Calibrating experts' probabilistic assessments for improved probabilistic predictions

Hanea, A.M.; Nane, G.F.

DOI
10.1016/j.ssci.2019.05.048

Publication date
2019

Document Version
Accepted author manuscript

Published in
Safety Science

Calibrating experts' probabilistic assessments for improved probabilistic predictions

Abstract

Expert judgement is routinely required to inform critically important decisions. While expert judgement can be remarkably useful when data are absent, it can be easily influenced by contextual biases which can lead to poor judgements and subsequently poor decisions. Structured elicitation protocols aim to: 1) guard against biases and provide better (aggregated) judgements, and 2) subject expert judgements to the same level of scrutiny as is expected for empirical data. The latter ensures that if judgements are to be used as data, they are subject to the scientific principles of review, critical appraisal, and repeatability. Objectively evaluating the quality of expert data and validating expert judgements are other essential elements. Considerable research suggests that the performance of experts should be evaluated by scoring experts on questions related to the elicitation questions, whose answers are known a-priori. Experts who can provide accurate, well-calibrated and informative judgements should receive more weight in a final aggregation of judgements. This is referred to as performance-weighting in the mathematical aggregation of multiple judgements. The weights depend on the chosen measures of performance. We are yet to understand the best methods to aggregate judgements, how well such aggregations perform out of sample, or the costs involved, as well as the benefits of the various approaches. In this paper we propose and explore a new measure of experts' calibration. A sizeable data set containing predictions for outcomes of geopolitical events is used to investigate the properties of this calibration measure when compared to other, well established measures.

Keywords: structured expert judgement, performance based weighting, calibration, probabilistic predictions

1 Introduction

Experts are consulted in a myriad of situations and inform all stages of the modelling process, from structuring the problem to estimating facts, and quantifying uncertainty. While consulting experts may be a valuable resource for decision-makers, it is crucial that decision-makers, stakeholders and experts play separate roles in the decision process. The experts' role should be limited to providing estimates of facts and predictions of event outcomes [25].

Expert judgement is used when empirical data are unavailable, incomplete, uninformative, or

conflicting. These judgements then inform critically important decisions. It is therefore important that such judgements are as defensible as possible. Research into the experts' performance when providing such judgements reveals that expert status is not correlated with the ability of an expert to give unbiased, error-free judgements [5]. Expert judgements are susceptible to a range of cognitive and motivational biases, to the expert's particular context, and to their personal experiences [23, 24, 17]. To counter such limitations, structured expert elicitation protocols have been developed.

A working definition of a structured protocol is given in [8] and slightly reformulated in [11]. It involves asking questions with operational meanings, following a traceable, repeatable and open to review process, mitigating biases and providing opportunities for empirical evaluation and validation. The above are guidelines rather than rules for what would one consider to be a structured and efficient protocol.

While expert interaction and the feedback provided are still discussed within the expert judgement community, one of the elements that everybody agrees with is that the judgements of more than one expert are essential in all situations. A diversity of opinions is always desirable and we use gender, age, experience, affiliation, and world view as proxies for diversity.

Here, we restrict our attention to eliciting expert judgements of event occurrences. Eliciting multiple judgements about the same event results in sets of probabilities, rather than one single probability of that event occurrence which is often what is needed in further modelling. It is not uncommon for these probabilities to differ, reflecting different knowledge bases and different mental models used by the experts when making their judgements. While these differences are crucial to the understanding of the problem in all its complexity and need to be recorded, they make the aggregation of different judgements (into a single one) somewhat cumbersome.

The two main ways of aggregating different judgements are behavioural aggregation [e.g. 18], and mathematical aggregation [e.g. 27, 8]. Behavioural aggregation typically involves face-to-face meetings of experts who, at the end of the meetings, agree on a judgement. Discussion and consensus seeking are often practised together. The main advantage of this approach is that experts share and debate their knowledge. Nevertheless such interaction is prone to biases including groupthink and halo effects [e.g. 13]. Sometimes experts do not agree to any possible consensus. When experts

disagree, even after a facilitated discussion, attempts to impose consensus may mask the group's diversity of opinion.

The alternative is to use a mathematical rule to aggregate the judgements. When mathematical aggregation is used, the interaction between experts is usually limited to training and briefing. Extensive discussion is discouraged because it may induce dependence between the elicited judgements [e.g. 19]. Very few studies have been undertaken in order to investigate this effect, and even fewer found it harmful to the process [e.g. 10]. Nonetheless, using a mathematical rule makes the aggregation explicit and auditable, and makes the results reproducible. Different rules satisfy different properties and unfortunately it is impossible to have all desirable properties in one rule [7]. One well established and used aggregation rule was formulated in [8], and it is a linear combination of judgements weighted by the experts' prior performance on similar tasks. Cooke's method, also called the Classical Model (CM) of structured expert judgement (SEJ), asks experts to give estimates that can be validated with data in a process that is transparent and neutral. This particular way of aggregation makes CM satisfy all the desiderata of a SEJ protocol.

Ideally after the aggregation, the resulting single probability reflects many of the experts' initial judgements, and even though they do not recognise it as their own, they should have no valid arguments against it. The consensus is not achieved by conferencing, but in a rather external way, thorough the mathematical aggregation. This sort of consensus is what Cooke calls rational consensus [8]. Achieving a rational consensus single probability may be very difficult in situations when experts strongly disagree and have very little interaction and feedback from their peers.

A structured protocol which strives to deal with such situations is the IDEA protocol [e.g. 12]. IDEA builds on CM, while using elements from the behavioural aggregation techniques, which makes it a mixed protocol for SEJ (similar to the well known Delphi protocol [22]). IDEA asks experts for their individual estimates without allowing them to interact, presents the anonymised set of judgements back to the group of experts, and encourages facilitated discussion and extensive interaction, while discouraging consensus. After experts share their reasoning, (sources of) data, and (mental) models they have the opportunity to privately modify their initial estimates (if they so wish) in accordance with the discussion. These second estimates are then mathematically ag-

gregated. The aggregation can be either an equally, or a differentially weighted linear combination. If differential weights are used in the IDEA protocol, they are always proportional with measures of prior performance on similar tasks [11].

An aggregated opinion can be viewed as that of a "virtual" expert. The same performance measures can be then used to both evaluate this virtual expert's performance, and to justify choosing the aggregation which performs best. Commonly used measures of performance are designed to be objective and conservative and focus on different attributes of good performance. They are measured on sets of questions to ensure sustained, rather than isolated good performance. Three of these attributes are long term accuracy, long term informativeness and calibration. Long term accuracy and informativeness are calculated per question and averaged across questions, hence they are average measures of performance. Accuracy measures how close an expert's estimate is to the truth, which is a difficult concept to interpret when the estimate is a probability of occurrence and the truth is the occurrence or non-occurrence of the event. Informativeness may measure the amount of entropy in what the expert says (independent of the actual occurrences of events), or the entropy in the expert's performance (without corresponding to the distribution that the expert, or anyone else believes) [8]. It may also measure the departure from the uniform distribution [10]. Rather than average measures for individual questions (variables) [8] proposes and discusses the advantages of measures for average probabilities. Calibration[1] rather than accuracy is proposed as a more appropriate measure of each experts' performance. Measures of performance are constructed using scoring rules. These scoring rules are random variables and analysing and comparing the scores' values requires knowledge about the scores' respective distributions. An important reason for Cooke's proposal is that the proposed score has a known asymptotic distribution, as opposed to (for example) the average Brier score for measuring accuracy [3], which does not. Moreover the score is asymptotically *proper*, which means that its expected pay-off is maximised only when experts express their true beliefs about the predicted event [e.g. 28]. Despite the very attractive theoretical properties of Cooke's calibration score, it only has a couple of real life applications for discrete variables [9, 2]. One reason for this lack of uptake may come from its asymptotic properties which

[1] In later publications Cooke started using the term "statistical accuracy" instead of the term calibration. We avoid this terminology in contexts where accuracy, defined as distance from truth, is also discussed.

imply the need of tens of questions in order to obtain reliable scores. These questions, commonly referred to as calibration questions, are additional to the questions of interest which are imperative, hence they are time consuming and add to the experts' fatigue. A reduced number is desirable. Another disadvantage of an asymptotic score is that comparing scores of experts who answered a different number of questions may be cumbersome and a power equalisation technique may be needed [8]. Both disadvantages point to the need for a score with an exact distribution.

In this paper we propose one such score (see Section 2), discuss its theoretical properties and compare it with Cooke's calibration score on a synthetic, simulated data set (in Section 3.1), as well as on large dataset containing predictions for outcomes of geopolitical events for the period $2011 - 2015$ (Section 3.2). The dataset together with the elicitation protocol used to elicit these data are described in Section 3.2.1. We conclude the paper in Section 4 with a discussion that outlines potential shortcomings of the new score and future research directions.

2 Methods

Assessing the probability of occurrence for certain events of interest equates to eliciting bivariate random variables. Nonetheless, the methodology presented in this paper can be extended to any discrete random variable. Let X be a bivariate random variable, such that $P(X = 1) = p$ and $P(X = 0) = 1 - p$. Experts are asked to estimate p, but instead of asking directly for this probability, they are asked to assign the event whose occurrence is modelled by X to a probability bin[2]. Any number $B > 2$ of bins can be used. For simplicity, we assume we present the experts with ten probability bins, denoted as $B_1, B_2, \ldots, B_{10}$. Every bin B_i is associated with a probability of occurrence p_i, for $i = 1, \ldots, 10$. An expert assigns an event to bin B_2 (for instance) if their best estimate (about the probability of occurrence) is anywhere between 0.1 and 0.2. We choose the middle value of this interval, that is 0.15, as the value of p_2 associated with B_2. Therefore, we choose $p_1 = 0.05, p_2 = 0.15, \ldots, p_{10} = 0.95$ and let $p = (p_1, p_2, \ldots, p_{10})$. Consider n binary random variables, $X_1, X_2, \ldots, X_n$ that model the probability of occurrence of n events. Assume an expert

[2]It is worth mentioning that this sort of elicitation is rarely used in practice; instead experts are asked about the direct probabilities [e.g. 12, 16] and their answers are used in conjunction with this construction in a rather artificial way. We will discuss this issue further in Section 3.2.

assigns n_i events to bin B_i, for $i = 1, \ldots, 10$. Then $n = \sum_{i=1}^{10} n_i$. Given the outcome of each event, for each bin B_i, we consider

$$s_i = \frac{\sum_{j=1}^{n} X_j \cdot \mathbf{1}_{\{X_j \in B_i\}}}{n_i},$$

for $i = 1, 2, \ldots, 10$, where $\mathbf{1}_{\{X_j \in B_i\}}$ is an indicator function such that $\mathbf{1}_{\{X_j \in B_i\}} = 1$ if X_j is assigned to B_i and 0 otherwise. Then s_i accounts for the frequency of occurring events in each bin. The vector $s = (s_1, s_2, \ldots, s_{10})$ is commonly referred to as the empirical probability vector. Ideally, s_i is close to p_i, for every i, hence the null hypothesis of $H_0 : s_i = p_i$ becomes a natural way of testing the hypothesis that the experts answer according to the designed bins and their respective probabilities. Nonetheless, some differences are to be expected in practice. Different approaches in quantifying these differences can be employed, which will lead to different calibration scores. We will first present Cooke's calibration score, which will be followed by our novel calibration score.

In Cooke's method [8], the discrepancy between s_i and p_i is measured using the Kullback–Leibler divergence, also called the relative information of s_i with respect to p_i. That is

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

for $i = 1, \ldots, 10$. Under the assumption of having n_i independent events with probability of occurrence p_i, which translates into the null hypothesis $H_0 : s_i = p_i$, the quantity $2n_i I(s_i, p_i) \sim \chi^2_1$, asymptotically, as $n_i \to \infty$. Furthermore, if all $X_1, \ldots, X_n$ are independent, and if for any $X_j \in B_i$, we have $P(X_j = 1) = p_i$, then

$$\sum_{i=1}^{10} 2n_i I(s_i, p_i) \sim \chi^2_{10},$$

as $n_i \to \infty$, for $i = 1, \ldots, 10$. Let F be the distribution function of $\sum_{i=1}^{10} 2n_i I(s_i, p_i)$ under the null hypothesis. The calibration score of expert e is then defined as

$$Cal_{\chi^2}(e) = 1 - F\left(\sum_{i=1}^{10} 2n_i I(s_i, p_i)\right). \tag{1}$$

The calibration score takes values from 0 to 1 and a small discrepancy between s and p yields a

small relative information of s with respect to p and hence a high calibration score. The calibration score has been proven to be a proper scoring rule asymptotically [8], that is, the experts maximize their expected scores, in the long run, if and only if they state their true beliefs. For n sufficiently large, we can use the approximation in (1). However, what "sufficiently large" means remains unclear. A rule of thumb has been proposed in [2]

$$n_i p_i \geq 4, \text{ and } n_i(1-p_i) \geq 4.$$

These inequalities pose serious constraints to the sample size, which would require hundreds of question for the above approximation to render reliable results. The question is, of course, how well does the Chi-square distribution approximate the distribution of $\sum_{i=1}^{10} 2n_i I(s_i, p_i)$. Simulations have been performed in Section 3.1, in order to investigate how accurate the χ^2 approximation is for small samples. Another approach proposed in [2] is to approximate F with an empirical distribution F_n obtained via simulations, but this idea has not yet been put into practice. We have considered it in this paper. We have performed 1000 simulations for random n_i assignments to the bins of the fixed n number of questions. Furthermore, the outcomes of the n events are randomly assigned to 0 or 1. An empirical distribution function is then obtained from the simulated data. The calibration score is then defined as

$$Cal_{edf}(e) = 1 - F_n\left(\sum_{i=1}^{10} 2n_i I(s_i, p_i)\right), \tag{2}$$

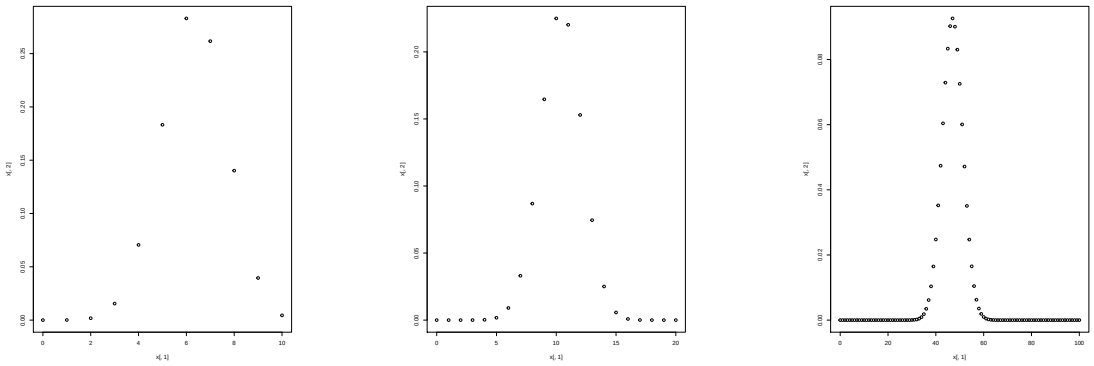
where F_n is the empirical distribution function obtained via simulations. We anticipate that the calibration score using the empirical distribution function is an asymptotically strictly proper scoring rule, since $F_n \to F$, almost surely, by the strong law of large numbers, as $n \to \infty$. The formal proof will be deferred to a later manuscript.

In contrast to Cooke's proposal, we employ a statistical hypothesis test whose test statistic has an exact rather than an asymptotic distribution. The null hypothesis concerns the equality between two probability vectors, i.e. $H_0 : s = p$, and, under H_0, one can observe that $s_i n_i$ follows

a binomial distribution with parameters n_i and p_i

$$s_i n_i \sim Bin(n_i, p_i).$$

Let $Y_i \sim Bin(n_i, p_i)$, and consider $Y = \sum_{i=1}^{10} Y_i$, which denotes the theoretical null distribution. Results of [6] provide the exact distribution of Y. The probability mass function of sums of binomial distributed random variables for different n is provided in the figure below.



(a) (b) (c)

Figure 1: The probability mass function of the exact distribution Y for $n = 10$ (a), $n = 20$ (b) and $n = 100$ (c).

We propose a calibration score based on the exact null distribution, that is, we compute the p-value of the hypothesis test. Note that the exact null distribution is discrete and not symmetric for small n, as depicted in Figure (1). We therefore employ a two-sided p-value that accounts for the the discreteness of the exact distribution of Y. A well-known approach is the two-sided mid-p-value [14, 15, 1],

$$\pi_{two}(a) = 2 \cdot \min\left(P(Y > a) + \frac{1}{2}P(Y = a), P(Y < a) + \frac{1}{2}P(Y = a)\right).$$

Then we define the calibration score of an expert e as

$$Cal_{bin}(e) = \pi_{two}\left(\sum_{i=1}^{10} n_i s_i\right). \tag{3}$$

Multiple approaches for two-sides p-values are available, some of which we initially considered. A good recent review can be found here [21]. We decided to use the mid-p-value due to its ease of interpretation and use, as well as its efficiency in dealing with the loss of power due to discreteness [21].

The main advantage of the proposed calibration score is its exact distribution, which will hopefully substantially reduce the necessary number of questions needed in order to provide reliable scores. As already mentioned, power equalisation techniques will also become unnecessary when comparing scores calculated from sets of questions of different sizes, so long as one set of questions is a subset of the other.

3 Results

We first interrogate an artificial data set to understand the behaviour of the three calibration scores when calculated for different number of questions. The theoretical properties of the calibration scores, such as the χ^2 approximation, rely on a sufficiently large number of questions. In an ideal situation, hundreds of questions should be answered in order to distinguish between the calibration scores of experts. In practice, it is very unlikely for experts to answer so many questions for calibration purposes. We expect the majority of studies to afford asking tens of questions at best.

After investigating the artificial situation, we turn our attention to a real data set collected during a forecasting tournament. We compare and investigate properties of the three calibration measures on this dataset.

3.1 Simulations

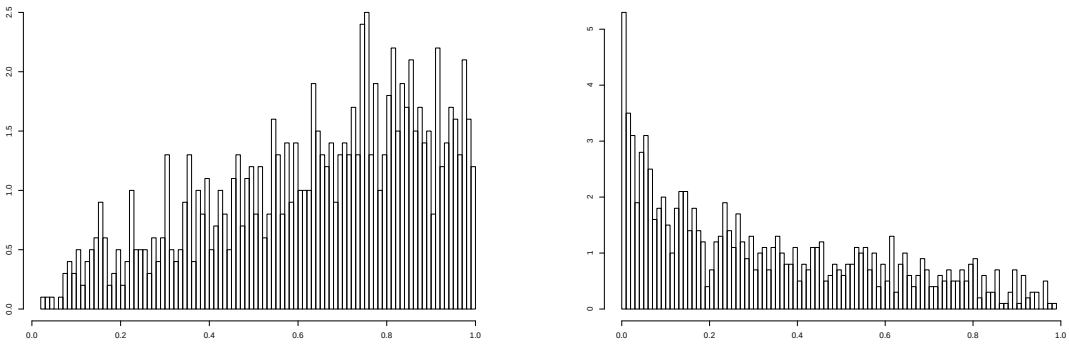
3.1.1 Simulated data

We have simulated expert data assuming random assignment to the ten bins, under the null hypothesis. The outcome of each variable in bin B_i has been assigned to one with probability p_i and to zero with probability $1 - p_i$. We used different dataset sizes, i.e. different numbers of calibration questions, and performed 1000 simulations for each dataset size. We show and discuss the findings

in the following section. We start by examining Cooke's calibration scores, developed using both the theoretical χ^2 distribution, as well as using the empirical distribution. Then we investigate the newly proposed calibration score based on the exact distribution of Y, the sum of ten binomial random variables.

3.1.2 Academic findings

Figure 2 shows calibration scores from simulations for ten and 100 questions, where the former represents the minimum allowable (and a very realistic[3]) number of calibration questions, and the latter represents the ideal (yet unachievable situation).



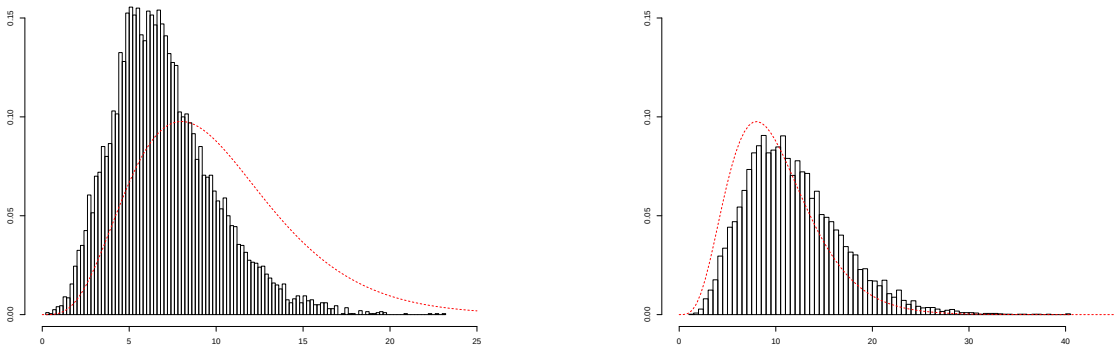
(a) Cal_{χ^2} distribution when using ten questions. (b) Cal_{χ^2} distribution when using 100 questions.

Figure 2: Histogram of calibration scores using the χ^2 distribution for $n = 10$ questions (a) and $n = 100$ questions (b).

Figure 2a shows that high calibration scores, that is, larger than 0.8 are more likely than small calibration scores, say smaller than 0.3. Figure 2b shows that small calibration scores will be more likely when answering 100 questions. Therefore, Figures 2a and 2b suggest very different (almost contradictory) behaviours of the calibration score in the two situations. These differences come from the differences in the approximations of the χ^2 distribution, and maybe from the fact that, when answering 100 questions, one is more likely to get low calibration scores by randomly assigning events to bins. On the other hand, for a small number of questions, one would expect any

[3]Between eight and ten calibration questions are considered to be enough in contexts where continuous variables and quantiles from their distributions are elicited instead [20].

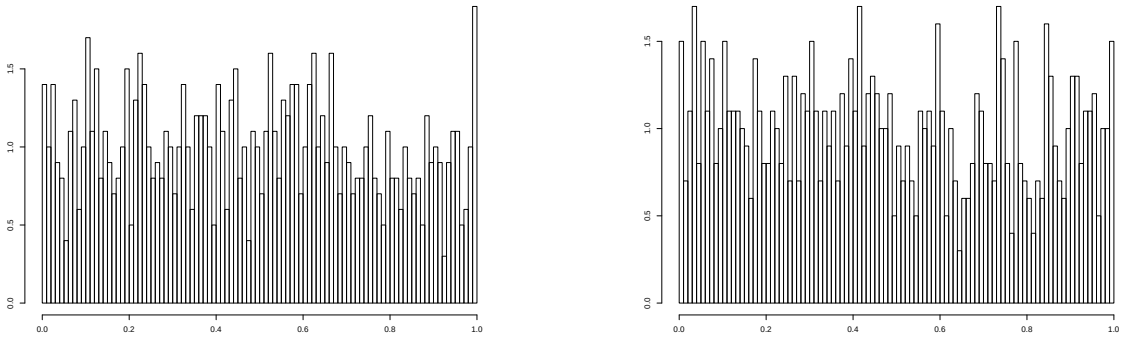
calibration score to be as likely as another. Cooke's calibration score seems more generous when using the χ^2 distribution. The question is of course whether the χ^2 distribution is appropriate for small sample sized data. The simulated data can bring some light into this matter.



(a) (b)

Figure 3: Histogram of $2n_iI(s_i, n_i)$ for $n = 10$ questions (a) and $n = 100$ questions (b).

Figure 3 shows this approximation for ten and 100 questions. As expected, the approximation of the χ^2 distribution, when only ten questions are used is very poor, as can be observed in Figure 3 (a), but it improves for 100 variables. Unfortunately asking 100 calibration questions is an academic rather than a practical situation. For small data sets, the empirical distribution is recommended and the histogram of the calibration scores using this empirical distribution is shown in Figure 4.

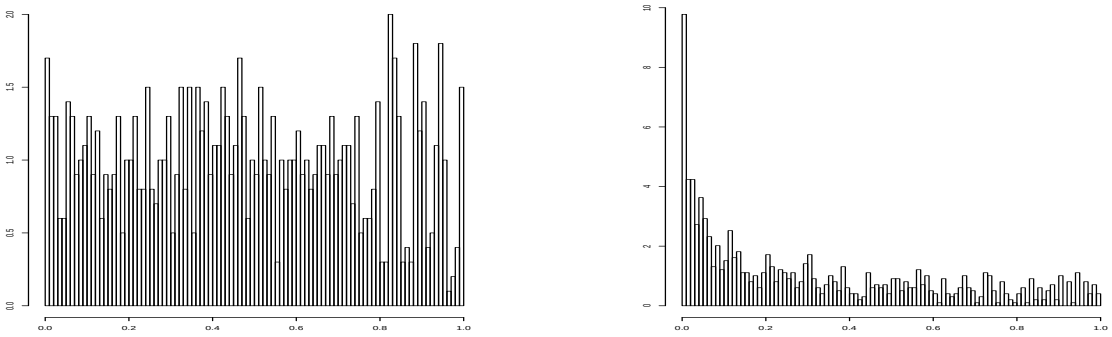


(a) (b)

Figure 4: Histogram of calibration scores using the empirical distribution function (edf) for $n = 10$ questions (a) and $n = 100$ questions (b).

Two observations are worth making when investigating Figure 4. The first is that the histograms for ten and 100 are surprisingly similar. The second observation relates to the comparison with Figure 2 which is supposed to be its theoretical analogue. The difference is striking, raising some questions about the appropriateness of the recommendation to use the empirical distribution instead or how appropriate the χ^2 approximation is.

For the new calibration score, developed using the exact distribution of Y, the simulation results are presented in Figure 5 below.



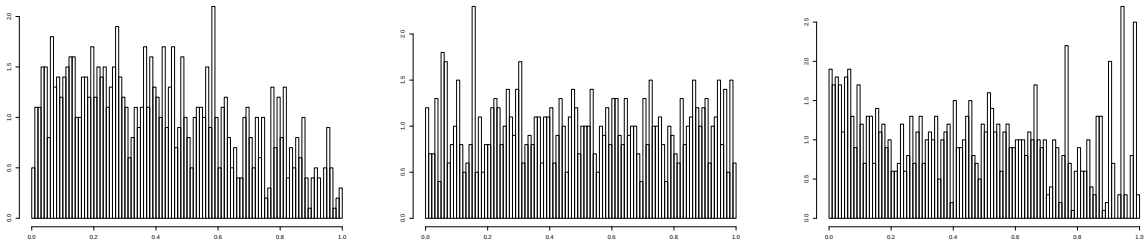
(a) (b)

Figure 5: Histogram of calibration scores using the exact distribution Y for $n = 10$ questions (a) and $n = 100$ questions (b).

The scores do not distinguish performance too well when only ten questions are used. We can see, from both plots in Figure 5, that the calibration score using the exact null distribution is somewhat conservative, that is, high calibration scores are less likely. The spikes in the histogram for $n = 10$ are due to the discreteness of the exact distribution, similar to the case where Cooke's calibration using the empirical distribution is calculated. When looking at 100 questions, the calibration scores using the exact distribution seem more conservative than the calibration score obtained by using the χ^2 distribution or the empirical distribution function, which is not surprising, given that this test is more powerful than the other two, in the sense that with so many questions it has the power to distinguish between calibration scores to a larger degree.

Assuming a set of 25 calibration questions corresponds to a realistic situation in a practical application, we compare the three calibration scores on the simulated data when $n = 25$. The

results are shown in Figure 6 below.



(a) Cal_{χ^2} (b) Cal_{edf} (c) Cal_{bin}

Figure 6: Histogram of calibration scores for $n = 25$, using the χ^2 (a), the empirical distribution (b) and the exact distribution Y (c).

As expected, the calibration score using the empirical distribution is not very different than the ones presented in Figure 4. The calibration score based on the exact test may be considered as a good compromise between Figures 5a and 5b if it wasn't for its unstable behaviour towards the high end of the calibration scores. The fluctuations for scores larger than 0.9 are due to the discreteness of the theoretical distribution; nonetheless they remain troublesome.

The advantage of using the new calibration score are, as pointed in the introduction, the use of an exact theoretical distribution rather than an asymptotic distribution. Furthermore, the new calibration score makes use of a p-value which takes into account (to some extent) the discreteness of the exact distribution, as well as its asymmetry. However, more research and possibly some adjustment are needed in order to understand and fix the above mentioned instability.

3.2 Application

As mentioned beforehand, we are not aware of any recent available data sets elicited by asking experts to place events in bins. Most applications ask experts to give their best estimate of the probability of an event occurrence. As part of the eliciting protocols, often experts are asked to first think of all the reasons why this probability may be low, then they are asked to think of all the reasons why the probability may be high and only then, balancing out the counter-factuals are they asked to provide a best estimate for the probability. The first two questions are meant to ensure that experts consider all possible scenarios and review all available information before

answering. They are also sometimes used as an indication of how uncertain the experts are when they make these estimates. The lower and upper bounds are very rarely used in the probabilistic analysis due to the lack of their operational definitions. However the length of these intervals may be compared with the length of the assumed bins, and this comparison may give an indication of how appropriate the binning of the probability scale is.

3.2.1 Data

The data used in this research was collected during a forecasting project that started in 2011 as an initiative of the US Intelligence Advanced Research Projects Activity (IARPA). Five university-based research teams entered this forecasting "tournament" and predicted hundreds of geopolitical events, with the aim of designing an effective elicitation protocol that predicts events with high accuracy. Real events that resolved in the near-future were used to test the accuracy of forecasts. Thousands of forecasters made over a million forecasts on hundreds of questions [26, 16].

The Centre of Excellence for Biosecurity Risk Analysis (CEBRA), at the University of Melbourne, was part of one of these teams in the first couple of years, after which the winner of the tournament was declared to be the Good Judgement Team (GJT)[4][26]. However the other teams still had access to the events used in the following couple of years, and as a consequence, CEBRA continued to collected data for a total of four years using internal resources and the questions posted by IARPA.

The collection of the data was done using the IDEA protocol ("Investigate", "Discuss", "Estimate", "Aggregate") developed at CEBRA and refined through this tournament [29]. The protocol distils the most valuable steps from existing structured protocols, and combines them into a single and practical protocol. The full protocol has been outlined in [12]; however, briefly, the key steps include:

1. Recruit a diverse group of experts.
2. Experts first *Investigate* the questions and clarify their meanings, then provide their private, individual best estimates and associated credible intervals.

[4]CEBRA was part of the team that came second with only a fraction of the GJT's participants.

3. Experts receive feedback on their estimates in relation to other experts.
4. With the assistance of a facilitator, experts *Discuss* the results, resolve different interpretations of the questions, cross-examine reasoning and evidence, and provide a second and final private *Estimate*.
5. The individual estimates are combined using mathematical *Aggregation*.

The data used in this research represent the answers to a subset of the questions developed by IARPA. All questions considered correspond to binary variables of the following sort: "Will the Turkish government release imprisoned Kurdish rebel leader Abdullah Ocalan before 1 April 2013?", which were answered using the so called three-step format [e.g. 4], which asks for the best estimate for the probability of the events' occurrence and for plausible upper and lower bounds for this same probability. All questions resolved within 12 months, allowing for validation exercises and accuracy calculations. The elicitation took place remotely, initially via email, and from the second year of the tournament through a website which provided the participants[5] with a platform where they could answer the questions and upload/download materials. Discussion was facilitated via a discussion board that could be used to share relevant resources from outside the platform, and to comment on and rate the quality of the information shared by others. The experts were divided in groups. Each year, new participants joined, and other participants left the project. There were 150 participants (in total) who answered at least one question (both rounds). A total of 155 questions were answered by at least one participant, with no one participant answering more than 96 questions; 84 participants answered at least ten questions; 39 participants answered at least 25 questions. For more details about the tournament, questions and participants we refer to [29, 12, 10, 11].

3.2.2 Practical Findings

Using the above described data set, we calculate the three calibrations scores for the participants who answered at least ten questions. The results are presented in Figure 7.

[5]We will use "participants" instead of "experts" because the requirements of expertise, when recruting, were not very strict. Anybody with an interest in the subject and a good understanding of the questions was allowed to participate.

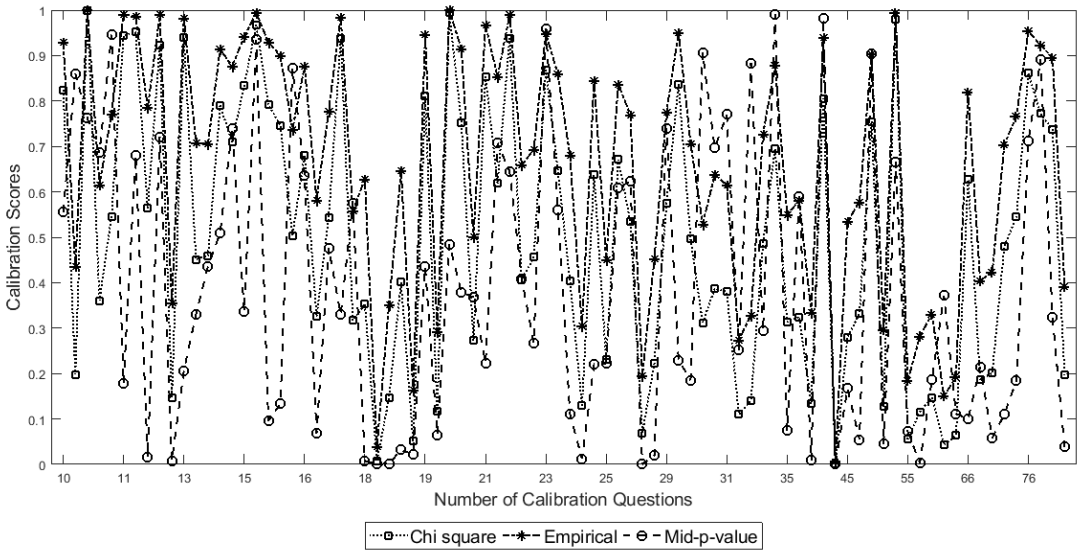
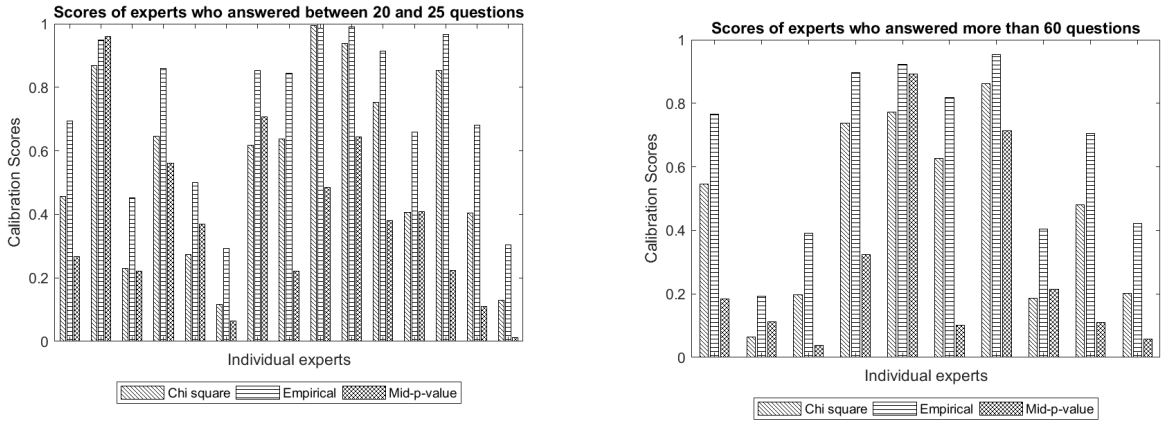


Figure 7: Calibration scores of participants who have answered at least ten questions ordered by the number of questions answered.

Often the scores agree, with the difference between the two calibration scores proposed by Cooke being smaller than the difference between each of them and the new proposed calibration. Very often, the new calibration score is smaller than the other two scores (79% of the times Cal_{bin} is less than Cal_{edf}, and 63% of the times is less than Cal_{χ^2}). For 82 of the 84 experts, the calibration score using the χ^2 approximation was smaller than the calibration score using the empirical distribution. For 34.5% (29) of the experts, the differences between Cal_{χ^2} and Cal_{edf} was smaller than 0.05. For 15.5% of the experts, the differences between Cal_{χ^2} and Cal_{bin} were smaller than 0.05. Interestingly, the same percentage, of 15.5% of the experts is registered when comparing Cal_{edf} and Cal_{bin}. Many of the variations are caused by using a small number of questions used (see the left hand side of Figure 7). When more than 20 calibration questions are used, less discrepancies are observed. However there are still a couple of differences larger than 0.7 between Cal_{edf} and Cal_{bin}. We shall investigate these (and other) differences on two subsets of scores, one calculated for experts who answered between 20 and 25 calibration questions (see Figure 8a), and the other calculated for experts who answered more than 60 questions (Figure 8b). The first subset corresponds to a realistic situation (for the calibration questions set size), whereas the second one corresponds to

a more theoretical one. Both these sets are subsets of the calibration scores showed in Figure 7, presented separately for better visibility.



(a) Scores of 15 experts answering no less than 20 and no more than 25 questions.

(b) Scores of ten experts answering no less than 60.

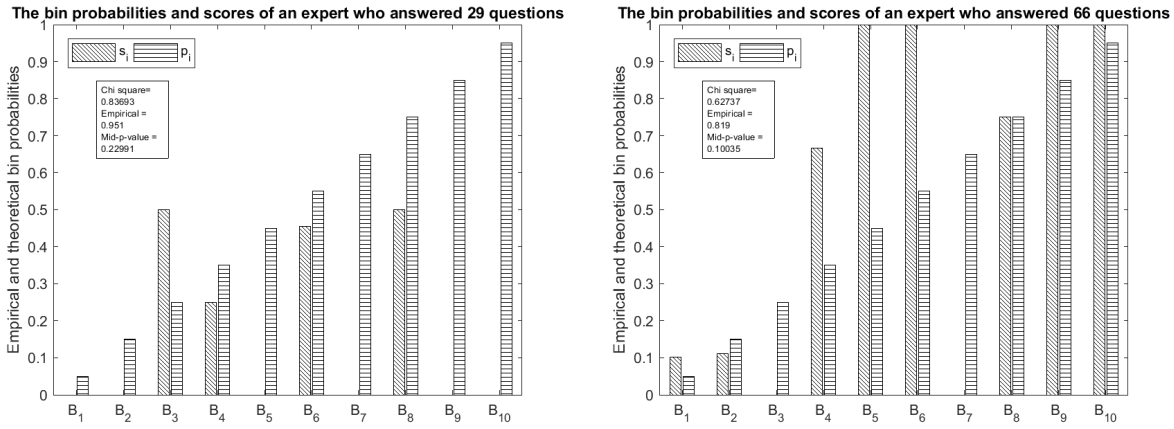
Figure 8: Cal_{χ^2}, Cal_{edf} and Cal_{bin} for experts answering between 20 and 25 questions (a) and more than 60 questions (b).

The large differences between scores are the ones which need further investigation. What properties does one score reward, while the other penalises? To better understand these sort of situations we look at the way events are placed in the probability bins and how well theoretical and empirical probabilities of occurrence align. Table 1 shows both examples of situations where the scores are in agreement (rightly or not), and situations where they are in disagreement, in a variety of cases when scores are calculated based on either very few, a realistic, or a large number of calibration questions.

Row	#	$s_1(p_1)$	$s_2(p_2)$	$s_3(p_3)$	$s_4(p_4)$	$s_5(p_5)$	$s_6(p_6)$	$s_7(p_7)$	$s_8(p_8)$	$s_9(p_9)$	$s_{10}(p_{10})$	Cal_{χ^2}	Cal_{edf}	Cal_{bin}
1	10	1 (0.05)	0 (0.15)	0 (0.25)	0 (0.35)	0 (0.45)	1 (0.55)	0 (0.65)	0 (0.75)	0 (0.85)	0 (0.95)	0.198	0.435	0.8605
2	13	0 (0.05)	0 (0.15)	0 (0.25)	0 (0.35)	0 (0.45)	0 (0.55)	0 (0.65)	0 (0.75)	0 (0.85)	1 (0.95)	0.939	0.983	0.206
3	15	0(0.05)	0.286 (0.15)	0(0.25)	0.334(0.35)	0(0.45)	0(0.55)	0(0.65)	0(0.75)	0(0.85)	0(0.95)	0.967	0.994	0.935
4	15	0(0.05)	0(0.15)	0(0.25)	0(0.35)	0(0.45)	0(0.55)	0(0.65)	0.667(0.75)	0(0.85)	0(0.95)	0.793	0.93	0.095
5	21	0(0.05)	0(0.15)	0(0.25)	0(0.35)	0(0.45)	0(0.55)	0.5(0.65)	1(0.75)	0(0.85)	0(0.95)	0.854	0.967	0.223
6	29	0 (0.05)	0(0.15)	0.5 (0.25)	0.25 (0.35)	0 (0.45)	0.455 (0.55)	0 (0.65)	0.5 (0.75)	0 (0.85)	0 (0.95)	0.837	0.951	0.230
7	41	0 (0.05)	0(0.15)	0 (0.25)	0.086 (0.35)	0 (0.45)	0 (0.55)	0 (0.65)	1 (0.75)	0 (0.85)	0 (0.95)	0.002	0.001	0.000
8	66	0.103(0.05)	0.111 (0.15)	0 (0.25)	0.667 (0.35)	1 (0.45)	1 (0.55)	0 (0.65)	0.75 (0.75)	1 (0.85)	1 (0.95)	0.627	0.819	0.1
9	75	0.03(0.05)	0.045(0.15)	0.334(0.25)	0.5(0.35)	0.5(0.45)	0.334(0.55)	1(0.65)	0.4(0.75)	1(0.85)	1(0.95)	0.545	0.766	0.183

Table 1: The three calibration scores obtained by several experts answering a different number of questions recorded in the second column (#) of the table.

When less than 25 questions are used, it is often hard to understand the behaviour of the scores. The first four rows of Table 1 exemplify this. The first row shows an example where all of the bins' probabilities differ considerably from their theoretical values, but both Cal_{edf} and Cal_{bin} fail to acknowledge that through their values, with the failure of the binomial test being more flagrant. In the second row, only the first and the last bin have recovered their theoretical probabilities, and yet Cal_{edf} rewards this with its almost maximal value. In contrast, the Cal_{bin} value seems a little more reasonable. In the third row all scores fail badly by giving very high calibration values to a very poorly estimated distribution of bins. The forth row shows an example where the number of questions answered is the same as in the previous row (15 questions), but the tests behave somewhat differently, in that the binomial test seems to capture the poor performance, unlike the other tests. A somewhat similar situation is shown in the fifth row, but this time, more questions are used (21 questions). The seventh row of the table shows a good example of perfect agreement and reasonable scores. Rows six, eight and nine are harder to discuss from the perspective of which score behaves better. Figure 9a corresponds to the sixth row of Table 1, and Figure 9b corresponds to the eighth row of Table 1.



(a) The empirical (s_i) and theoretical (p_i) bins' (B_i) probabilities, and the corresponding calibration scores for an expert who answered 29 calibration questions.

(b) The empirical (s_i) and theoretical (p_i) bins' (B_i) probabilities, and the corresponding calibration scores for an expert who answered 66 calibration questions.

Figure 9: Examples of he empirical (s_i) and theoretical (p_i) bins' (B_i) probabilities and corresponding calibration scores.

Investigating the way the empirical and theoretical bins' probabilities match for the example depicted in Figure 9a suggests an inflated Cal_{edf} and an appropriate value of Cal_{bin}. However, when we shift our attention to Figure 9b, Cal_{edf} seems appropriate, but Cal_{bin} is inexplicably small. It is hard to say if the score penalises a certain behaviour, or the small value comes from the observed instability of the score in the upper tail. We suspect the latter. Because the test we use for Cal_{bin} is exact, we do not necessarily expect the score behaviour to improve for a larger number of samples. However, if the 29 questions answered by the expert from Figure 9a were a subset of the 66 questions answered by the expert from Figure 9b, then a fair expectation is for the score to improve.

4 Discussion

4.1 Concluding remarks

A new score for measuring how calibrated experts' assessments of probabilities are, was introduced and discussed. This score has a known exact distribution. We have investigated the score's theoretical properties and practical performance and discussed a number of its positive and negative attributes. The identified need for a score which uses an exact rather than an approximate distribution of its test statistic is satisfied, but this comes with other shortcomings, some identified in this paper, some yet to be discovered.

There are many unanswered questions and many properties of the new calibration score that need to be properly investigated. The theoretical properties of the score are promising, but its practical properties need to be equally good.

One important practicality which seems to be better when considering the new calibration versus Cooke's calibration is the number of calibration questions needed, since calibration scores requiring hundreds of questions may be very interesting academically, but very impractical.

However, the instabilities Cal_{bin} exhibits in its upper tail need to be further investigated and resolved.

4.2 Future work

Our current research interests include proving that Cal_{bin} is a proper scoring rule, and accounting for the non-independence of $s_i n_i$, due to the $\sum_{i=1}^{10} n_i = n$ constraint.

Another research interest, that we have touched upon when analysing the data from the Intelligence Game, is that the probabilities of occurrence are not usually elicited using the bins, but they are asked for in a direct manner using prompts like an upper plausible and lower plausible bound before asking for a best estimate. When investigating those bounds, which can be thought of as the size of the bins, very rarely do they equal 0.1. This suggests that experts would be fairly uncomfortable with assigning events to bins of the size we investigated. Designing elicitations where experts are asked about placing events in probability bins, but complementing that with experiments to establish "bin size comfort" may be very valuable.

References

[1] Alan Agresti. *Categorical data analysis*, volume 482. John Wiley & Sons, 2003.

[2] B. Bhola and R.M. Cooke. Expert opinion in project management. *European Journal of Operational Research*, 57:24 – 31, 1992.

[3] G.W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3, 1950.

[4] M.A. Burgman. *Trusting judgements: how to get the best out of experts*. Cambridge University Press, 2015.

[5] M.A. Burgman, M. McBride, R. Ashton, A. Speirs-Bridge, and L. et al. Flander. Expert status and performance. *PLoS ONE*, 6:e22998, 2011.

[6] Ken Butler and Michael A Stephens. The distribution of a sum of independent binomial random variables. *Methodology and Computing in Applied Probability*, 19(2):557–571, 2017.

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

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

[9] R.M. Cooke, M. Mendel, and W. Thijs. Calibration and information in expert resolution. *Automatica*, 24(1):87–94, 1988.

[10] A. Hanea, M. McBride, M. Burgman, and B. Wintle. Classical meets modern in the idea protocol for structured expert judgement. *Journal of Risk Research*, 2016.

[11] A. Hanea, M. McBride, M. Burgman, and B. Wintle. The value of discussion and performance weights in aggregated expert judgements. *Risk Analysis*, 2018.

[12] A. Hanea, M. McBride, M. Burgman, B. Wintle, F. Fidler, L. Flander, S. Mascaro, and B. Manning. $I_{nvestigate}D_{iscuss}E_{stimate}A_{ggregate}$ for structured expert judgement. *International Journal of Forecasting*, 33(1):267–279, 2016.

[13] V.B. Hinsz, R.S. Tindale, and D.A. Vollrath. The emerging conceptualization of groups as information processors. *Psychological Bulletin*, 121(1):43–64, 1997.

[14] HO Lancaster. The combination of probabilities arising from data in discrete distributions. *Biometrika*, 36(3/4):370–382, 1949.

[15] HO Lancaster. Significance tests in discrete distributions. *Journal of the American Statistical Association*, 56(294):223–234, 1961.

[16] B. Mellers, E. Stone, P. Atanasov, N. Rohrbaugh, S.E. Metz, L. Ungar, M.M. Bishop, M. Horowitz, E. Merkle, and P. Tetlock. The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied*, 21:1–14, 2015.

[17] G. Montibeller and D. von Winterfeldt. Cognitive and motivational biases in decision and risk analysis. *Risk Analysis*, 35 (7):1230–1251, 2015.

[18] A. O'Hagan. Elicitation. *Significance*, 2:84–86, 2005.

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

[20] John Quigley, Abigail Colson, Willy Aspinall, and Roger M. Cooke. Elicitation in the classical model. In Luis C. Dias, Alec Morton, and John Quigley, editors, *Elicitation: The Science and Art of Structuring Judgement*, pages 15–36. Springer International Publishing, Cham, 2018.

[21] Isabelle Rivals, Léon Personnaz, Lieng Taing, and Marie-Claude Potier. Enrichment or depletion of a go category within a class of genes: which test? *Bioinformatics*, 23(4):401–407, 2006.

[22] G. Rowe and G. Wright. Expert opinions in forecasting: the role of the Delphi technique. *In Principles of forecasting: A handbook for researchers and practitioners*, Norwell: Kluwer Academic Publishers:125–144, 2001.

[23] K. Shrader-Frechette. Value judgments in verifying and validating risk assessment models. In C.R. Cothern, editor, *Handbook for environmental risk decision making: values, perception and ethics*, pages 291–309, London, 1996. CRC Lewis Publishers, Boca Raton.

[24] P. Slovic. Trust, emotion, sex, politics, and science: surveying the risk-assessment battle field. *Risk Analysis*, 19:689–701, 1999.

[25] William J Sutherland and Mark Burgman. Policy advice: use experts wisely. *Nature News*, 526(7573):317, 2015.

[26] L.H. Ungar, B. Mellers, V.A. Satopaa, J. Baron, P.E. Tetlock, J. Ramos, and S. Swift. The good judgment project: A large scale test of different methods of combining expert predictions. AAAI Fall Symposium Series, (AAAI Technical Report FS-12-06), 2012.

[27] L.J. Valverde. Expert judgment resolution in technically-intensive policy disputes. In *Assess-*

ment and management of environmental risks, pages 221–238. Kluwer Academic Publishers, Norwell, 2001.

[28] R.L. Winkler and A.H. Murphy. "Good" probability assessors. *Journal of applied Meteorology*, 7:751 – 758, 1968.

[29] B. Wintle, M. Mascaro, F. Fidler, M. McBride, M. Burgman, L. Flander, G. Saw, C. Twardy, A. Lyon, and B. Manning. The Intelligence Game: Assessing Delphi groups and structured question formats. In *Proceedings of the 5th Australian Security and Intelligence Conference*, 2012.