An in-depth perspective on the Classical Model

A.M. Hanea[a, *] & G.F. Nane [b]
[a] *Centre of Excellence for Biosecurity Risk Analysis & Centre for Environmental and Economic Research , University of Melbourne, Australia*
[b] *Delft Institute of Applied Mathematics, Delft University of Technology*

Abstract

The Classical Model (CM) or Cooke's method for performing Structured Expert Judgment (SEJ) is the best known method that promotes expert performance evaluation when aggregating experts assessments of uncertain quantities. Assessing experts' performance in quantifying uncertainty involves two scores in CM, the calibration score (or statistical accuracy) and the information score. The two scores combine into overall scores, which, in turn, yield weights for a performance-based aggregation of experts' opinion. The method is fairly demanding, and therefore carrying out a SEJ elicitation with CM requires careful consideration. This chapter aims to address methodological and practical aspects of CM into a comprehensive overview of the CM elicitation process. It complements the chapter "Elicitation in the Classical Model" in the book *Elicitation* [27]. Nonetheless, we regard this chapter as a stand-alone material, hence some concepts and definitions will be repeated, for the sake of completeness.

1 The Classical Model: overview and background

Structured expert elicitation protocols have been deployed in many different areas of applications [e.g. 13, 26, 2, 17] and Part 4 of this book. Even though most are guided by similar methodological rules, they differ in several aspects, e.g., the way interaction between experts is handled, and the way an aggregated opinion is obtained from individual experts.

As mentioned in the introductory chapter of this book, the two main ways in which experts' judgements are aggregated are: behaviourally (by striving for consensus via facilitated discussion), and mathematically (by using a mathematical rule to combine independent individual expert estimates). Mathematical rules provide a more transparent and objective approach. A weighted linear combination of opinions is one example of such a rule. While evidence shows that equal weighting frequently performs well relative to unequal, performance-based weighting methods for reliably estimating central tendencies [e.g. 8], when uncertainty quantification is sought, differential weighting provides superior performance [9].

A widely used version of a differential weighting scheme is the Classical Model (CM) for structured expert judgement (SEJ) [11]. CM was developed and used in numerous professional applications[1] involving the quantification of various uncertainties required to aid rational decision making. These uncertain quantities usually refer to unknown variables measured on continuous scale. Point/"best" estimates are not sufficient when the quantification of uncertainty is the main aim, since they do not give any indication of how much the actual (unknown) values may plausibly differ from such point estimates. Expert uncertainties are thus quantified as subjective probability distributions. Experts

*Correspondence to: A.M.Hanea, CEBRA, University of Melbourne, Parkville, VIC 3010, Australia, email: anca.hanea@unimelb.edu.au

[1] We call a professional application one for which the problem owner is distinct from the analyst.

are, however, not asked about full distributions, or parameters of distributions, but rather about a fixed and finite number of percentiles (usually three) of a distribution. From these percentiles, a minimally informative non-parametric distribution is constructed. Parametric distributions may be fitted instead, but these will add extra information to the three percentiles provided by the experts, when compared to the minimally informative non-parametric distribution. This extra information may or may not be in accordance to experts' views.

Experts are elicited individually, and face-to-face interviews were recommended in the CM's original formulation. Variants of the CM's elicitation protocol involve workshops (ranging from half a day to three days), remote elicitations, or a combination of these. Each method has its advantages and disadvantages. Having all experts in one (potentially virtual) room may permit facilitated discussion prior to the actual elicitation with the aim of reducing ambiguity, providing feedback on practice questions, and better understanding of the heuristics to be avoided in order to reduce biases. However, these may come to the price of group biases, halo - effects, dominating or recalcitrant personalities, etc.

Rather than consensus, CM advances the idea of rational consensus, in which the parties (experts and facilitators) pre-commit to a scientific method for aggregating experts' assessments. CM operationalizes four principles which formulate necessary conditions for achieving rational consensus (the aim of rational decision making). These principles are detailed in the introductory chapter of this book and repeated here for convenience: *scrutability/accountability*, *empirical control*, *neutrality* and *fairness*. Cooke argues that a rational subject could accept these principles, but not necessarily accept a method implementing them. If this were the case, such a rational subject "incurs a burden of proof to formulate additional conditions for rational consensus which the method putatively violates." [10]. Even though part of the expert judgement community does not regard CM as an appropriate method for expert judgement [6, 7], to the best of our knowledge, no additional conditions for rational consensus, as proposed by Cooke, were formulated or identified as being violated. We note that there are numerous other sets of axioms proposed within the literature, see, e.g., [14].

The empirical control requirement is essential to the CM and, some would argue, e.g., [23], to any elicitation protocol which calls itself *structured*. It is this requirement that justifies the use of seed (calibration) variables to derive performance-based weights, providing an empirical basis for validating experts judgements that is absent in other approaches. We note however, that other methods, lacking empirical control, but eliciting expert judgments in a structured manner, following a rigorous protocol, are also considered SEJ protocols [5]."Seed" (or calibration) variables are variables taken from the problem domain for which, ideally, true values become known post hoc [2]. However, this is rarely feasible in practice, hence variables with known realizations (values) are used instead. The questions about the seed variables that the experts need to answer are called seed questions. Experts are not expected to know the answers to these questions precisely, but they are expected to be able to capture them within informative ranges, defined by ascribing suitable values to the chosen percentiles (usually the 5^{th}, 50^{th} and 95^{th}).

The theoretical background and mathematical motivation for many of the modelling choices which define the CM are detailed in [11]. However interesting and technically complete this book is, many CM neophytes find it difficult to decipher or navigate. For excellent short descriptions of the CM, written for practitioners and less technically inclined audiences, we recommend [27, 1].

CM is implemented in the software Excalibur, freely available from http://www.lighttwist.net/wp/excalibur. Excalibur is a fully functioning application (if somewhat old) which was originally developed at Delft University of Technology and it is now maintained by Lighttwist Software.

This chapter aims to complement the existing CM descriptions, draw attention to methodological and practical aspects which were not covered in the aforementioned descriptions, update recommendations made when the CM protocol was originally designed, and clarify assumptions and misconceptions. As we will emphasize throughout the chapter, some issues arise from necessary theoretical requirements,

while others are reasonable pragmatic assumptions. We stress that theoretical requirements define the rigorous setting of the Classical Model, and the pragmatic assumptions allow for model flexibility that can be explored by a more experienced user.

The remainder of this chapter is organised as follows: Section 2 discusses several elements that need to be organised prior to the elicitation and dwells on aspects which may be problematic or are critical for a successful elicitation. Section 3 details some steps of the elicitation protocol, from constructing an expert's distribution from elicited percentiles to evaluating experts performance using a calibration score, an information score and a combined score. These performance measures are discussed from a theoretical, practical and intuitive view point. Section 4 discusses different mathematical aggregations of experts' distributions and ways to evaluate them. Section 5 concludes the chapter with a few remarks.

2 Pre-elicitation for the Classical Model

If decision making is supported by quantitative models and the modelling is associated with uncertainties, then assessing uncertainty over the model inputs is essential. Assume a model is chosen appropriately (i.e., in accordance to needs and resources) and the sources of uncertainty are identified. Next, the modelers and analysts should collate and evaluate the available resources (e.g., data, prior studies, related literature). After completing this step, the data gaps will become apparent and the requirements for expert input can be formulated. With this we are entering what is often called the pre-elicitation stage. Many elicitation guidelines cover this stage [e.g. 10, 5], so in this section we will merely complement the existing guidelines by addressing only a few, less discussed, aspects.

2.1 Formal documents

Sometimes research which involves collecting subjective data from human participants needs a *human ethics approval*. Moreover, some journals require such approval to publish research informed by subjective data. Although less common in Europe, and the United States[2], this is very often a requirement in New Zealand and Australia.

A *project description* is another useful document. This will be outlining the purpose of the project, the relevant time-frames, the required expert input, and potential payments. A *consent form* sometimes accompanies the project description, and it is sent to participants to formalise their agreement to take part in the elicitation and to disclose any conflict of interests.

A *briefing document* guides participants through the elicitation, including the specific way to answer questions, the reasons behind asking the questions in a particular format, and the ways in which the answers are evaluated. An example of such document is [1].

The project description and briefing document are sometimes combined into one single document as recommended in [5]. As an alternative, the authors of [4] compiled a much larger document and made it available prior to the elicitation. This document is an extended version of the briefing document, augmented with background information and available literature, especially useful to inform assessments about the target variables. However, the available literature made available should not contain the answers to the seed variables, as this would invalidate the calibration exercise.

[2]In some instances it has been ruled that experts in an elicitation are not experimental subjects. If needed, human ethics only applies if the number of subjects is larger than nine, and only if the elicitation conducted by the Federal Government (R.M. Cooke, personal communication 2018).

2.2 Framing the questions

The most common format of asking experts to quantify their uncertainty about a continuous variable is eliciting three percentiles, normally the 5^{th}, 50^{th} and the 95^{th} percentiles. Eliciting five percentiles has also been used in practice [e.g., 25], where the 25^{th} and 75^{th} percentiles are elicited additionally to the three percentiles mentioned beforehand. Eliciting other percentiles or other number of percentiles (i.e., four percentiles) is nonetheless possible, posing no theoretical or practical problems. Excalibur supports formats with three, four or five elicited percentiles, which can be specified by the analyst.

However, for certain types of questions this is easier said than done. The difficulties can arise from several reasons, and we will touch upon three of these: 1) the underlying elicited variables are not continuous, 2) the questions are not about variables that experts are familiar with, but rather they address transformation of these variables, and 3) the experts are not statistically trained. The following discussion applies to both seed and target variables. Specific seed variables issues are discussed in a dedicated sub-section.

2.2.1 Modelling discrete data with continuous variables

Modelling discrete data with continuous random variables is not an unfamiliar practice in statistics, i.e., age of patients or months since surgery. Similarly, when eliciting bounded variables measured on a countable scale, most practitioners assume a continuous approximation of these variables and use the percentile elicitation procedure. This can be challenging for the experts. For example, assume a population of 10 healthy coral reefs. The experts are then asked about the number of future diseased coral reefs. Assume an expert's best estimate (corresponding to their median, the 50^{th} percentile) is one. The only value strictly less than one that they can estimate as their 5^{th} percentile is zero. However, that means that there is a one in 20 chance for the number of diseased coral reefs to be negative, which is physically impossible.

Situations like the one in the above example may lead experts to assign equal values for two or even all three percentiles, or to assign physical bounds instead of the extreme percentiles, even though they understand in theory, that the percentiles of a continuous variable have to be distinct, and different than the bounds.

2.2.2 Unfamiliar framing

Framing the question in a way that is different to the context experts are familiar with, dramatically increases the cognitive load, and should be avoided whenever possible.

For example, asking for three percentiles of variable X in relation to something normally expressed as a ratio, say 1/X, can be awkward. It is even worse if the expert thinks in terms of something which is naturally expressed as a different ratio, say, X = Y/Z.

2.2.3 Statistical proficiency

The assumption of an underlying continuous distribution comes with very clear theoretical constrains, among which: the extreme (upper and lower) elicited percentiles should not equal the physical bounds of the support of the variable, and the three percentile values should be strictly increasing. Above, we touched upon a situation where these constrains may be violated because the modelled variable is not in fact continuous (but rather approximated with a continuous variable). We now want to draw attention to situations were these constrains are violated because of the difficulty of the questions, coupled with an inadequate probabilistic and statistical training of the experts.

Let us consider the example of eliciting percentages which are though to be extreme. When experts need to estimate a very small or a very large percentage, they may assess the 5% percentiles to be 0%

or the 95% percentile to be 100%. It is the analyst's job to emphasize that the elicited quantity is uncertain and to try to guide the expert through probabilistic thinking. Advising experts to reason in terms of relative frequencies may sometimes be a solution. However, if it does not help, the experts assessments are usually slightly modified (i.e., by adding or subtracting a very small number such as 10^{-8}) to comply with the theoretical restrictions.

In certain situations, experts will assign equal values for two (or all three) percentiles even after a brief probabilistic training. If time allows, we advice that during training, an example should be used to emphasize why equal percentiles are problematic and less desirable for modelling distributions of continuous random variables. To exemplify this, consider expert's assessments for an unknown variable X to be 3 for the 5^{th} percentile, 3 for the 50^{th} percentile and 10 for the 95^{th} percentile. Then, the probability that the true percentage is 3 is 0.45, that is $P(X = 3) = 0.45$. Nonetheless, X is assumed to be a continuous random variable and the probability that X attains any specific value is zero, hence $P(X = 3)$ should be zero. Obviously, the expert does not acknowledge that her assessments do not correspond to a continuous random variable. And it is the analyst' job to clarify the setting. Finally, the requirement of strictly increasing percentiles has also been implemented in Excalibur.

The facilitators and analysts need to be aware of these issues when framing the questions. Sometimes, certain, possibly problematic formats cannot be avoided. Then, the experts need to be made aware of these difficulties and, if needed, be contacted after the elicitation for re-assessment.

2.3 Seed variables

The seed questions/variables are an essential element, since one of the main assumptions of CM is that prior performance on seed questions is a good predictor of future performance on the target variable/questions of interest[3]. When building the differential weighted aggregated distributions, these aggregations are basically fitted to seed questions and the entire model is calibrated on them. Their importance is paramount. A strong recommendation for analysts and facilitators is to consult a couple of domain experts when looking for and formulating seed variables (see also the dry run section below). Given their involvement with the seed questions, these experts' judgements cannot be formally elicited during the elicitation.

Seed variables and the purposes they serve are also discussed in detail in Section 2.3 of [27]. We reiterate below the main four types of seed variables (domain-prediction, domain-retrodiction, adjacent-prediction and adjacent-retrodiction), as cathegorised in [12], and qualify their desirability.

	Prediction	*Retrodiction*
Domain / Subject matter	Most desirable	Reasonably desirable
Adjacent / Contingent subject matter	Reasonably desirable	Last resort

Table 1: Types of seed variables and their desirability. The reasonably desirable options are the ones usually used in practice.

As mentioned beforehand, the answers to seed questions should not be known by experts during the elicitation. Table 1 provides general guidance for selecting seed variables. Ideally, the analyst should have access to ongoing studies or domain data which become available shortly after the elicitation. These make great sources for formulating domain-prediction variables. Examples can include data from official reports which will become available shortly after the elicitation takes place. Suppose experts are asked several questions about the percentage of unvaccinated children in Europe, in the

[3]From here on we will call *questions of interest* the questions related to the target variables.

period 2015-2018. The elicitation takes place in November 2019, and the WHO official report, which is the only source for these questions is due to appear in December 2019. Since one of the questions of interest regards the percentage of unvaccinated children in Europe in 2030, we regard the seed questions to be domain questions.

However, this not always possible, and data from recent studies within the subject matter or, less desirable, in adjacent subject matters are often the only option. Typically, data from official, yet not public, reports are used do define calibration questions. For example, existing confidential reports that document outbreaks of Salmonella in different provinces in The Netherlands could be used to define seed questions. If the questions of interest regard the number of cases of infection with Salmonella in the same provinces, then the seed questions are seen as being retrodictions and from the same domain. If, on the other hand, the question of interest regards the number of cases of infection with Salmonella at the national level, or even at the European Union level, the seed questions can regarded as being from an adjacent subject matter. Even though the question of interest refers to the same bacteria, it is defined in a different context than the calibration question and can therefore be seen as from an adjacent subject matter. Another, more clear, example is the following. Suppose the question of interest refers to the effects of Bonamia ostreae parasite in Ostrea chilensis oysters. Since this parasite-host combination is new, data are lacking and domain calibration questions are not possible. Calibration questions have been chosen to study the effects of different parasite-hosts combinations, i.e., Bonamia ostreae parasite in Ostrea edulis and Bonamia exitiosa parasite in Ostrea chilensis.

Often, elicitations need to involve two or more sub-disciplines. The set of seed questions should have then a balanced selection of items from each discipline. However, the boundaries between sub-disciplines are sometimes blurry and we are yet to learn how well can experts extrapolate their knowledge to answer questions from adjacent domains. This should be carefully dealt with prior to the elicitation and, if resources allow, consider separate panels of experts to answer different (sub-domain specific) seed questions.

Not only the domain of the seed variables is important, but also their formulation. We argue that the seed questions should be asked in exactly the same format as the questions of interest. There is no reason to believe that good performance on a certain type of task is transferable to different tasks. On the contrary, a couple of studies [28, 24] comparing the performance of experts when quantifying one-dimensional distributions using percentiles, with quantifying dependence between these one-dimensional margins, indicated a negative relationship.

Given that the domain and the formulation of the seed questions are appropriate, the next thing to consider is what sort of thinking they trigger for the experts. Answering the seed questions should certainly not be a memory test about factual knowledge alone. To be able to differentiate expert performance better, the seed questions should also be as diverse as possible. Experts need to be able to make judgements of appropriate uncertainties, hence the seeds should require experts to think about composite uncertainties, in the same way they would need to do when answering the questions of interest.

The seed questions may be asked before the questions of interest and feedback may be presented to the experts before they start answering the questions of interest. Another format of the questionnaire may have all questions in random order. Some (retrospective) seed questions will be identified as such by the experts, however, the predictive ones may not stand out as seeds. An argument for having a questionnaire where seed questions and questions of interest are randomly intermixed relates to the level of experts' fatigue, as increased fatigue affects the ability of experts to concentrate towards the end of the elicitation exercise.

For continuous quantities, between eight and ten seed questions were recommended [11] independent of the number of questions of interest. We argue that a minimum of 15 should be used when there are no more that 35 questions of interest and at least a one day workshop. These are of course guidelines derived from experience and practice, rather than results of proper studies on experts behaviour and

fatigue.

Many of the more recent studies using CM published all questions as supplementary material, but some of the older studies did not necessarily do so. As a future recommendation, aligned with the need for transparency imposed by Cooke's principles of rational consensus, we suggest all questions to be made available. Moreover, identifying and reporting the type of seed variables used, as characterised in Table 1 is highly recommended.

2.4 Dry-run

A dry run of the elicitation is strongly encouraged. Such an exercise is essential in decreasing the linguistic uncertainty (ambiguity), which is almost certainly present in the project description and, most importantly, in the formulation of the questions (of interest and seeds). It is also a good exercise for checking if all relevant information is captured and properly conveyed (in a language which is familiar to the experts). One or two domain experts should be asked to provide comments on all available documents, the questions, the additional information given for each question appreciated, and to estimate a reasonable time required to complete the elicitation.

2.5 Elicitation format

There is no single best way to carry on an expert elicitation using CM. The original setting proposed in [11] involves a face-to-face individual interaction between the facilitator and the expert. That is, the facilitator meets separately with each expert, trains them if necessary, discusses practice question(s), and then proceeds to guide the expert through the elicitation questions. Willy Aspinall (personal communication) carried out many of his numerous elicitations in a workshop setting. More recently, a number of elicitations have also been performed remotely, using one-to-one Skype interviews. In such cases a teleconference with all experts may be held prior to the individual elicitation interviews. During this teleconference the procedure, scoring and aggregation methods should be explained, and a couple of a practice questions should be answered [19, 4].

If the elicitation is done remotely and the seed questions are retrospective, the calibration exercise needs to be done "face-to-face" and the experts should work with the facilitator (e.g. in individual Skype sessions). The questions of interest can be then finished on a more relaxed time-frame and without the facilitator's virtual presence. However if all the seeds are predictive, individual (remote) interviews are not a requirement.

Special attention needs to be given to experts' uncertainty training. Reasoning with uncertainty and expressing uncertainty prove to be a challenging endeavor. Practice questions are therefore desirable. Some practitioners choose practice questions from the same domain as the seed variables and questions of interest. Others choose a different subject matter, e.g., questions regarding weather, in order to focus primarily on how experts express their uncertainty.

For more details on the elicitation format, we refer to Section 2.4 from [27].

3 Elicitation with the Classical Model

The many details decided upon in the pre-elicitation stage determine the elicitation itself. These include: the number and type of questions, the number and expertise of experts, the type of feedback given to, and interaction permitted between experts. Once the required estimates are elicited, they are scored and the scores are used to form weights. Several weighted combinations are calculated; they form several so-called decision makers (DM) distributions. It is worth mentioning that a decision maker in

this context represents a mathematically calculated distribution which corresponds to a virtual expert. The real decision maker would adopt one of the DMs distribution as their own.

3.1 From assessments to distributions

It is important to stress again that CM, as largely known from the literature, applies to continuous variables. That is, the elicited seed variables, as well as the variables of interest are modelled as continuous variables and the questions are formulated in terms of percentiles of continuous distributions. Moreover, all major CM applications made use of continuous variables. As already emphasized in places, this chapters provides an in-depth perspective on the Classical Model when using continuous random variables. Eliciting discrete random variables, in terms of the probabilities of their states, and scoring the experts' performance, even though proposed in [11], has scarce applications and it has not been implemented in Excalibur[4]. It is also noteworthy that CM should not be used for mixed types of questions, that is both discrete and continuous. Moreover, the questions (seed and of interest) should be either all continuous or all discrete.

The rest of this chapter refers solely to eliciting continuous random variables.

It is worthwhile discussing first how expert's distribution is actually constructed from the expert's assessed percentiles within the CM. In order to specify expert's distribution, we first need to determine the support of the distribution. Assume N experts provide their assessments. Denote expert's e_i assessments for a given question as q_5^i, q_{50}^i and q_{95}^i for the 5^{th}, 50^{th} and 95^{th} percentiles, respectively, and $i = 1, 2, \ldots, N$. The range $[L, U]$ is given by

$$L = \min_{1 \leq i \leq N} \{q_5^i, \text{realization}\},$$
$$U = \max_{1 \leq i \leq N} \{q_{95}^i, \text{realization}\},$$

for a given seed variable. Note that L denotes the minimum among all experts' lower bounds and the realization, whereas L denotes the maximum between all experts' upper bounds and the realization. For the questions of interest, the lower and upper bounds are determined exclusively by the experts' percentiles, i.e., $L = \min\{q_5^i\}$ and $U = \max\{q_5^i\}$, for $i = 1, \ldots, N$. The support of experts' distributions is then determined by the so-called intrinsic range

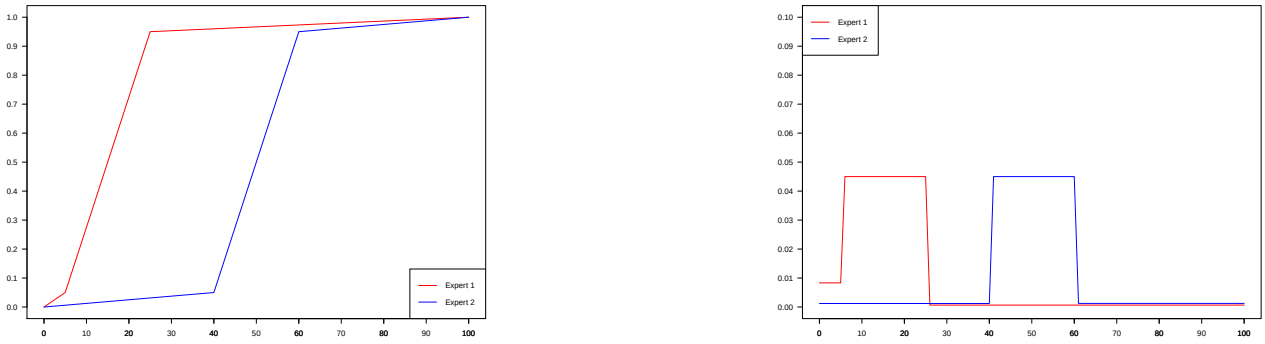
$$[L^*, U^*] = [L - k \cdot (U - L), U + k \cdot (U - L)],$$

where k denotes an overshoot and is chosen by the analyst (usually $k = 10\%$, which is also the default value in Excalibur). The intrinsic range therefore allows for an extension of the interval determined by the interval $[L, U]$. The extension is symmetrical for simplicity. For some questions, the intrinsic range can be specified a priori by the analyst[5]. For example, when eliciting percentages, the intrinsic range can safely be chosen to be $[0, 100]$.

Each of the expert's distribution is constructed then by interpolating between expert's percentiles such that mass is assigned uniformly within the inter-percentile ranges. Consequently, by assuming an

[4]The performance scores are calculated differently for discrete variables. Informativenes is replaced with entropy and the calibration score, even though still based on a simmilar test statistic, is different as well, and it requires many more seed variables for reliable estimation. The interest in this topic has been revived recently with a theoretical research on calibration scores [15].

[5]This is however not possible in Excalibur. Unrealistic ranges obtained in Excalibur need to be truncated externally.



(a) (b)

Figure 1: Cumulative distribution functions (1a) and probability distribution function (1b) for two experts whose assessments are $(5, 15, 25)$ (for Expert 1) and $(40, 50, 60)$ (for Expert 2).

uniform background measure, the distribution of expert e_i is given by

$$F_i(x) = \begin{cases} 0, \text{ for } x < L^* \\ \frac{0.05}{q_5^i - L^*} \cdot (x - L^*), \text{ for } L^* \leq x < q_5^i \\ \frac{0.45}{q_{50}^i - q_5^i} \cdot (x - q_5^i) + 0.05, \text{ for } q_5^i \leq x < q_{50}^i \\ \frac{0.45}{q_{95}^i - q_{50}^i} \cdot (x - q_{50}^i) + 0.5, \text{ for } q_{50}^i \leq x < q_{95}^i \\ \frac{0.05}{U^* - q_{95}^i} \cdot (x - q_{95}^i) + 0.95, \text{ for } q_{95}^i \leq x < U^* \\ 1, \text{ for } x \geq U^*. \end{cases}$$

The distribution is piecewise linear on the four intervals determined by the assessed percentiles. Note that the cumulative distribution F_i is continuous. The cumulative distribution and the corresponding density function for two experts with assessments $(5, 15, 25)$ (for Expert 1) and $(40, 50, 60)$ (for Expert 2) are depicted in Figure 1. The intrinsic range has been assumed $[0, 100]$, which is appropriate as the quantities are percentages.

The above construction of distributions is arguably the most popular method of constructing distributions.

3.2 Measures of performance

CM measures experts' performance as uncertainty assessors. Performance may be regarded as being determined by properties of experts' assessments that we value positively. Three of these properties are accuracy, calibration and informativeness. Often, in the judgement and decision-making literature, accuracy is understood as the distance from the "best estimate" to the true, realised value [e.g. 22, 18]. In the CM the best estimate is operationalised as the median (the 50^{th} percentile). To avoid difficulties related to estimating average accuracy across multiple seed variables, which will unavoidably be measured on different scales, the CM does *not* score accuracy as defined above. In turn it scores calibration and informativeness.

Confusingly, from a terminological point of view (in the context outlined above), the CM calibration is also called statistical *accuracy*[6]. We recall the technical definitions of calibration and informativeness and provide accompanying intuitive explanation.

[6]The terminology was changed from calibration to statistical accuracy because of another potential terminological clash with the engineering interpretation of the term calibration.

3.2.1 Calibration

Assume there are N experts, $e_1, e_2, \ldots, e_N$ and M seed variables/questions $SQ_1, SQ_2, \ldots, SQ_M$. Denote expert's e_i assessments on question j as $q_5^{i,j}$, $q_{50}^{i,j}$ and $q_{95}^{i,j}$ for the 5^{th}, 50^{th} and 95^{th} percentiles, respectively; the index j is sometimes omitted for convenience to denote the percentiles assessed for a random question (rather than for a given question j). The notation will then reduce to q_5^i, q_{50}^i and q_{95}^i. For each question, and each expert, the probability range is divided into four inter-percentile intervals, corresponding to inter-percentile probability vector $p = (0.05, 0.45, 0.45, 0.05)$. Suppose the realisations of these seed questions are x_1 for SQ_1, $\ldots, x_M$ for SQ_M. We may then form the sample distribution of expert e_i's inter-percentile intervals by simply counting how many of the M realizations fall within each inter-percentile interval. Formally, let

$$s_1(e_i) = \frac{|\{k|x_k \leq q_5^i\}|}{M} = \frac{\sum_{k=1}^{M} \mathbf{1}_{\{x_k \leq q_5^{i,k}\}}}{M},$$

$$s_2(e_i) = \frac{|\{k|q_5^i < x_k \leq q_{50}^i\}|}{M} = \frac{\sum_{k=1}^{M} \mathbf{1}_{\{q_5^i < x_k \leq q_{50}^i\}}}{M},$$

$$s_3(e_i) = \frac{|\{k|q_{50}^i < x_k \leq q_{95}^i\}|}{M} = \frac{\sum_{k=1}^{M} \mathbf{1}_{\{q_{50}^i < x_k \leq q_{95}^i\}}}{M},$$

$$s_4(e_i) = \frac{|\{k|q_{95}^i < x_k\}|}{M} = \frac{\sum_{k=1}^{M} \mathbf{1}_{\{q_{95}^i < x_k\}}}{M}.$$

where

$$\mathbf{1}_{\{x \leq a\}} = \begin{cases} 1, & \text{when } x \leq a \\ 0, & \text{otherwise} \end{cases}$$

is the indicator functions. Then $s(e_i) = (s_1(e_i), s_2(e_i), s_3(e_i), s_4(e_i))$, i.e., the empirical distribution for expert i. Note that if the expert assesses the uncertainty effectively, then we expect the distribution of the M counts to multinomial, with parameters $0.05, 0.45, 0.45$ and 0.05. Alternatively, if the realizations are indeed drawn independently from a distribution with percentiles as stated by the expert, then the quantity

$$2MI(s(e_i), p) = 2M \sum_{l=1}^{4} s_l(e_i) \ln \frac{s_l(e_i)}{p_l}, \tag{1}$$

is asymptotically distributed as a chi-square random variable with 3 degrees of freedom. Hence we can score expert e_i as the statistical likelihood of the hypothesis

H_{e_i} : the inter-percentile interval containing the true value for each variable is drawn independently from probability vector p.

In equation (1), M is the number of seed questions, and $I(s(e_i), p)$ is the Kullback-Leibler divergence [21], which Cooke calls the relative information of one distribution with respect to another [e.g. 13]. The relative information score measures how one distribution, s in this case, diverges from another distribution, p here. In other words, if the experts would indeed give values which correspond to the

$5^{th}, 50^{th}$ and 95^{th} percentiles of distributions, on the long run, their sample distribution s should be equal to p. Then $I(s(e_i), p) = 0$ and this should correspond to the highest possible calibration score. As s starts diverging from p the value of $I(s(e_i), p)$ increases, and the calibration measure should decrease, penalizing the fact that the experts are not answering corresponding to the stated percentiles. A simple test for this hypothesis uses the test statistic defined by equation 1.

The p-value of this hypothesis is defined as the calibration (or statistical accuracy) score

$$Cal(e_i) = Prob\{2MI(s(e_i), p) > r | H_{e_i}\},$$

where r is the value of the expression from equation (1) based on the observed values[7] $x_1, \ldots, x_M$. It is the probability, under hypothesis H_{e_i}, that a deviation at least as great as r should be observed on M realizations if H_{e_i} were true.

With a finite, relatively small number of questions, often s cannot equal p. Most of the times, they differ, because of, for example, M being an odd number. An even number of seed questions does not guarantee equality either, for example for the most commonly used number of questions, ten, an expert can achieve a maximum calibration score of 0.83 when $s = (0.1, 0.4, 0.4, 0.1)$[8]. This is important when comparing calibration scores. How different should calibration scores be to conclude that one is much better than another? The answer to this question is not straightforward. The following example illustrates an interesting situation which is slightly unrealistic, but not impossible.

On the right hand side of Figure 2, Expert e_2 gave their percentiles for ten seed questions. The left and right ends of each line correspond to the 5^{th} and the 95^{th} percentiles, respectively. The blue dots corresponds to the 50^{th} percentiles, and the crosses correspond to the realisations of the seed variable. The crosses are blue if they are captured within the 90% credible interval, and red is they fall outside this interval. In this example $s(e_2) = (0.1, 0.4, 0.4, 0.1)$ and expert e_2 achieves the maximum possible calibration score of 0.83. Expert e_1 gave exactly the same estimates for all the questions with the exception of four medians, which happened to coincide with the realisations of those variables (see the left hand side of the same figure). The empirical distribution of expert e_1 is $s(e_1) = (0.1, 0.6, 0.2, 0.1)$. Expert e_1 is thus penalised as an artefact of the way the empirical distribution is constructed and achieves what seems to be a much lower calibration score of 0.39.

This sort of examples are useful to understand what these differences in calibration scores can mean. In this case, both experts are well calibrated and the 0.44 difference between calibration scores should not be used to say that expert e_2 is much better calibrated than expert e_1. However, when calibration scores are low with one of them below 0.05, the former should be considered as an indication of better performance. For example, if the empirical distribution of an expert is $s(e_3) = (0.3, 0.2, 0.2, 0.3)$, their calibration score is with approximately 0.3 less then expert e_1 calibration, making it of order 10^{-2}. Expert e_3 placed most of the mass in the tails of the distribution, which should make one confident in considering them poorly calibrated.

[7] if s is equal to p, then $r = 0$ and $Cal = 1$.

[8] The minimum number of questions needed to obtain a calibration score of 1 is 20.

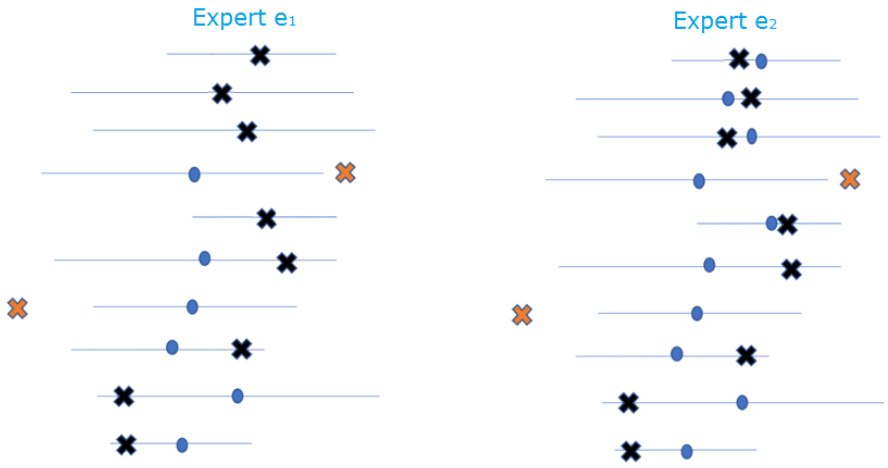
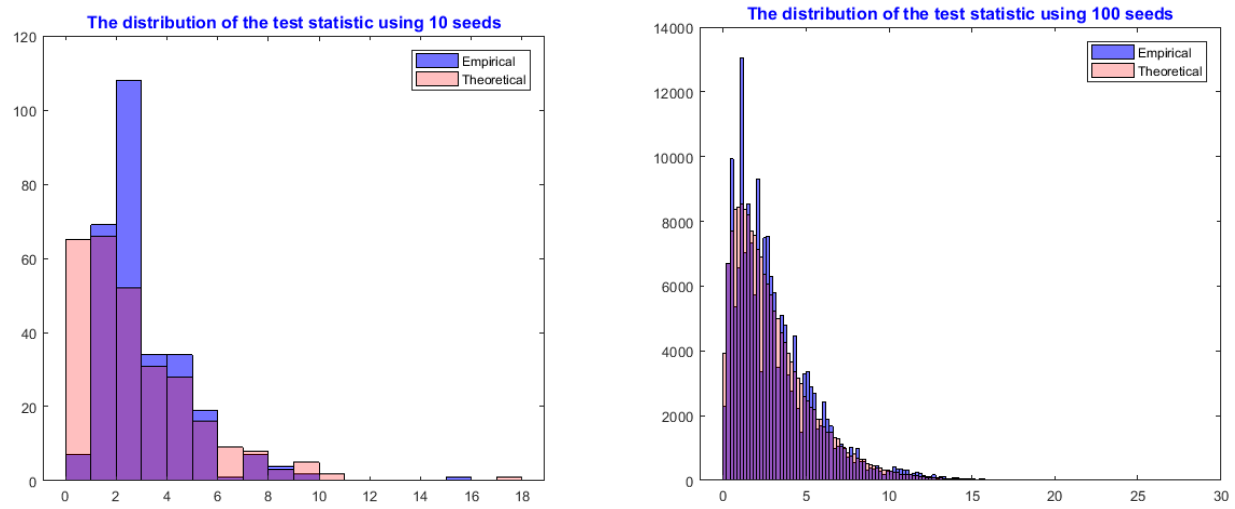


Figure 2: Two experts assessments on 10 seed questions. The starting and ending points of any line in this graph correspond to the 5^{th} and the 95^{th} percentiles, the blue dot corresponds to the 50^{th} percentile and the **x** corresponds to the true value of the true value of the seed variable. The blue dot is not visible when it coincides with the realisation.

The discussion above about the significance level aims to stress that any calibration above a certain threshold (often chosen to be the familiar 0.05 from classical statistical testing) may be considered a good calibration, and that the calibration score should not be used to differentiate among very fine levels of calibration, but provide rather indicative levels. This is, again, similar to conducting a hypothesis testing, where one does not compare different p-values concluding that a higher p-value produces more evidence to accept the null hypothesis, but one rather compares the p-values with the significance level of, say, 0.05. Consequently, the conclusion is either enough or not enough evidence to reject the null hypothesis H_0.

Another reason for not taking the actual calibration scores and the differences between them too seriously is the asymptotic nature of the test. For ten seed variables, the distribution of the test statistic is quite far from a chi squared distribution. This is illustrated in Figure 3, where the histogram of the test statistic is calculated empirically and compared with the histogram obtained by sampling from a chi squared distributed variable. The figure on the left hand side uses ten seed variables and the one on the right hand side uses 100 which is of course not feasible in practice. The right hand side histograms in Figure 3 agree not only on a visual level, but also when comparing them using statistical tests. We repeatedly used the two-sample Kolmogorov-Smirnov and the two-sample Cramer-Von Mises tests, and the null hypothesis that the data in the two samples came from the same continuous distribution, was not rejected in 98% of the cases.



(a) Ten seed variables.

(b) 100 seed variables

Figure 3: Histograms of $2MI(s(e_i), p)$ under null hypothesis that the inter-percentile interval containing the true value for each variable is drawn independently from probability vector p (blue), versus a random sample from a chi-square variable with 3 degrees of freedom (pink).

Calibration scores are absolute scores and can be compared across studies, if these studies use the same number of seed questions. In other words, before comparing calibration scores, it is appropriate to equalize the power of the different hypothesis tests by equalizing the effective number of seed variables. Because the calibration score uses the asymptotic distribution of the $2MI(s(e_i, P)$, we adjust the power by leaving s calculated on M questions but replacing $2M$ by $2M'$, with $M' < M$, M' representing the smallest number of seed variables. In this way we use all the M seed variables, but *pretend* that the relative information is based on M' rather than M variables. The ratio $\frac{M'}{M}$ is called the power of the calibration test (called *calibration power* in Excalibur). When the number of the seed questions increases, the calibration scores decrease, but are still distinguished if the numerical implementation of the scores are accurate enough. However, Cooke argued in [11] that the degree to which calibration scores are distinguished should be a model parameter one can optimise for, and that reducing the power may be important in situations when all experts are very poorly calibrated. When all experts are poorly calibrated (e.g., with calibration scores of order less than or equal to 10^{-4}, spanning three or more orders of magnitude) with one being better calibrated than the rest, all the weight may go to this one (still very) poorly calibrated expert. By reducing the power, several other combinations may be found optimal and the best of them should be used[9]. However, the accumulation of evidence since 1991, seems to suggest that in such cases an equally weighted combination of experts' distributions will be a much better choice than a combination based on optimising the calibration power.

To close our little parenthesis on the calibration power, we advise reducing the calibration power *only* for comparing calibration scores across studies with different number of seed questions.

To give an indication of the range of experts' calibration scores in professional applications, Figure 4 presents just over 300 of experts'calibration scores extracted from the studies collected in the Delft dataset, prior to 2006. The horizontal line corresponds to the calibration score of 0.05, and it is quite

[9]If you do elect to optimize weights using reduced calibration power, you should evaluate performance by introducing these weights as user weights and compare with other combinations *without* power reduction.

clear that the majority (73%) of individual calibration scores are below this level[10].

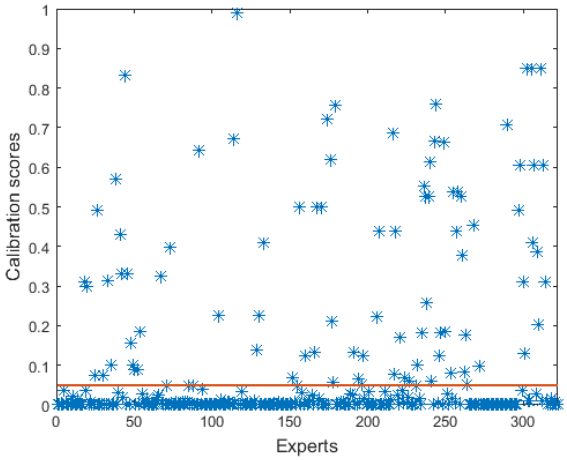


Figure 4: The calibration scores of 322 experts across the pre-2006 studies available in the TU Delft dataset. The red line denotes the 0.05 significance level.

A completely different picture will emerge when, in a later section (Subsection 4.1) of this chapter, we will investigate the magnitude and spread of combinations of experts. Figure 9 reveals the improved performance, in terms of the calibration score, of the combination of experts.

3.2.2 Informativeness

Along with the calibration score, experts' assessments are evaluated with respect to an information score. The information score is intrinsically connected with determining experts' distribution, given the three percentiles specified by the expert, as was constructed in Subsection 3.1. The information score reflects how informative the expert's distribution is with respect to the background measure used to construct the distribution. If that measure was the uniform distribution, then informativeness is calculated with respect to the uniform. However, when the intrinsic range spans many orders of magnitude, the log-uniform measure is used to construct the distributions. The informativeness of such a constructed distribution is then evaluated with respect to the log-uniform background measure as well.

Both background measures are available in Excalibur and the analyst should choose between the two measures. As a rule of thumb, when the range of experts' assessments for a question spans over four orders of magnitude, then it is advised to use a log-uniform background measure[11].

The background measure is assumed, for now, to be the uniform distribution over the intrinsic range $[L^*, U^*]$

$$U(x) = \frac{x - L^*}{U^* - L^*}, \text{ for } L^* \leq x \leq U^*.$$

One can derive the probability that an uniform random variable with distribution U lies within each of the inter-percentile intervals. Experts assessments with respect to the uniform background measure

[10] Similar pictures presented in a slightly different format are shown in [9].

[11] There is no theory behind the choice of the background measure. It is chosen on the basis of experiences and can later be subjected to sensitivity analysis.

for each of the four inter-percentile intervals thus yield

$$\begin{aligned}
r_1 =& U(q_5^i) - U(L^*) = \frac{q_5^i - L^*}{U^* - L^*}, \text{ for } x \in [L^*, q_5^i], \\
r_2 =& U(q_{50}^i) - U(q_5^i) = \frac{q_{50}^i - q_5^i}{U^* - L^*}, \text{ for } x \in (q_5^i, q_{50}^i], \\
r_3 =& U(q_{95}^i) - U(q_{50}^i) = \frac{q_{95}^i - q_{50}^i}{U^* - L^*}, \text{ for } x \in (q_{50}^i, q_{95}^i], \\
r_4 =& U(U^*) - U(q_{95}^i) = \frac{U^* - q_{95}^i}{U^* - L^*}, \text{ for } x \in (q_{95}^i, U^*].
\end{aligned}$$

With respect to expert's distribution $F(\cdot)$, let

$$\begin{aligned}
f_1 &= F(q_5^i) - F(L^*) = 0.05, \\
f_2 &= F(q_{50}^i) - F(q_5^i) = 0.45, \\
f_3 &= F(q_{95}^i) - F(q_{50}^i) = 0.45, \\
f_4 &= F(U^*) - F(q_{95}^i) = 0.05,
\end{aligned}$$

The information score of expert e_i for question j is then determined by

$$I_j(e_i) = \sum_{k=1}^{4} f_k \ln \frac{f_k}{r_k}.$$

Writing the information score in terms of expert's assessments and the intrinsic range gives

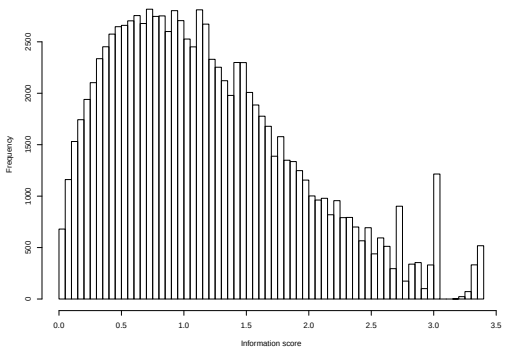
$$I_j(e_i) = 0.05 \ln \frac{0.05(U^* - L^*)}{q_5^i - L^*} + 0.45 \ln \frac{0.45(U^* - L^*)}{q_{50}^i - q_5^i} + 0.45 \ln \frac{0.45(U^* - L^*)}{q_{95}^i - q_{50}^i} + 0.05 \ln \frac{0.05(U^* - L^*)}{U^* - q_{95}^i},$$

which can be re-written somewhat more compactly

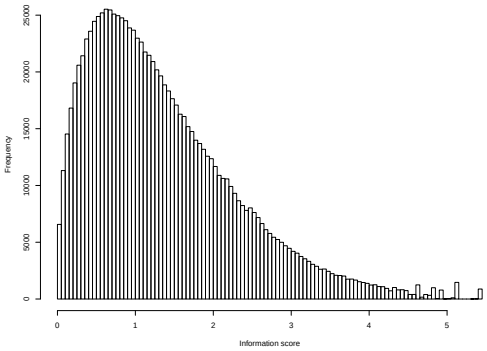
$$I_j(e_i) = 0.05 \ln \frac{0.05}{q_5^i - L^*} + 0.45 \ln \frac{0.45}{q_{50}^i - q_5^i} + 0.45 \ln \frac{0.45}{q_{95}^i - q_{50}^i} + 0.05 \ln \frac{0.05}{U^* - q_{95}^i} + \ln(U^* - L^*), \quad (2)$$

as in [11]. The information score is a strictly positive function, which can take, in principle, arbitrarily large values. It can be observed in (2) that the closer expert's assessments are, the larger $I_j(e_i)$ will be. One would wonder, however, how large can the information score be, in practice, and how does the distribution of information scores looks like. We have investigated the behaviour of information scores from simulated data, as well as from expert elicitations data from previous studies.

Firstly, the simulations have been performed assuming an intrinsic range of $[0, 100]$, as for the elicitation of percentages, and are depicted in Figure 5a. Only integer values have been assumed for the experts' assessments, in order to simplify calculations. Furthermore, simulations of information scores over an intrinsic range of $[0, 1000]$ and the histograms can be found in Figure 5b.



(a) Intrinsic range of $[0, 100]$.



(b) Intrinsic range of $[0, 1000]$.

Figure 5: Histograms of information scores over an intrinsic range of $[0, 100]$ (5a) and $[0, 1000]$ (5b).

While for an intrinsic range of $[0, 100]$, information scores obtained are not larger than 3.5, when the intrinsic range extends to $[0, 1000]$, the maximum observed information score is around 5.8. Repeated simulations have produced similar results for the information scores. As mentioned beforehand, the intrinsic range of $[0, 100]$ corresponds to integer percentage assessments, whereas the intrinsic range is $[0, 1000]$ corresponds to eliciting percentages up to the first decimal.

The information score of an expert over all seed questions is defined as the average of information scores

$$I(e_i) = \frac{1}{M} \sum_{j=1}^{M} I_j(e_i).$$

Notice that the information score can be computed for the seed questions as well as for the questions of interest, whereas the calibration score can only be computed for the seed questions. Moreover, note that, while the calibration score of each expert is computed independently of other experts assessments, the distribution of experts, and hence the information score depends on all experts assessments, which makes informativeness a group dependent measure.

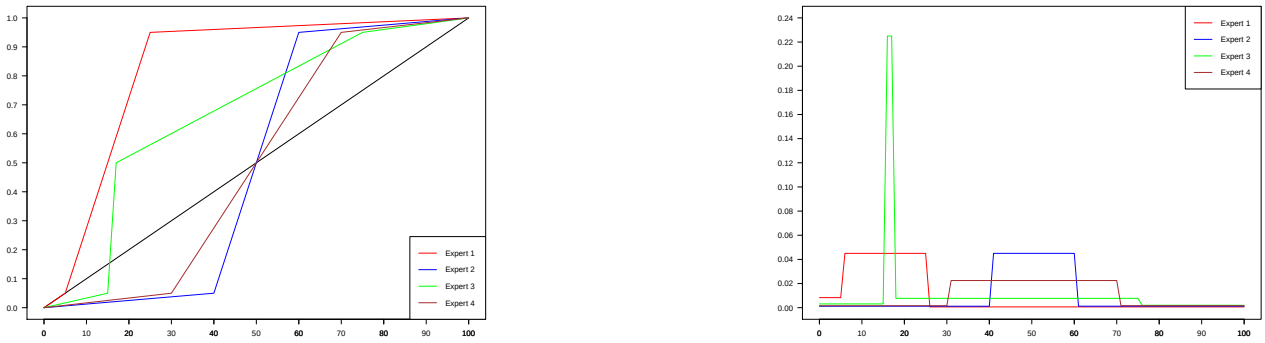
Finally, it should be once more emphasized that the information score reflects how informative expert's distribution is with respect to the background measure, which is usually assumed to be the uniform distribution. While the information score could be thought as associated with how spread the expert's assessments are, that is, in fact, not quite the case. Consider the following examples of experts assessments, as depicted in the table below.

	5%	50%	95%	Information score
Expert 1	5	15	25	1.21
Expert 2	40	50	60	1.14
Expert 3	15	17	75	1.15
Expert 4	30	50	70	0.55

Table 2: Example of four experts percentage assessments.

Even though Expert 3 assessments are quite spread, the percentiles result in a skewed distribution, which is quite informative with respect to the background measure. The information score is almost the same as for Expert 2, where the probability mass function is concentrated between 40 and 60. There is a significant difference in the information score between experts $1, 2, 3$ and Expert 4. Whereas

the highest information score is attained by Expert 1, the difference with Expert 2 and 3 is not that large. The cumulative distribution function and the probability density function of the 4 experts are depicted in the Figure 6.



(a) (b)

Figure 6: Cumulative distribution functions (1a) and probability density function (1b) for four experts whose assessments are included in Table 2.

The information score can now be heuristically tied with expert's distribution, namely with how discrepant expert's distribution is from the uniform distribution. For example, it is quite obvious that Expert 4 (brown) is the least discrepant from the uniform distribution (black). Similarly, Expert 1 (red) is the most discrepant and has therefore the highest information score among the 4 experts. Additionally, it is quite hard to evaluate and compare the information scores of experts 2 (blue) and 3 (green). Their cumulative distribution functions are quite distinct, whereas the information scores are the same.

Obviously, the higher the information score, the more informative the expert is and an expert with high information score is preferred over an expert with a low information score, assuming they have the same calibration. One can however wonder when is an information score low, that is, when is an expert considered uninformative. Of course an expert whose assessments coincide with the percentiles of the uniform distribution will have an information score of zero. When the assessments differ from the uniform percentiles, one could think that a test can determine whether the differences are statistically significant or not. A number of tests can quantify the difference between two distributions. Cramér-von Mises test, for example, evaluates the integrated quadratic difference between two distributions. The distributions of all four experts whose assessments are included in Table 2 are statistically significantly different from the uniform distribution, according to the Cramér-von Mises test, when using 100 or 1000 observations. An inspection of several examples lead to the conclusion that information scores as low as 0.15 lead to the rejection of the null hypothesis that expert's assessments come from an uniform distribution. Furthermore, an assessment of $10, 35$ and 90 for the three percentiles leads to an information score of 0.1, and the p-value of the Cramér-von Mises test is 0.21. However, it should be born in mind that these results dependent on the intrinsic range, which has been chosen $[0, 100]$.

Another question that might arise is whether information scores are significantly different from a statistical point of view. This is nicely exemplified with the four experts assessments above, that is, whether an information score of 1.15 is significantly higher than an information score of 0.55. Cramér-von Mises test between Expert 2 distribution (blue) and Expert 4 (brown) distribution leads to a p-value of 0.25, whereas the p-value for the test between Expert 3 and Expert 4 is less than 2.2×10^{-16}. This shows that determining statisticall significant differences between information scores is arguably an important question that would need, nonetheless, more refined metrics.

To get an idea about the possible values and spread of information scores from expert elicitation data, we plotted information scores obtained by the experts taking part in the studies collected in the Delft dataset, prior to 2006. All scores are between 0.25 and 3.81 and half of these scores are larger than 1.47.

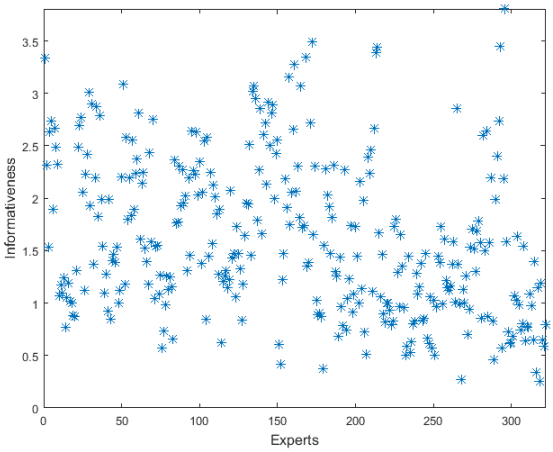


Figure 7: The information scores of 322 experts across the pre-2006 studies available in the TU Delft dataset.

3.3 Combined scores to form global and item weights

Measuring performance serves multiple purposes. Apart from differentiating between experts' performance, scores can be used to form weights which will then be used to construct a differentially weighted linear combination of distributions over the target variables. These mathematically aggregated distributions are considered to be the rational consensus distributions. They can be thought of as virtual experts whose "opinions" incorporate all experts' opinions, weighted according to their validity. An equally weighted linear combination is another virtual expert. These virtual experts can be treated as any other expert and their constructed opinions can be scored in the same way as experts' opinions. The final aim of this exercise is to find the virtual expert who performs the best. Before discussing the different virtual experts, let us return to how the scores presented in the previous sections can be combined and used as weights.

CM accounts for both calibration score and informativeness and proposes a combined score, which is the product of the calibration and the information score and it uses a cutoff level α, below which calibration scores are undesirable. The calibration score is often described as being a *fast* function, which means that its value changes quickly with the addition of every seed question and its associated response. Informativeness, on the other had is said to be a *slow* function, which means that it is less sensitive to small change in the number of questions. When multiplied, the calibration will dominate the value of the combined score, therefore CM values the calibration score more in comparing experts. This is also intuitively desired, as one would not prefer an informative over a poorly calibrated expert, which reflects only overconfidence. The combined score for expert i is given by

$$CS(e_i) = Cal(e_i) \cdot I(e_i) \cdot \mathbf{1}_{\alpha}\left(Cal(e_i)\right),$$

for $i = 1, \ldots, N$ and $\alpha \geq 0$; the weight of expert i will be proportional to their score

$$w_i = \frac{CS(e_i)}{\sum_{k=1}^{N} CS(e_k)}, \tag{3}$$

for $i = 1, \ldots, N$. Experts with calibration scores below α will receive weight zero and their judgements will not be directly used in the final linear combination of opinions. However, all experts' assessments determine the support of all variables, therefore all experts contribute to virtual expert's distribution. A value α larger than zero ensures that the weights are asymptotically strictly proper. For detailed information on scoring rules, see [11].

Note that the information score is actually calculated per question (item), and then averaged across all questions. This suggests that a combined score can be computed for each expert and seed variable

$$CS_j(e_i) = Cal(e_i) \cdot I_j(e_i) \cdot \mathbf{1}_\alpha \left(Cal(e_i)\right),$$

for $j = 1, \ldots, M$ and $i = 1, \ldots, N$. The information score $I_j(e_i)$ denotes how informative expert i is on question j. This combined score leads to the weights

$$w_i^j = \frac{CS_j(e_i)}{\sum_{k=1}^{N} CS(e_k)},$$

for expert i and question j, where $i = 1, \ldots, N$ and $j = 1, \ldots, M$. The weights are called "item weights", and they are calculated per item, per expert. Thus an expert can receive different weights for each seed variable. It should be born in mind, however, that the calibration score remains the same for each seed variable, therefore dramatic changes in the item weights should not be expected, especially for experts with very low calibration scores. Furthermore, these weights are potentially more attractive, as they allow an expert's weight to be higher or lower for individual items/questions/variables, according to their knowledge about each question. Knowing less is usually translated into choosing percentiles further apart, and by doing that, lowering the information score for that item. The combined score for expert i is then different for each question j.

In contrast, the weights in (3) are referred to as global weights. For both global and item weights, calibration dominates over informativeness; the information score serves to modulate between more or less equally calibrated experts, with one exception, which will be discussed in the next section.

4 Post elicitation

As mentioned in the previous section, the performance-based weights are used in CM to combine experts' judgements using a linear pool. The aggregation of expert distributions is usually referred to as a Decision Maker (DM). We reiterate that a DM in this context is a mathematically calculated distribution which corresponds to a virtual expert. The real decision maker would adopt this distribution as their own, representing rational consensus.

The performance-based weights distinguish between global and item weights, which lead to two DMs, the Global Weight Decision Maker (GWDM) and the Item Weight Decision Maker (IWDM). Moreover, different GWDM and IWDM combinations can be obtained by choosing different values for the cutoff α parameter. The α values which lead to distinct GWDM and IWDM are, in fact, the calibration scores of the experts. Using α equal to the smallest calibration score results in the combination of all experts' assessments into the DMs. Choosing the next larger calibration value

translates into forming DMs using all but one expert. Choosing the largest calibration as a cutoff level translates into DMs which are the same as the best calibrated expert. We distinguish between GWDM and optimized GWDM; GWDM uses $\alpha = 0$ (but it is essentially the same as using α equal to the smallest calibration which is usually larger than zero), therefore accounts for all experts' assessments, whereas optimized GWDM uses α such that the combined score of GWDM is maximum. Similarly, we have IWDM and optimized IWDM.

For the IWDM, the weights are different for each question, hence IWDM uses a set of weights. If GWDM uses a vector of weights, IWDM uses a matrix of weights, where each row represents the vector of weights corresponding to each question, of interest or calibration. Concluding, for GWDM, experts' weights are constructed exclusively based on the calibrations questions. IWDM uses, alternatively, weights that are constructed both on calibration questions, as well as on questions of interest. More specifically, the weights for each question of interest is computed using the calibration score and experts' information score of the question of interest.

The aggregation of expert distributions can also be done by using equal weights, which gives the equal-weight decision maker, denoted by EWDM.

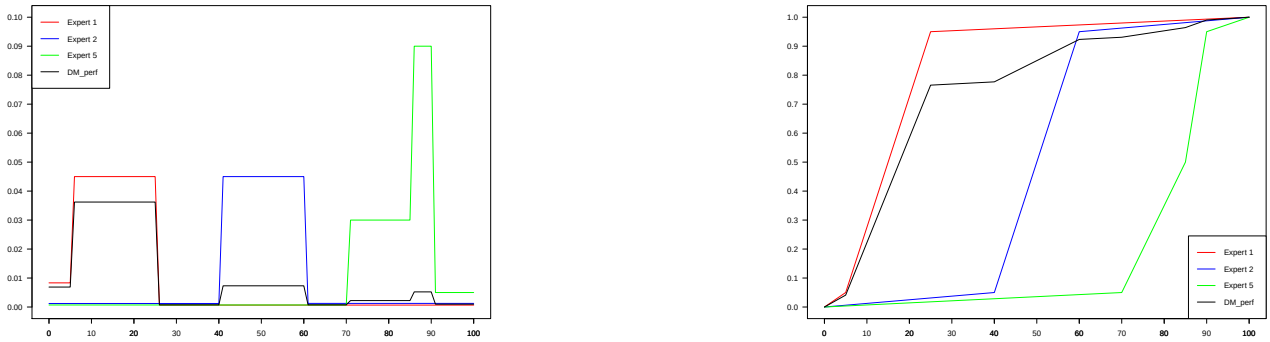
Finally, it is worth mentioning that even though CM aggregates experts' distributions, other approaches are possible, such as aggregating experts' percentiles. A discussion between emerging differences in DM's distributions as well as DM's performance when aggregating distributions versus percentiles has been addressed in [9].

4.1 DMs and their scores

The final, and perhaps most important use of the performance based scores is to evaluate the performance of the many DMs and be able to choose the best one, as measured by performance, which is expressed in terms of the combined score defined in (3.3). This is arguably the only valid way of motivating one choice of aggregation over others available.

DM distributions for the questions of interest are used as a final output of the elicitation study. DM can however be regarded as an expert itself, albeit virtual, and therefore one can derive its assessments also for the seed questions. These assessments can be evaluated with respect to the calibration and information score, just as for any other expert. The calibration score and informativeness of DM can be compared to single experts' performance. Moreover, both GWDM and IWDM can be optimised by choosing the value of α which maximizes the combined score of the resulting DM. The combined scores of GWDM, IWDM and EWDM can be compared; the combined scores are available in Excalibur and they are a standard output of CM studies.

Excalibur also allows the users to export the DMs percentiles, which can then be used to derive the DMs distribution and plot it along with the other experts' distributions. Figure 8 presents the cumulative distribution functions and the density functions of three experts along with the GWDM. Expert 1 and 2's assessments can be found in Table 2, whereas Expert 5's assessments are 70, 85 and 90. The normalized weights are 0.8, 0.15 and 0.05, for Expert 1, Expert 2 and Expert 5 respectively.



(a) (b)

Figure 8: Cumulative distribution functions (8a) and probability distribution function (8b) of three experts along with DM.

DMs distributions as the ones above can be evaluated in terms of the performance scores. The range of DMs' calibration scores in professional applications can be seen in Figure 9, where the scores for EWDMs, the optimised GWDM and optimised IWDMs of 74 studies from the Delft dataset are shown[12]. The horizontal line corresponds to a calibration score of 0.05 and, contrary to the individual scores (see Figure 4), the minority (6.7%) of DMs' calibration scores are below this level.

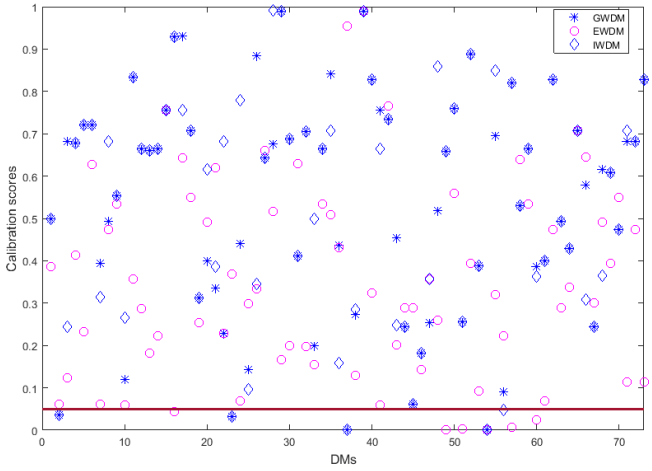


Figure 9: The calibration scores of 222 DMs (74 EWDM, 74 optimized GWDM and 74 optimized IWDM) across studies available in the Delft dataset. The red line denotes the 0.05 significance level.

We consider separately the EWDMs and the GWDMs and analyse their performance. This evaluation of the performance is usually referred to as an in-sample validation. That is, the performance of DMs is evaluated on the questions that were used to determine the DMs. Figure 10 shows the GWDM scores on the x-axis and the EWDM scores on the y-axis. The horizontal and vertical lines indicate the 0.05 significance level, which can be regarded as a threshold for the calibration score. Very rarely one combination is below this threshold while the other is above. The main diagonal represents equal performance from the calibration view point, and again the two DMs are equally calibrated in very few

[12]There are 79 professional studies for which the DMs' scores were reported in [9] and [13]. We were able to identify, re-run and reproduce scores for 74 of them.

cases. Given the discussion in Section 3.2.1 about small differences in the calibration scores, we may consider a region around the main diagonal, where we cannot distinguish between calibration scores (see the area bounded by dashed lines in Figure 10). We consider only the studies which used at least ten seed variables (63 out of the 74 used above). It results that 41.27% of the scores fall within that region, and in 50.79% of the cases, the GWDM calibration score is clearly better than the EWDM's calibration score. In only 7.94% of the studies was the EWDM's calibration better than the GWDM's. Some would consider this as irrefutable evidence that the optimised GWDM combination is either as good or better than the EWDM.

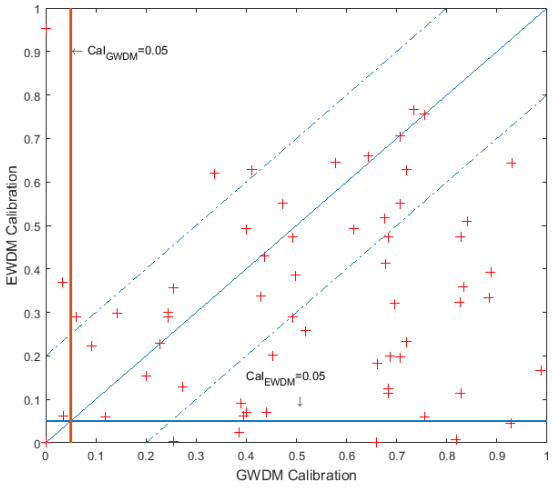
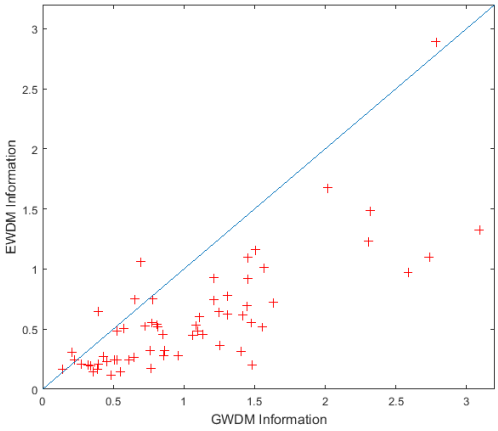
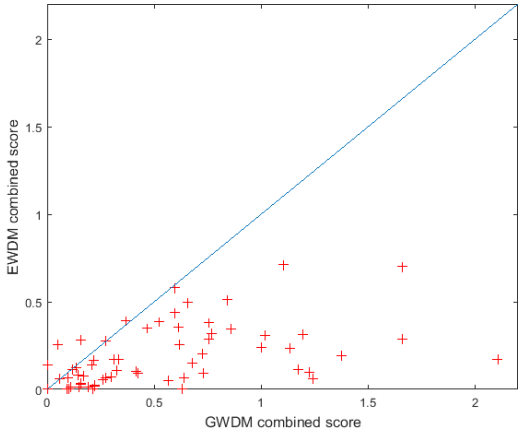


Figure 10: Pairs of 63 calibration scores for optimised GWDMs versus EWDMs across the studies from the Delft dataset using at least 10 seed questions.

The picture changes dramatically when we consider the information scores. These are shown in Figure 11a. The vast majority of the scores are higher for the GWDM, pattern which is repeated when looking at the combined score (ses Figure 11b).



(a) Pairs of 63 information scores for optimised GWDMs versus EWDMs across the studies from the Delft dataset using at least 10 seed questions.



(b) Pairs of 63 combined scores for optimised GWDMs versus EWDMs across the studies from the Delft dataset using at least 10 seed questions.

Figure 11: Optimised GWDMs versus EWDMs information scores (11a) and combined scores (11b) across the studies from the Delft dataset using at least 10 seed questions.

Item weights sometimes improve over global weights. In the same dataset of 74 professional studies (that is all studies we initially considered and not just those with more than ten seed questions), the informativeness of the IWDM is larger than the informativeness of GWDM in 57.1% of the studies, IWDMs' calibrations is only 20.6% of the times larger than that of the GWDMs. IWDMs' combined scores are larger that the PWDMs score for 41.3% of the studies.

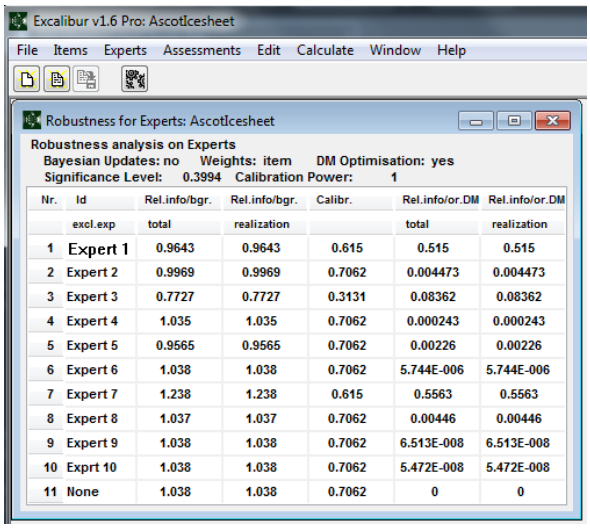
Of course the above analysis only serves as an in-sample validation of our intuition that performance based combinations are at least as, or more calibrated than, and certainly more informative than the equally weighted combinations. Out of sample validation studies confirming the same results have been published in [9]. Add reference, when available of Tom and Roger's chapter on random experts. An ultimate proof that the observed differences in scores are indeed important would be the possibility to use the different combinations in their respective decision problems and confirm that such differences in performance result in differences in decisions. Unfortunately this does not seem to be possible. Maybe future SEJ studies should follow up with such an analysis.

4.2 Optimised DMs

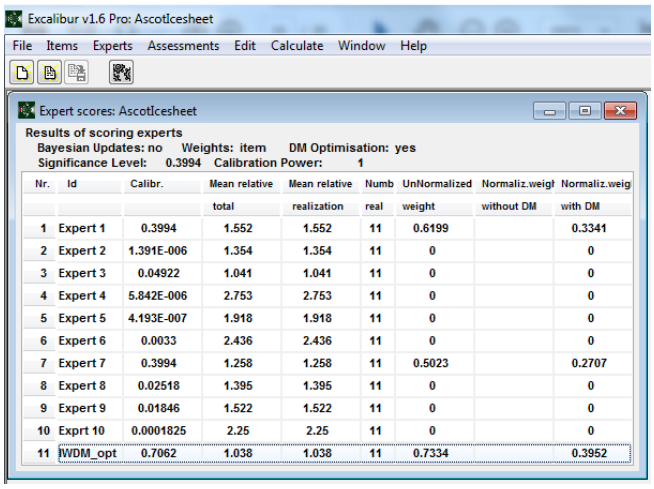
Optimized performance-based DM's have been considered in the analysis of the professional studies in the previous sub-section. Even though clarified and discussed with every opportunity, the optimisation procedure (which ensures that we are using a proper scoring rule, at least asymptotically) seems to still make analysts and young facilitators nervous, because this procedure is perceived as excluding experts (by assigning them zero weight) from the final combination of judgements.

Weight zero does ***not*** mean value zero. Most of the times this means that those experts' knowledge was already contributed by other experts. The value of un-weighted experts is seen in the robustness of the answers against loss of experts. Excalibur has the option to perform such a robustness analysis and to recalculate the scores that would have been obtained if experts were completely excluded (rather than weighted zero) from the analysis. One of the very important contributions experts make is in determining the support of the variables. All experts contribute to these ranges and, when one expert's

assessments are not taken into account both the calibration scores and the information scores of the remaining experts may change. This sometimes results in a worse calibrated DM. Below is one such example from the ice sheet application published in Nature Climate Change.



(a) Robustness analysis for experts in the ice sheet application detailed in [3].

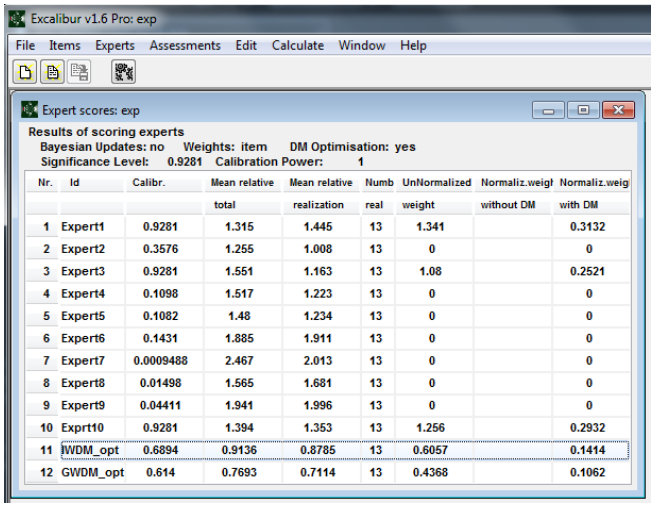


(b) Item Weights Decision Maker Optimised combination for the ice sheet application detailed in [3].

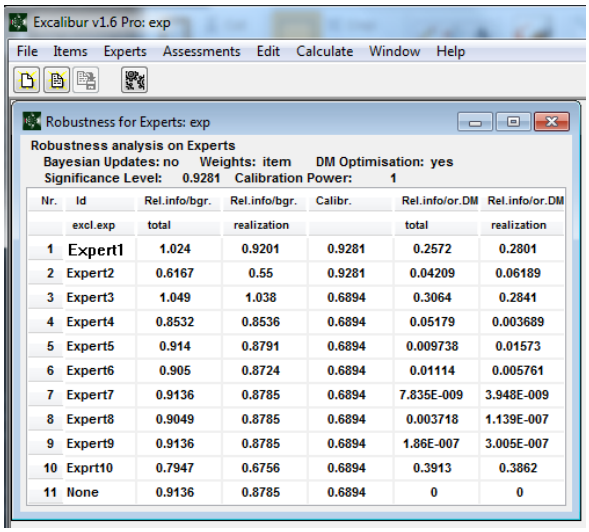
Figure 12: Weight zero does ***not*** mean value zero.

Figure 12a shows a snapshot from Excalibur obtained when clicking on the Robustness (experts) button. Row i corresponds to the scores that would have been obtained if Expert i were not part of the expert panel. The last row shows the scores obtained when all experts are involved. Figure 12b shows the optimal combination of experts when item weights are assigned. Only experts 1 and 7 are weighted in the optimal combination, however, the robustness analysis shows that if one of them is removed from the analysis, there is only a slight, irrelevant (given the number of seeds) decrease in calibration. However, if expert 3, whose weight is zero in the combination, is completely removed from the panel, the calibration drops from 0.7 to 0.3.

In the example above, the optimised IWDM (and GWDM) uses a combination of the two best calibrated experts from the panel. In this case, as in many other cases, the optimised combination affords a higher calibration score than the two experts individually. Even though this seems intuitive it is not always the case. Hence, there are cases when the optimised DM performs worse than the best expert. The reason behind this is the following: when the optimised DM is used, the optimisation is based on the calibrations scores alone. When there are two (or more) experts with the same best calibration, the optimised DM takes them all in the final combination, independent of the differences between their information scores. Their respective weights will be differentiated using their information scores, but this may still result in "optimal" DM whose calibration (or even combined score) is worse that the best experts' calibration. An explanation for this behaviour may be what Cooke calls a "peculiar" sort of correlation, which "has never been observed in practice" (pg 197 from [11]). However, since the book was written, this phenomenon was observed in practice, even though in a different context than the one explored in [11]. We conjecture that these situations occur when the experts answers are correlated in a certain way; however, it is not clear yet what this "certain way" may mean. In a recent application detailed in [16], there were three experts who received the best possible calibration (0.928) score to be obtained on 13 seeds (which is the number of seeds used for this elicitation). Even though it is common for two (or even three) experts to have the same calibration score, it is rather unusual for three of the experts to have the same best calibration score.



(a) Optimised Decision Makers for a recent defence application detailed in [16].



(b) Robustness analysis for experts in a recent defence application detailed in [16].

Figure 13: The optimised DM is not alway optimum.

The combination of the three best experts (experts 1, 3, and 10) leads to a poorer performance for both the GWDM, and the IWDM. However, taking one of the best calibrated experts out of the combination, restores the score of the DMs to equal that of the best calibrated experts. This is true only when we take expert 1 out of the analysis, as shown in Figure 13b. The dependence structure between these three experts is depicted in Table 3.

	Expert 1	Expert 3	Expert 10
Expert 1	1	-0.07	0.55
Expert 3	-0.07	1	-0.24
Expert 10	0.55	-0.24	1

Table 3: Correlation matrix of the three best calibrated experts.

Expert 1's assessments seem to be positively correlated with those of expert 10 and uncorrelated with those of expert 3. The two experts whose combination would be better calibrated seem to be slightly negatively correlated (even though on 13 samples this correlation is not significantly different than zero). The correlation values were calculated based on the medians of the experts rather than all three quantiles, in a similar way to the calculations performed in other studies that investigated dependence between experts' assessments (see [20, 29]).

There is an unequivocal need for more research into these issues and more awareness of the possibilities.

5 Closing remarks

This chapter draws attention to some (maybe less discussed) aspects of the theoretical background of CM. One of these aspects is the misinterpretation of the differences between calibration scores. Another one regards the intuitive relation between the wideness of the uncertainty bounds and the information score. It also aims to provide a thorough overview of practical aspects and choices that practitioners face before and during the elicitation process.

"The qualifier *structured* means that expert judgment is treated as scientific data, albeit scientific data of a new type" [11]. The name of the method itself, the "Classical Model" emphasizes the close connections with classical statistics. Furthermore, the method has auspiciously laid grounds for further statistical endeavors, such as goodness of fit and validation. If one regards the DM's performance as a goodness of fit measure, then the optimized DM's distributions are constructed such that they best fit the expert data. The evaluation of the performance-based DM has also been referred to as an in-sample validation. Furthermore, notable effort has been undertaken [9] to validate CM. The scores of performance-based DM's are hence evaluated on questions that have not been used to construct DM's distributions.

Despite the demanding nature of CM, the results from the studies show that the effort of forming and using performance-based combination of experts distributions is definitely worthwhile.

References

[1] W. Aspinall. Expert judgement elicitation using the classical model and excalibur. *Briefing notes*, 2008.

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

[3] W. Aspinall and J. L. Bamber. An expert judgement assessment of future sea level rise from the ice sheets. *Nature Climate Change*, 3:424, 2013.

[4] E. Beshearse, B.B. Bruce, G.F. Nane, R.M. Cooke, W. Aspinall, T. Hald, S.M. Crim, P.M. Griffin, K.E. Fullerton, S.A. Collier, K.M. Benedict, M.J. Beach, A.J. Hall, and A.H. Havelaar. Source attribution of illnesses transmitted commonly by food and water in the united states using structured expert judgment. *Submitted to Emeging Infectious Diseases*, 2018.

[5] F. Bolger, A. Hanea, A. O´ Hagan, O. Mosbach-Schulz, J. Oakley, G. Rowe, and M. Wenholt. Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment. *EFSA Journal*, 12(6):Parma, Italy, 2014.

[6] F. Bolger and G. Rowe. The aggregation of expert judgment: Do good things come to those who weight? *Risk Analysis*, 35:5–11, 2015.

[7] F. Bolger and G. Rowe. There is Data, and then there is *Data*: Only Experimental Evidence will Determine the Utility of Differential Weighting of Expert Judgment. *Risk Analysis*, 35:21–26, 2015.

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

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

[10] R. Cooke, B. Kraan, and L. Goossens. Rational consensus under uncertainty: Expert judgment in the EC-USNRC uncertainty study. In Kjell Andersson, editor, *NEI-SE–308*. Sweden, 1999.

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

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

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

[14] S. French. Group consensus probability distributions: a critical survey. *Bayesian Statistics. Eds Bernardo J.M., De Groot M.H., Lindley D.V. and Smith A.F.M.*, Elsevier North Hollan:182–201, 1985.

[15] A.M. Hanea and G.F. Nane. Calibrating experts' probabilistic assessments for improved probabilistic predictions. *Safety Science*, 118:763 – 771, 2019.

[16] V. Hemming, A.M. Hanea, Armstrong. N, and M.A. Burgman. Improving expert forecasts in reliability. application and evidence for structured elicitation protocols. *Quality and Reliability Engineering International*, accepted in September, 2019.

[17] Victoria Hemming, Terry V. Walshe, Anca M. Hanea, Fiona Fidler, and Mark A. Burgman. Eliciting improved quantitative judgements using the idea protocol: A case study in natural resource management. *PLOS ONE*, 13(6):1–34, 06 2018.

[18] Einhorn H.J., R.M. Hogarth, and E. Klempner. Quality of group judgment. *Psychological Bulletin*, 84(1):158, 1977.

[19] Sandra Hoffmann, Brecht Devleesschauwer, Willy Aspinall, Roger Cooke, Tim Corrigan, Arie Havelaar, Frederick Angulo, Herman Gibb, Martyn Kirk, Robin Lake, Niko Speybroeck, Paul Torgerson, and Tine Hald. Attribution of global foodborne disease to specific foods: Findings from a world health organization structured expert elicitation. *PLOS ONE*, 12(9):1–26, 09 2017.

[20] M.J. Kallen and R.M. Cooke. Expert aggregation with dependence. In *In, Proceedings of the 6th International Conference on Probability Safety and Management*, pages 1287–1294, 2002.

[21] S. Kullback and R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[22] R.P. Larrick and J.B. Soll. Intuitions about combining opinions: Misappreciation of the averaging principle. *Management science*, 52(1):11–27, 2006.

[23] Anca M. Hanea, Marissa McBride, Mark Burgman, and Bonnie Wintle. The value of performance weights and discussion in aggregated expert judgments. 38, 03 2018.

[24] O. Morales-Napoles, A.M. Hanea, and D.T.H Worm. Experimental results about the assessments of conditional rank correlations by experts: Example with air pollution estimates. In R.D.J.M. Steenbergen, P.H.A.J.M. van Gelder, S. Miraglia, and A.C.W.M. Vrouwenvelde, editors, *Safety, Reliability and Risk Analysis: Beyond the Horizon*, pages 1359–1366. CRC Press, 2014.

[25] Van Elst NP. Betrouwbaarheid beweegbare waterkeringen [reliability of movable water barriers]. In *Delft University Press*. WBBM report Series 35, 1997.

[26] 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.

[27] J. Quigley, A. Colson, W. Aspinall, and R.M. Cooke. Elicitation in the classical model. In Dias L.C., Morton A., and Quigley J., editors, *Elicitation: The Science and Art of Structuring Judgement*, pages 15 – 36. International Series in Operations Research & Management Science, Springer, Cham, 2018.

[28] C. Werner, A.M. Hanea, and O. Morales-Napoles. Eliciting multivariate uncertainty from experts: Considerations and approaches along the expert judgement process. In Dias L.C., Morton A., and Quigley J., editors, *Elicitation: The Science and Art of Structuring Judgement*, pages 171 – 210. International Series in Operations Research & Management Science, Springer, Cham, 2018.

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