride MF, Garnett ST, Szabo JK et al. Structured elicitation of expert judgments for threatened species assessment: A case study on a continental scale using email. *Methods in Ecology and Evolution*, 2012; **3** (5):906-20.
35. Mukherjee N, Huge J, Sutherland WJ et al. The delphi technique in ecology and biological conservation: Applications and guidelines. *Methods in Ecology and Evolution*, 2015; **6** (9):1097-109.
36. Linstone HA, Turoff M. The delphi method. *Techniques and applications*, 2002; **53**.
37. Colson AR, Cooke RM. Cross validation for the classical model of structured expert judgment. *Reliability Engineering & System Safety*, 2017; **163**:109-20. http://dx.doi.org/10.1016/j.ress.2017.02.003
38. Clemen RT. Comment on cooke's classical method. *Reliability Engineering & System Safety*, 2008; **93** (5):760-5.
39. van Gelder T, Vodicka R, Armstrong N. Augmenting expert elicitation with structured visual deliberation. *Asia & the Pacific Policy Studies*, 2016; **3** (3):378-88. 10.1002/app5.145
40. Surowiecki J. *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. London, United Kingdom: Little, Brown; 2004. xxi-295
41. Budescu DV, Chen E. Identifying expertise to extract the wisdom of crowds. *Management Science*, 2014; **61** (2):267-80.
42. Speirs-Bridge A, Fidler F, McBride M et al. Reducing overconfidence in the interval judgments of experts. *Risk Analysis*, 2010; **30** (3):512-23. 10.1111/j.1539-6924.2009.01337.x
43. Hudson EG, Brookes VJ, Ward MP. Assessing the risk of a canine rabies incursion in northern australia. *Frontiers in Veterinary Science*, 2017; **4**:141. 10.3389/fvets.2017.00141
44. Cooke RM, Goossens LL. Tu delft expert judgment data base. *Reliability Engineering & System Safety*, 2008; **93** (5):657-74.
45. Lightwist. Excalibur http://www.lighttwist.net/wp/excalibur.
46. Bamber J, Aspinall W, Cooke R. A commentary on “how to interpret expert judgment assessments of twenty-first century sea-level rise” by hylke de vries and roderik sw van de wal. *Climatic Change*, 2016; **137** (3):321-8. 10.1007/s10584-016-1672-7
47. Quigley J, Colson A, Aspinall W et al. Elicitation in the classical model. In: LC Dias; A Morton; J Quigley, editors *Elicitation: The science and art of structuring judgement*. Cham, Switzerland: Springer International Publishing; 2018; p. 15-36.
48. Lichtenstein S, Fischhoff B, Phillips LD. Calibration of probabilities: The state of the art. In: *Decision making and change in human affairs*: Springer; 1977; p. 275-324.
49. Lin S-W, Bier VM. A study of expert overconfidence. *Reliability Engineering & System Safety*, 2008; **93** (5):711-21. http://dx.doi.org/10.1016/j.ress.2007.03.014
50. Bedford T, Cooke RM. *Mathematical tools for probabilistic risk analysis*. Cambridge, United Kingdom: Cambridge University Press; 2001.
51. Mosleh A, Bier VM, Apostolakis G. A critique of current practice for the use of expert opinions in probabilistic risk assessment. *Reliability Engineering & System Safety*, 1988; **20** (1):63-85. https://doi.org/10.1016/0951-8320(88)90006-3
52. Lewis TL, Mazzuchi T, Sarkani S. A statistical approach to the development of progress plans utilizing bayesian methods and expert judgmentA publication of the Defense Acquisition University.
53. Tetlock P, Gardner D. *Superforecasting: The art and science of prediction*. New York: Random House; 2015. 340 p.
54. Mannes AE, Soll JB, Larrick RP. The wisdom of select crowds. *Journal of personality and social psychology*, 2014; **107** (2):276.
55. Ale B, Bellamy L, Cooke R et al. Causal model for air transport safety. *Final Report, July*, 2008; **31**.

56. Lorenz J, Rauhut H, Schweitzer F et al. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 2011; **108** (22):9020-5.
57. Sarewitz D. How science makes environmental controversies worse. *Environmental science & policy*, 2004; **7** (5):385-403.
58. Merkhofer MW. Quantifying judgmental uncertainty: Methodology, experiences, and insights. *IEEE Transactions on Systems, Man, and Cybernetics*, 1987; **17** (5):741-52.
59. Warwick Manufacturing Group. Section 12: Product excellence using six sigma: Failure modes, effects & criticality analysis. Conventry, United Kingdom: School of Engineering, University of Warwick.
60. Wallsten TS, Budescu DV, Rapoport A et al. Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General*, 1986; **115** (4):348.
61. Kent S. Words of estimative probability. *Studies in Intelligence*, 1964.
62. Sutcliffe A, Sawyer P. *Requirements elicitation: Towards the unknown unknowns*. In Requirements Engineering Conference (RE), 2013 21st IEEE International: IEEE; 2013. 92-104 p.
63. Kaiya H, Saeki M. *Using domain ontology as domain knowledge for requirements elicitation*. In Requirements Engineering, 14th IEEE International Conference: IEEE; 2006. 189-98 p.
64. Schwartz HA, Giorgi S, Kern ML et al. More evidence that twitter language predicts heart disease: A response and replication. 2018.
65. Curtis B, Giorgi S, Buffone AE et al. Can twitter be used to predict county excessive alcohol consumption rates? *PloS one*, 2018; **13** (4):e0194290.
66. Wittmann ME, Cooke RM, Rothlisberger JD et al. Use of structured expert judgment to forecast invasions by bighead and silver carp in lake erie. *Conservation Biology*, 2015; **29** (1):187-97. 10.1111/cobi.12369
67. Hemming V, Lane S, Hanea A. Classical model calculator. In., Series Classical model calculator. The Open Science Framework; 2019. DOI 10.17605/OSF.IO/BGYZU.
68. Cooke R, Mendel M, Thijs W. Calibration and information in expert resolution; a classical approach. *Automatica*, 1988; **24** (1):87-93.
69. Spetzler CS, Stael von Holstein C-AS. Exceptional paper—probability encoding in decision analysis. *Management science*, 1975; **22** (3):340-58.

12 Tables and Figures

Table 1 The multinomial distributions for statistical accuracy scores of experts and aggregations.

Description	Q1	Q2	Q3	Q4	Statistical Accuracy (SA)
Highest SA possible	**1**	6	6	**1**	0.968
Highest SA for aggregations	**0**	7	6	**1**	0.659
Lowest SA for aggregations	**0**	8	5	**1**	0.527
Highest SA for an expert	**1**	6	4	**3**	0.156
0.05 threshold	**2**	8	2	**2**	0.054
Lowest SA for an expert	**4**	2	1	**7**	2.95E-08
Lowest SA possible	**0**	0	0	**14**	0

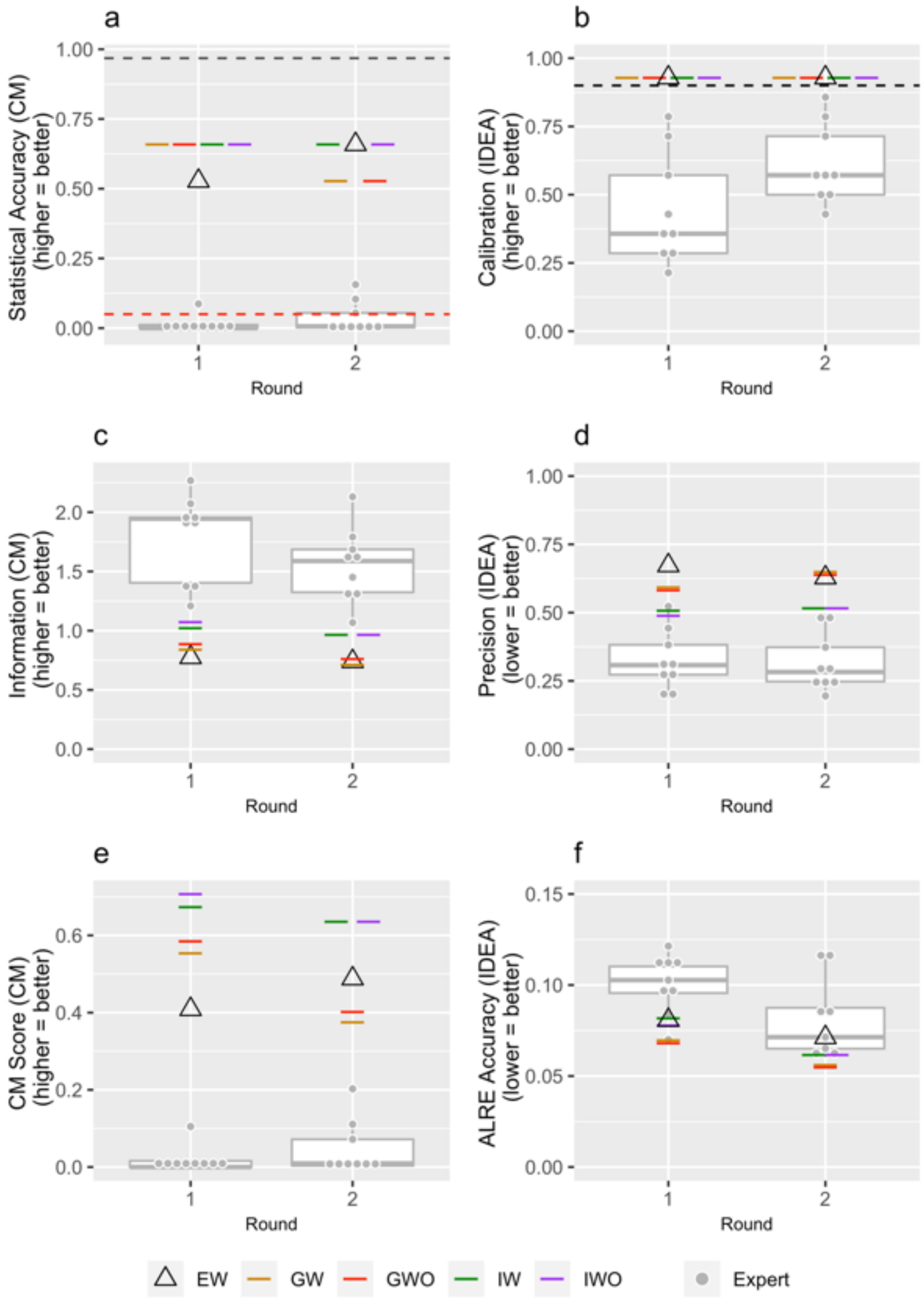
a
Statistical Accuracy (CM)
(higher = better)
Round
b
Calibration (IDEA)
(higher = better)
Round
c
Information (CM)
(higher = better)
Round
d
Precision (IDEA)
(lower = better)
Round
e
CM Score (CM)
(higher = better)
Round
f
ALRE Accuracy (IDEA)
(lower = better)
Round
EW
GW
GWO
IW
IWO
Expert

Figure 1. Scores for Classical Model (CM), and the IDEA protocol, for individual experts and aggregations, in Round 1 and Round 2 on the 14 calibration questions. The boxplots represent the median and interquartile ranges (IQR) of the individual experts. In Figure 1a the red dashed line represents a statistical accuracy level of 0.05, dots below the line are considered statistically inaccurate. In Figure 1a the black dashed lined represents perfect statistical accuracy (CM), and in 1b it represents perfect calibration (IDEA) on 14 questions. The accuracy score extends from 0 (most informative) to 0.31 (least informative), to show the differences between Round 1 and Round 2, the graph represents half of the possible scale. Key: EW= Equal Weights, IT=Item Weights, IWO= Item Weights Optimised, GW=Global Weights, GWO=Global Weights Optimised.