

*Investigate**Discuss**Estimate**Aggregate* for structured expert judgement

A.M. Hanea^{a,1} M.F. McBride^{a,b} M.A. Burgman^a B.C. Wintle^a F. Fidler^c
L. Flander^a C.R. Twardy^d B. Manning^a S. Mascaro^e

^a *University of Melbourne, Australia*

^b *University of Helsinki, Finland*

^c *RMIT University, Australia*

^d *George Mason University, United States*

^e *Bayesian Intelligence, Australia*

Abstract

This study presents the results of an approach to predicting the outcomes of geopolitical events, which we term the IDEA protocol. Participants investigate the background and causal factors behind a question, predict the outcome, discuss their thinking with others, and then make a second, private and anonymous judgement of the probability of the event that is subsequently aggregated mathematically. The method performed well compared to an equally weighted linear pool and a prediction market and is relatively simple to implement. The results indicate the value of discussion to remove arbitrary linguistic uncertainty and to share and debate knowledge, thereby improving judgements. Weighting judgements by prior performance using Cooke's method improved group judgements. Even though some of the results are not statistically significant, the study may not have had sufficient power to detect some important effects. Nevertheless, the results help us formulate conjectures, which can be further investigated.

Keywords: Probability forecasts, Performance, Elicitation, Aggregation, Structured Protocol

1. Introduction

The use of expert judgement in forecasting and decision making is well established (French, 2012). The role of experts can vary from reviewing reports or quantifying uncertain model parameters, to making final decisions. Nevertheless, the elicitation of expert judgements is often unstructured (e.g. Morgan, 2014). This results in judgements that are not only unreproducible

¹Correspondence to: A.M.Hanea, Centre of Excellence for Biosecurity Risk Analysis, University of Melbourne, Parkville, VIC 3010, Australia, email: anca.hanea@unimelb.edu.au

and unreliable, but sometimes disastrous (e.g. Shine, 2010; Rabinowitz, 1998; Bergstrom et al., 2009).

Compelling research demonstrates that cognitive biases prevail during unstructured elicitations (Kahneman & Tversky, 1984; Slovic, 1999), and that structured approaches provide substantial improvements (Cooke & Goossens, 2008; Aspinall, 2010). Meaningful elicitation and aggregation require a framework that combines foundational elements of decision theory and mathematics with procedures to mitigate the most pervasive and debilitating psychological and contextual frailties of expert judgement identified in the cognitive psychology literature (O’Hagan et al., 2006; Keeney & von Winterfeldt, 1991).

This paper describes recent developments in a structured protocol (based broadly on other procedures) for gathering and combining expert judgements of quantitative estimates of uncertain model parameters called IDEA. IDEA encourages experts to Investigate, Discuss, and Estimate, and it concludes with a mathematical Aggregation of judgements scheme. IDEA combines features of three major classes of protocols for eliciting quantitative expert knowledge.

The first, classical approach to obtaining a single estimate from several experts employs behavioural aggregation (e.g. O’Hagan, 2005), in which the experts typically meet face-to-face and agree on a judgement. In behavioural aggregation, discussion and consensus seeking are often practised together. In a structured behavioural aggregation, a facilitator encourages interaction and discussion. The main advantage of this approach is that experts share and debate their knowledge. Moreover, during a workshop the facilitator can ensure that experts have the same understanding of what is being asked. If they apply procedural rules for fairness skilfully, facilitators can anticipate and reduce the impact of some of the more serious sources of psychological bias, including dominance, anchoring and halo effects (e.g. Phillips & Phillips, 1907; Meyer & Booker, 1991). Nevertheless such interaction is prone to biases including overconfidence, polarisation of judgements and groupthink (e.g. Hinsz et al., 1997). When experts strongly disagree, attempts to impose agreement may mask the group’s diversity of opinion.

The second broad class of protocols use mathematical methods to aggregate judgements. In these approaches, interaction between experts is usually limited to training and briefing (e.g. Valverde, 2001; Cooke, 1991). It is generally believed that interaction between experts may harm mathematical aggregation because it induces dependence between their elicited judgements (e.g.

O'Hagan et al., 2006). Investigating this belief is one of the authors' current research themes. The main advantage of mathematical aggregation is that it makes the aggregation explicit and auditable. The choice of the aggregation rule is nevertheless difficult. Different rules possess different properties and it is not possible to have all desirable properties in one rule (Clemen & Winkler, 1999). One well established and widely used aggregation rule was formulated in Cooke (1991), a linear combination of judgements weighted by the experts' performance on test questions. This method asks experts to give estimates that can be validated with data in a process that is transparent and neutral.

The third class of protocols, called mixed methods, combines behavioural and mathematical aggregation techniques. The most common protocol in this class is the Delphi protocol (Rowe & Wright, 2001), which provides feedback from experts over successive questionnaire rounds. The experts do not interact with one another directly. Instead, a facilitator provides feedback between rounds of estimation. Experts remain anonymous to reduce the social and political pressures to accept judgements that can occur in face-to-face interactions of groups. The original Delphi method strives to reach consensus after a relatively small number of rounds. Although evidence is mixed, some critical reviews suggest that even though there is some convergence of individual judgements after a number of rounds, this convergence does not necessarily increase accuracy (e.g. Murphy et al., 1998).

The IDEA protocol synthesizes specific elements from all the approaches described above. In doing so, it aims to minimize the disadvantages of existing approaches and optimize their respective advantages. It is worth mentioning that the majority of elements that characterise IDEA are not new, it is the combination of these elements and the structured approach to this combination that make the IDEA protocol novel.

The IDEA protocol and the data used in this paper are introduced in Section 2. General concepts underlying the protocol are presented in Section 2.1, and the details of data collection are detailed in Section 2.2. Similar protocols are briefly contrasted with the IDEA protocol in the same section. The details of the expert data sets in terms of framing the questions, the number of questions, and the number of experts are presented in Section 2.3. Section 3 presents the analysis of the expert data collected using IDEA. The accuracy and confidence of individual and aggregated estimates are assessed (Sections 3.1 and 3.2). The relationship between accuracy

and confidence is investigated in Section 3.3, and Section 3.4 looks into explanations of patterns of experts' confidence. Concluding remarks are summarised in Section 4.

2. IDEA Protocol and Data Collection

2.1. The Protocol

IDEA can be thought of as a Delphi-like protocol in the sense that experts give their individual judgements in subsequent rounds of elicitations and facilitators provide feedback. In contrast to classical Delphi, IDEA does not seek consensus. Instead, a diverse group of experts first answer questions without engaging in discussion. Experts are then provided with the judgements of their peers and have the opportunity to discuss differences of opinion (unlike a Delphi protocol), allowing people to reconcile the meanings of words and context.

This design is motivated by literature that suggests that diverse interacting groups perform well at judgement tasks, and that judgements often are improved by reconciling arbitrary linguistic misunderstandings (e.g. Hong & Page, 2004; Carey & Burgman, 2008). Previous research also indicates that receiving feedback about other peoples' judgements improves the accuracy of individuals (Yaniv & Choshen-Hillel, 2012; Wintle et al., 2013), and group interaction is thought to be beneficial in quantitative judgement tasks (Schultze et al., 2012).

The discussion and debate is encouraged by facilitators and it may be remote or face-to-face. Individual judgements remain anonymous even during face-to-face discussion, thereby circumventing potential for dominance and halo effects that derail some group deliberations. It is possible that during a face-to-face discussion, people may declare the identity of their estimates to others, exposing them to dominant opinions and the pressure to conform, or that conversely, they may restrain from defending their opinion in order to avoid the risk of identification. Making the second estimate strictly anonymous helps to mitigate these phenomena. This promotes the benefits of behavioural aggregation, while guarding against some of the most debilitating elements of group elicitation. Details on how the discussion is coordinated are given in Section 2.3. At the end of the second round, the expert estimates are mathematically aggregated. The IDEA protocol also resembles the estimate-talk-estimate protocol (Gustafson et al., 1973), another mixed method, but with an extra, essential (Delphi-like) feature of providing feedback on first round responses.

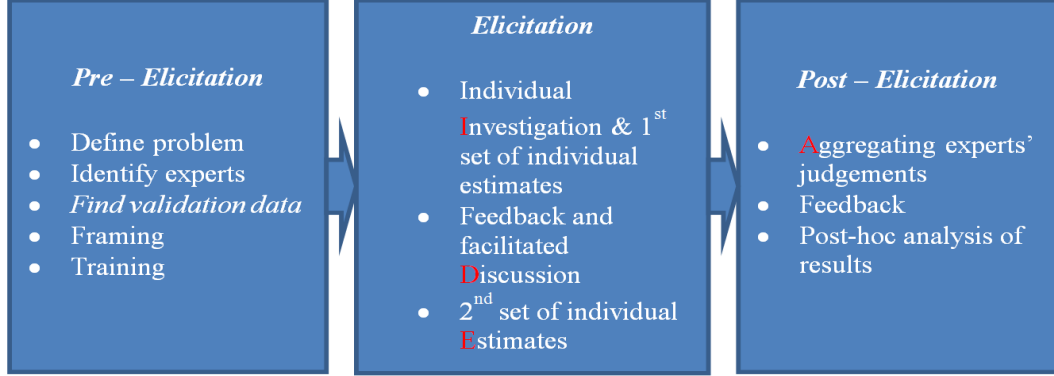


Figure 1: The proposed IDEA framework for structured expert judgement

Figure 1 gives a schematic representation of the IDEA process. Many of the pre-elicitation steps are characteristic of most structured protocols for expert elicitation. One essential feature (highlighted in Figure 1) of IDEA’s pre-elicitation stage is that it provides opportunities for empirical evaluation and validation. If expert opinion is used as scientific data, then it should be subjected to methodological rules for quality assurance, similar to those that are applied to other kinds of empirical data. This is important because too often experts abilities are assumed rather than established on the basis of past performance (e.g. Weiss & Shanteau, 2012; McKenzie et al., 2008), and because without proper quantitative tracking, it is hard to know how well individuals actually perform (e.g. Tetlock, 2005).

One way to do this involves asking experts for estimates on a set of (domain relevant) test questions whose values are (or soon will become) known. Test questions provide a way to measure experts’ performance. Measures of performance can then be used to enhance the mathematical aggregation of the experts’ final estimates. Different aggregation rules will produce different estimates that can be compared and evaluated based on the validation data provided by the test questions. To our knowledge, such empirical validation requirements are advanced in only one other structured expert judgement protocol designed by Cooke (1991).

The IDEA protocol is designed for eliciting quantitative estimates and the uncertainty around them. For variables measured on a continuous scale, the elicitation component of the procedure uses four questions, termed a four-step format (Figure 2, right). Three questions are used to elicit three values of the variable, corresponding to three different quantiles. The fourth question is used to identify the probabilities corresponding to the upper and lower quantiles specified by

the expert². In asking experts about events, three questions are used, a three-step format, to elicit the probability of occurrence and the uncertainty around it (Figure 2, left).

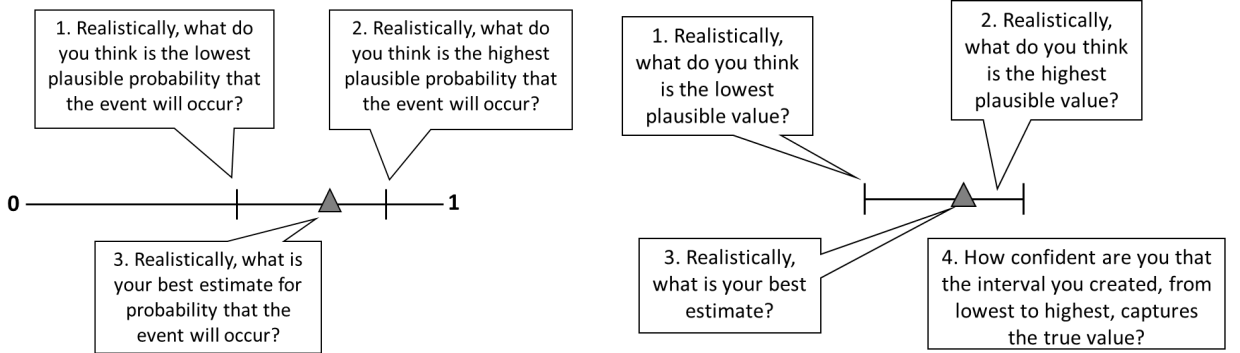


Figure 2: The three-step and the four-step format for occurrences of events and quantities, respectively

As mentioned above, the IDEA protocol draws heavily, but not exclusively, on findings from the psychology literature. Crucial elements of the IDEA protocol were first formulated, combined and used in Speirs-Bridge et al. (2010), Burgman et al. (2011) and McBride et al. (2012). However, the literature offers many competing recommendations for how to best elicit judgements and counter bias in predictive settings (e.g. Larrick, 2004; Moore et al., 2015b; Soll et al., 2015). More comparisons against alternative formulations of knowledge-resampling and group interactions would be beneficial. Thus, at this stage the IDEA protocol’s development is ongoing, and it will hopefully add to the work of others (e.g. Mellers et al., 2015a) on how structured group interactions and probability forecasting formats contribute to improved performance.

We restrict the following discussion to the elicitation of the probabilities of event occurrences, since these are the judgements collected in the data set to be introduced in Section 2.3. Each expert’s “best estimate” represents his or her subjective degree of belief in the probability that the event occurs. Prior to eliciting the “best estimate”, experts are asked to provide upper and lower bounds. We treat the bounds as an uncertainty interval.

The last step of IDEA involves the mathematical aggregation of the estimates. The choice of an aggregation rule is not straightforward, and decision scientists disagree about the most appropriate form of aggregation (e.g Bolger & Rowe, 2015). Aggregation rules can be evaluated using validation data and choices can be made application specific.

²It is worth pointing out that even though the fourth question asks for “confidence”, its answer helps defining a credible interval.

IDEA was developed primarily for eliciting judgements from experts. In this paper, we outline results of a study conducted with people who may be considered to be non-expert participants, although we use a broader interpretation of expertise to include self-taught individuals with specialist knowledge (Burgman et al., 2011). For this reason, the remainder of this paper uses the term *participant* instead of *expert*.

2.2. The Intelligence Game

The IDEA protocol was refined and tested as part of a forecasting “tournament” that started in 2011 as an initiative of the US Intelligence Advanced Research Projects Activity³. Five university-based research teams were involved in predicting hundreds of geopolitical, economic and military events, with the goal of finding the key characteristics of efficient protocols for eliciting and aggregating accurate probabilistic judgements. The project used real events that resolved in the near-future to test the accuracy of forecasts. Thousands of forecasters made over a million forecasts on hundreds of questions (Ungar et al., 2012; Mellers et al., 2015a).

IDEA was developed by researchers at the Centre of Excellence for Biosecurity Risk Analysis (CEBRA), at the University of Melbourne. CEBRA was part of the DAGGRE team (Decomposition-based Aggregation) led by George Mason University in the US. In the first years of the tournament, the George Mason team ran a prediction market alongside IDEA. Prediction markets provide an interesting alternative to the most common elicitation methods. They may be particularly effective at tracking group opinion (Sunstein, 2004); they have a long history of use, and recently their popularity has grown (Graefe & Armstrong, 2011). Prediction markets create contracts whose payoffs depend on the outcomes of uncertain events (Wolfers & Zitzewitz, 2004). Essentially, participants invest in their beliefs about outcomes, and trade relevant contracts. In principle, based on their performances, participants can be compensated. Therefore they have an interest in producing more accurate estimates, which should improve consensus probability. Graefe & Armstrong (2011) compared prediction markets to other elicitation techniques, some similar to the IDEA protocol, but were unable to distinguish any clear differences in performance. Our results add to the small but growing data on such comparisons.

The current paper presents and evaluates the IDEA protocol. The results are based on a subset of the questions developed by IARPA. In addition, IDEA forecasts were made for only

³<http://www.iarpa.gov/index.php/research-programs/ace>

two snapshots in time, while the broader competition was concerned with average performance over the entire prediction period for which questions were active (i.e. from the day IARPA posed the question to its resolution as either *true* or *false*). Hence, comparisons with the results obtained by other forecasting teams are outside the scope of this paper.

It is nevertheless worth pointing out key differences between IDEA and the protocols advanced by the other teams, including the Good Judgment Project (GJP) team who won the tournament (Mellers et al., 2015a). The IDEA protocol elicited uncertainty in the form of upper and lower bounds around probability estimates, whereas the main competition focused only on scoring point probabilities. Other teams encouraged similar *think again* and *estimate lower and upper bounds* style practices (e.g. Mellers et al., 2015b), motivated by literature on enhancing probability judgements, though to our knowledge, the IDEA protocol was unique in eliciting multiple, uncertainty judgements for each estimate. As opposed to other protocols, the IDEA protocol combines both feedback and interaction. Moreover the interaction is done with the help of a facilitator. Participants in the GJP teams for example, received training on how to interact effectively as a group, but engaged without external oversight. Finally, mathematical aggregation of participants’ opinions were investigated by other teams (e.g. Baron et al., 2014), but none considered the performance weighting (as in Cooke (1991)) used in the IDEA protocol.

2.3. The Data

The data presented and analysed in this paper were collected during two cycles of the forecasting tournament. The first cycle covered the period from June 2012 until April 2013 (here referred to as Year 1, even though it corresponds to the second year of the forecasting tournament⁴); the second cycle started in September 2013 and lasted until March 2014 (Year 2). Each month, a new set of questions was released. All questions considered here correspond to Bernoulli variables. Two example questions are: “Will the Turkish government release imprisoned Kurdish rebel leader Abdullah Ocalan before 1 April 2013?”, “Before 1 May 2014, will Iran install any new centrifuges?”. All questions usually resolved within 12 months, hence they are suited for empirical validation studies. Information about the total number of questions asked and answered is given in Table 1.

⁴The first year of the tournament is not covered in this analysis because the feedback and the facilitated discussion were handled differently.

A website (<http://intelgame.acera.unimelb.edu.au/>) was set up for the participants to answer the questions, discuss and upload/download necessary materials. Participants had IDs that could disclose or not their real names. Participants were allocated to subgroups and the discussion between the rounds of estimates took place within these subgroups. The way participants were recruited and placed into groups will be discussed later in this section.

Initial training of the participants took place before the game started. Some of the participants engaged in initial face-to-face training (in Year 1), where they learned about how the questions would be asked, why they were asked in this manner, and most importantly, the cognitive biases and group issues that can occur during an elicitation, and ways to mitigate them. Participants who did not receive face-to-face training, received online or telephone training. Training materials/documents that outlined and explained the issues above were also uploaded to the website for access and reference. Even though probabilistic training was not offered, many probabilistic concepts were introduced through practice questions that were part of the training.

A trained facilitator played a key role. Each month of the game followed the same structure, namely once the monthly set of questions was released the facilitator sent an email to all participants announcing the opening of the first round of estimates and the schedule of the month. Participants were encouraged to investigate the question context and were given approximately two weeks to provide their initial, individual estimates. After about ten days the facilitator sent another email announcing the imminent closing of the first round. During these first two weeks, all comments, links and resources were invisible to others. Participants were encouraged to save their sharing and discussion until the second round. After two days break, the facilitator sent an email announcing the completion of the between-rounds processing and the opening of the (one week long) second round. Participants were encouraged to participate in the discussion by logging in more than once, observing how the discussion developed for various questions, adding, viewing and rating resources. The debate and information sharing took place within groups. The facilitator supervised and encouraged (if necessary) the discussion for each group and question. Feedback from specific participants was asked for, usually when forecasts stood out, either as very confident, very high or low, or contrary to the group's majority opinion. It was nevertheless emphasised in personal communications that the participants should not feel singled out as *wrong*, but rather as the ones who potentially provided evidence that others

missed. Hence, rationales for answers and sharing resources were encouraged. Towards the end of the week, participants were encouraged to provide their second individual estimates.

Participants were found through social networking, initiated from the authors’ personal networks. Additional contacts were recruited from among their connections, etc. There were no constraints on participation, beyond a declared interest in world events and global politics.

In Year 1, participants included Australians, Americans and (a mix of Australian and US) college students. The Australian participants were divided among five groups (Group 1 to Group 5), the US participants were divided between two groups (Group 6 and Group 7). The students formed three groups (Group 8 to Group 10). The students were undergraduates at colleges in the US and Australia, enrolled in a range of programs including engineering, mathematics, politics and intelligence analysis.

For Year 2 the groups were redefined. A “Supergroup” (G1) was formed⁵ composed of the best performing participants from the previous year⁶. This construction is meant to help us accumulate evidence of prior performance being a useful guide to future performance on estimation tasks. Three more groups, G2, G3, and G4 of new and returning participants were formed. The new participants were assigned randomly to the remaining groups.

More details about the exact number of participants and the number of participants who answered at least one question in both rounds are given in Table 1. It is worth pointing out that the participants who followed the IDEA protocol for answering questions are different to those who followed the protocols of the other teams from the forecasting tournament.

Table 1 summarises data about the number of (active) participants and the number of questions. Participants from within a group answered subsets of different questions. Even though they were encouraged to answer all questions, the participants only answered those they felt comfortable responding to⁷. Therefore the average number of questions answered by a group was larger than the average number of questions answered by a participant.

⁵The GJP team tested an idea similar to the supergroup of top forecasters formed in Year 2, but on a considerably greater scale, i.e. with the benefits of many more participants.

⁶Performance was measured using the average Brier score introduced and discussed in the next section.

⁷This is less likely to happen in a more focused elicitation. Nevertheless the lesson learned is that the importance of answering all questions should be better stressed in future elicitations.

Table 1: Participants, groups and questions

	Year 1	Year 2
Total no. of participants	161 ^a	50 ^b
No. of groups	10	4
Total no. of questions	55	34
No. of participants who answered at least 1 question	121	37
Average no. of participants per group answering at least 1 question	12	9
Average no. of questions answered by a group	29	33
Average no. of questions answered by a participant	11	17

^a91 Australian participants, 35 US participants and 35 college students

^b45 participants from Year 1 and 5 new participants.

Using the data collected from the forecasting tournament, we investigate two main features of participants’ forecasts: how accurate (Section 3.1) and how confident (Section 3.2) they are, individually and as a group. Both qualities are useful in measuring individual/group performance. We also investigate the relationship between these two qualities (Section 3.3) and factors that might influence them. Specifically, we investigate the influence of discussion between rounds on individual/group accuracy and confidence, how confidence changes as a function of the best guess, and how experience (with the protocol) influences the performance of participants.

3. Analysis

3.1. Measuring Accuracy

In assessing the performance of probability forecasters, we are interested in assessing their accuracy, which we define here as a measure of the degree of correspondence between the participants’ predictions (which are the probabilities of event occurrences) and the observed outcomes.

We wish to point out that our definition of the accuracy of forecasts differs from the commonly used measure of hit rate; this is because, in our case, the forecasts are of the probability of event occurrence rather than event occurrence or non-occurrence per se⁸.

Using the IDEA protocol, participants are asked for three estimates per question including the participant’s best guess for the probability of occurrence, and it is this best estimate that we

⁸Our definition of accuracy also differs from others in the literature on evaluating subjective probabilities (e.g Yaniv & Foster, 1997; Lichtenstein et al., 1982; Moore et al., 2015a)

will focus on in evaluating judgement accuracy. One very commonly used measure for evaluating the accuracy of forecasts (e.g. Murphy & Winkler, 1992; Yates et al., 1996; O’Hagan et al., 2006) is the Brier score (Brier, 1950), which is calculated as twice the squared difference between an estimated probability (best guess) and the actual outcome; hence it takes values between 0 and 2. Consider question/event i with two possible outcomes j . The Brier score of participant k assessing event i is calculated as follows:

$$BrierScore_i^k = \sum_{j=1}^2 (p_{ij}^k - x_{ij})^2,$$

where p_{ij}^k is participant k ’s probability for event i , outcome j , and x_{ij} is 1 if outcome j occurs and 0 otherwise. The above formula measures the accuracy of one estimate (of one participant) for one question. The lower the score, the better the accuracy (a perfectly accurate expert would receive a score of zero). In practice, this means that good scores can be achieved if an expert assigns large probabilities to events that occur, or small probabilities to events that do not occur. Usually, participants’ accuracy is measured over many questions (N) and averaged to represent long term accuracy:

$$BrierScore^k = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^2 (p_{ij}^k - x_{ij})^2$$

When we consider groups of participants, we are interested in the group accuracy as measured by the average Brier score of the group. For each group, the best estimates of the participants are averaged per question, and a Brier score per question, per group is calculated. The average Brier score (for all questions answered by at least one participant from that group) is reported as the accuracy measure of the group.

In Year 1, we compared an equally weighted combination (average of all best estimates) of all participants after the first round of estimates (prior to discussion) to the equally weighted estimates of the three group types (the Australian participants, the US participants and the college students) after discussion.

Figure 3 compares the average Brier scores of these combinations with those of the prediction market⁹ (with 95% confidence intervals). Similar performance is achieved by two IDEA groups

⁹Slight differences between the prediction market scores, reported here and reported previously by the

and the prediction market. DAGGRE used the Logarithmic Market Scoring Rule (LMSR) (Hanson, 2003, 2007) with an “edit based” interface: there was no buying or selling of shares, only seeing and adjusting a probability. The cost of the adjustment was computed using the LMSR, which can be roughly imagined as a bowl-shaped function where the cost of further changes climbs faster as the change becomes bigger. The LMSR is thought to be particularly well-suited to DAGGRE’s combinatorial market (Sun et al., 2012).

All combinations except the college students outperformed the equally weighted combination of first round estimates. This indicates that discussion improves performance in most cases. This hypothesis will be further discussed in the following sections. The reason for the poor performance of the students group is not obvious, but may have been because they were less highly motivated or more homogeneous groups, lacking the diversity of age, background, and life experience of the other groups.

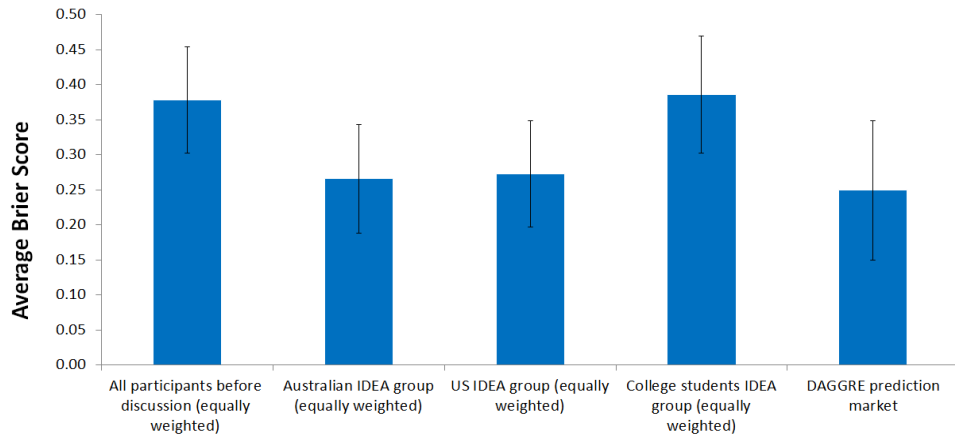


Figure 3: Year 1: Average Brier Scores

Figure 4 shows a similar comparison as Figure 3, but for Year 2. The DAGGRE prediction market closed in June 2013, due to a change in focus; hence for Year 2, no prediction market comparison is available.

The equally weighted combination of first round estimates was outperformed by all other combinations except G3. The Supergroup G1 outperformed the other groups of participants, indicating that prior performance is a useful guide to future performance on similar prediction and estimation tasks. A slightly better performance (although not statistically significant as

DAGGRE team, are due to slightly different data sets (due to different subsets of questions, different subsets of participants) used in the analysis.

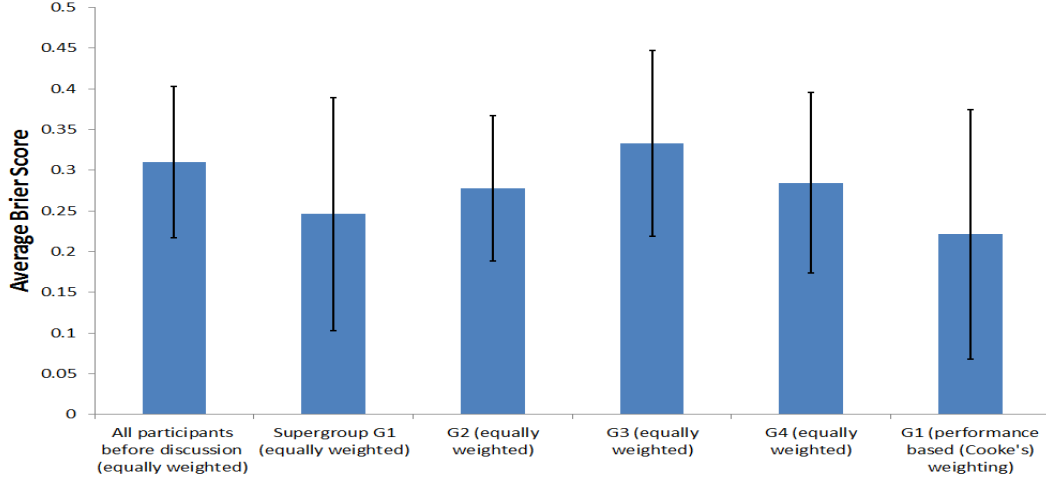


Figure 4: Year 2: Average Brier Scores

indicated by the confidence intervals) is achieved by using a performance based weighting scheme. The non-zero weights are calculated using the Cooke’s (classical) method (Cooke, 1991). In this method, participants are scored based on measures of calibration and informativeness. Participants’ best estimates are assigned to probability bins of the following form (0.1,0.9), (0.2,0.8), (0.3,0.7), etc. For example, a participant assigns a variable to the (0.2,0.8) bin if his or her best estimate (about the probability of occurrence) is anywhere between 0.1 and 0.2. From the set of events assigned to a probability bin, only some occur, allowing us to calculate an empirical distribution of each bin. Calibration scores are based on the difference between the empirical and the theoretical distribution of each bin. Informativeness scores are based on Shannon’s entropy, which is often taken as a measure of lack of information in a distribution. Two entropy measures are discussed in Cooke (1991), the average response entropy and the average sample entropy. The response entropy measures the entropy in what the expert says and it does not depend on the actual occurrences of events. The sample entropy measures the entropy in the expert’s performance, but it does not correspond to the distribution that the expert (or anyone else) believes. For more details and precise definitions we refer to Cooke (1991).

Having compared the scores from Figures 3 and 4, it is also worth mentioning that the differences between most of these scores (within each year) are less than 0.1. The visual inspection of the confidence intervals associated with the scores, indicates that the sample sizes were insufficient to detect statistically significant differences of this magnitude. The size of the confidence

intervals is nevertheless determined by more than one factor, namely the sample size, the large individual differences, score fluctuations which depend on the number of variables assessed. All these issues and the way they influence the reliability of the above comparisons are the subject of the authors' current research.

One of IDEA's distinctive features is the way participants' interaction/discussion is handled. Figure 5 shows the average Brier scores of all participants in Year 1 (left) and Year 2 (right), *before discussion*, plotted against the scores *after discussion*. The main diagonal is plotted for better visualisation of the points' distribution. In both cases, most of the points fall below the main diagonal, indicating that average Brier scores decrease (accuracy improves) after discussion.

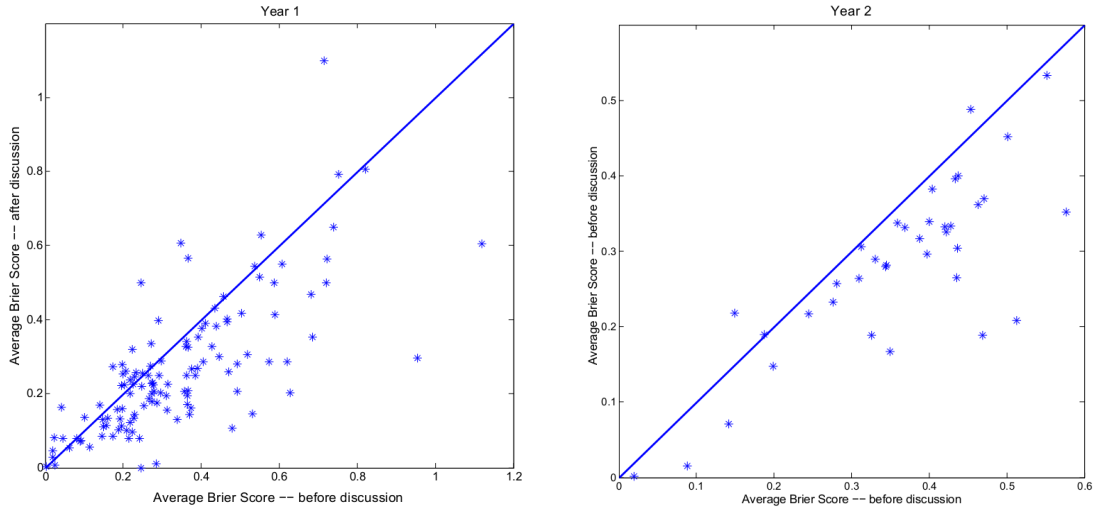


Figure 5: Participants' average Brier scores before discussion vs. average Brier scores after discussion

We also investigate the change in the average Brier scores per group following discussion between rounds. The average Brier scores per group, for Year 1 are presented in Wintle et al. (2012). The improvement in group performance in Year 2 is shown in Figure 6.

Figure 6 shows a consistent reduction in the group scores after discussion (denoting improvement in accuracy). However, the size of the 95% confidence intervals around the estimated average Brier scores indicate the change is not statistically significant for any given group. Nevertheless, on an individual scale, 91% of the participants changed their estimates in the direction of the true outcome.

There are 30 people who participated in both years of the tournament and answered at least one question in both rounds. They are called *returning participants*. We expected that

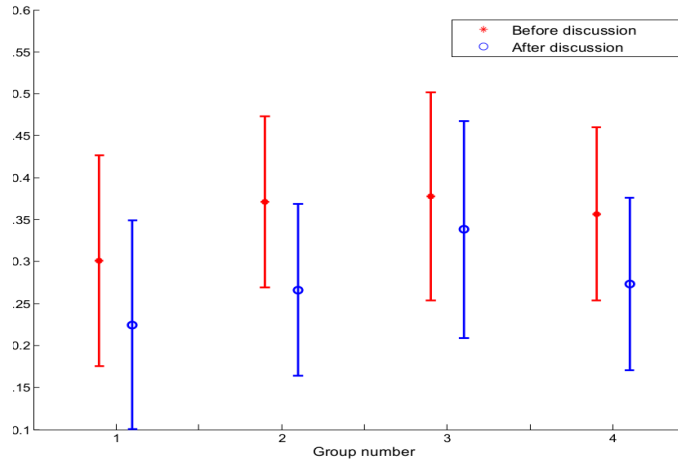


Figure 6: Groups' average Brier scores (and 95% CI) before and after discussion in Year 2

experience and practice would improve performance. Surprisingly, the accuracy of the majority of returning participants worsened. Almost 77% of the returning participants had higher Brier scores (decreased performance) in round one of Year 2 when compared to their scores in round one of Year 1. Almost 84 % were in this situation for round two, with 20 % of the scores' differences being larger than 0.2. One possible explanation is that the question difficulty increased between years. Nevertheless, question difficulty is rather difficult to define, let alone to measure.

The mean of the average Brier scores for the 30 returning participants (in round two) is 0.18 in Year 1 and 0.29 in Year 2. The means of average Brier scores of random groups of 30 participants in Year 1 are all very close to 0.28, which indicates that the 30 returning participants performed better than a random group of participants in Year 1. The similarity between the average performance of a random group in Year 1 (0.28), and the performance of the returning participants in Year 2 (0.29) could be taken as a counter-argument for the hypothesis that differences between question difficulty levels for Year 1 and Year 2 were responsible for the increase observed in average Year 2 Brier Scores . It could also reflect regression to the mean.

Another hypothesis is that the returning participants were more involved and focused in Year 1 than they were in Year 2, due to the decreased novelty of the exercise. This is supported by a decrease in activity between the rounds of elicitation, as measured by a decreasing average number of viewed and rated resources between years.

3.2. Measuring Confidence

The calculations above explore the accuracy of participants' best estimates, but a feature of using the IDEA protocol, is that participants also provide a separate uncertainty interval around each of their best estimates.

The length of these intervals can be interpreted as a measure of the participants' confidence in their estimates, or lack thereof. Under this interpretation, a larger (wider) interval denotes decreased confidence, while a smaller (narrower) interval denotes an increased level of confidence.

To avoid ambiguity, it is worth clarifying the use of the word *confidence* throughout the paper. Firstly, referring to the length of the uncertainty intervals as confidence when the three-step format is used is similar to referring to the length of credible intervals as a participant's confidence when the four-step format is used. In neither of the two cases is the meaning of confidence formalised. Secondly, this use of *confidence* is obviously unrelated to the frequentist confidence intervals calculated in the previous section. And finally, and perhaps most importantly, this use and interpretation of the word *confidence* differs from more standard uses in the body of work on subjective probabilities (e.g. Lichtenstein et al., 1982; Keren, 1991; Harvey, 1997; Griffin & Brenner, 2004). In particular, a large body of literature that addresses questions where participants choose between two alternatives and give a probability that their answer is correct, has used participant *confidence* to refer to the probabilities assigned to the likelihood of an event occurring (equating to what in this paper we have termed the best estimate). Under this research paradigm, a participant is more confident in their responses when they assign more extreme probabilities with the likelihoods of events occurring or not. Critiques of this experimental paradigm can be found in Griffin & Brenner (2004); Moore & Healy (2008). This paradigm has primarily been focused on confidence in knowledge, with most questions focusing on general knowledge statements, rather than on estimating probabilities of events' occurrences. Another use of the word *confidence* is in conjunction with the overconfidence analysis based on placing events in probability bins and calculating calibration rather than accuracy.

Having clarified the context, and given that we would like to define a measure of confidence which is, in some sense, equivalent with that used in the forecasting of continuous variables¹⁰,

¹⁰Assessing the width of uncertainty intervals when forecasting continuous variables has been established as a useful way to measure the uncertainty or confidence a participant associates with their knowledge.

we chose the length of the uncertainty interval as a measure of confidence.

Other authors refer to such length of uncertainty intervals as precision (e.g. Lyon et al., 2015), but we chose to avoid employing the word *precision* which might be too easily misinterpreted as associated with imprecise probabilities.

The analysis in Section 3.1 suggest that the discussion between the rounds improves the performance of participants. In this section, we investigate the relationship between the uncertainty intervals before and after discussion in the 10 groups from Year 1, and in the four groups from Year 2. Figure 7 shows the percentage of participants by group whose confidence increased (i.e. the length of the uncertainty interval decreased) after the group discussion. Most of participants' mean confidence levels increased after the discussion between rounds. Fortunately, the increased confidence is coupled with an increased accuracy between rounds.

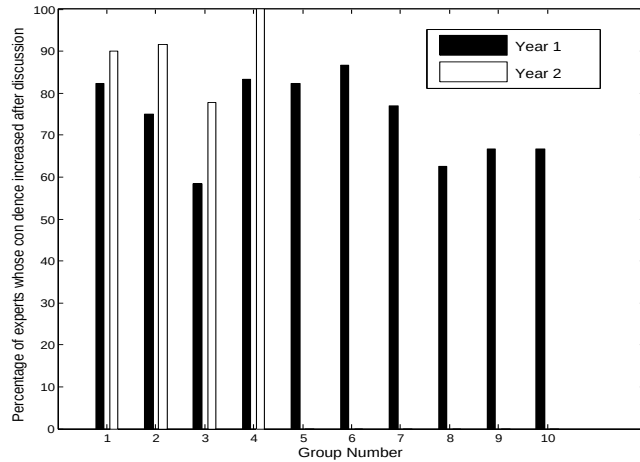


Figure 7: Percentage of participants whose confidence increased after discussion by group

If we ignore group membership and combine the data for all participants per year, we can investigate the magnitude of individual change (see Figure 8).

Looking at the returning participants' behaviour in terms of confidence change over the years, the situation is different. 60% (63%) of the participants exhibited decreased confidence when comparing interval lengths from Year 1, round 1 (round 2) to Year 2, round 1 (round 2). It may be that participants' previous experience reduced their confidence in their estimates.

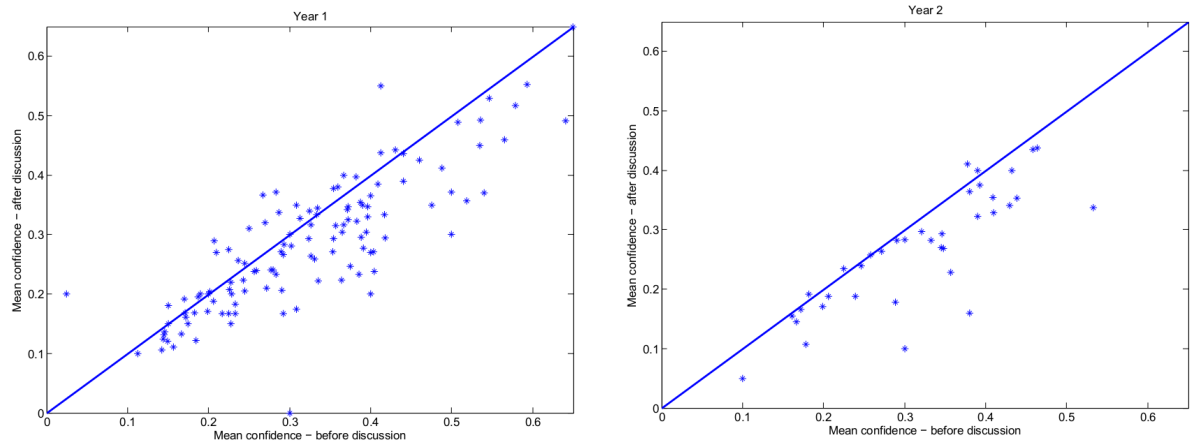


Figure 8: Mean confidence change after discussion

3.3. Confidence versus Accuracy

The question of whether people’s confidence in their judgements increases in line with improvements in performance (measured as accuracy or calibration¹¹) has long been of interest for researchers looking at subjective probabilities (e.g. Lichtenstein et al., 1982; Moore et al., 2015a).

The often used *calibration curve* (e.g. Lichtenstein et al., 1982) for example, provides a visualization of the degree to which a subjective probability judgement is matched by the observed relative frequency. Overconfidence (underconfidence) occurs when the observed relative frequency is lower (higher) than the subjective probability. While ideally, any increase in a participant’s confidence would be matched with an increase in performance, research across many different measures of confidence and performance has found that subjective confidence is often poorly correlated with performance (e.g. Lichtenstein et al., 1982; Griffin & Tversky, 1992; Moore et al., 2015b; Aspinall & Cooke, 2013).

In this next section we are interested in examining the relationship between performance (measured as accuracy) and confidence (as measured by uncertainty interval length) more generally. While our measures of confidence and accuracy do not correspond to those typically used in examining performance/confidence relationships, given that the participants knew they were being assessed using an average Brier score, it is interesting to see whether their actual performance as assessed by their average Brier score is related to the average degree of confidence (as

¹¹an alternative widely used performance measure when eliciting subjective probabilities

measured by uncertainty interval length) they provided.

Figure 9 shows the linear correlation between mean confidence and the average Brier score, after discussion, per group, per year. In Year 1, the correlations range from -0.37 (with $(-0.9, 0.6)$ as 90% confidence interval) to 0.3 (with $(-0.2, 0.7)$ as 90% confidence interval) and all confidence intervals include zero, rendering no statistically significant correlation.

Two of the values for Year 2 tell a similar story, those for Groups 3 and 4 in the sense that the estimated correlation coefficients indicate no obvious relationship between the average Brier scores and the mean length of the confidence intervals. For the other two groups larger absolute correlations were found. For Group 1 a correlation of -0.65 indicates a negative correlation between accuracy and confidence; whereas for Group 2 a correlation of 0.59 indicates a positive association. Nevertheless, both correlations are found to be on the verge of statistical significance (the upper bound for Group 1 is -0.038 and the lower bound for Group 2 is 0.026). Averaging across groups, we failed to detect a relationship between confidence and accuracy. However, we acknowledge that a larger sample may have detected an effect.

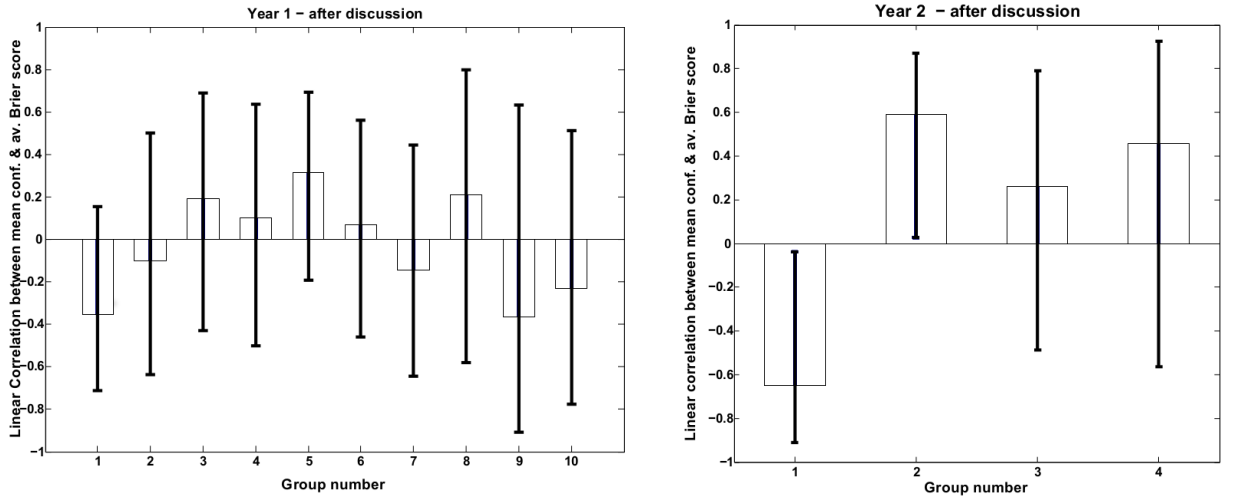


Figure 9: Linear correlation between confidence and accuracy.

3.4. Confidence versus Best Estimate

A participant's confidence might change depending on the associated probability of the uncertainty interval. The choice (alone) of a more extreme best guess (i.e. assigning a probability close to 0 or 1) can also indicate confidence. This should produce a $\cap$ -shape when each participant's best guess is plotted against the length of the interval associated with it. Before defining

a more general (and different) measure of confidence to reflect and quantify such behaviour, we will investigate individual patterns.

We considered the set of participants who answered at least 25 questions in each year and at least 40 questions in both years (for returning participants). This set had 21 participants, 10 of whom exhibited random patterns. From the remaining 11 participants, the estimates of five followed the $\cap$ -shape, and the estimates of the others followed a $\cap$ -shape that seemed incomplete because their best guesses were never larger than 0.8. Figure 10 shows three examples.

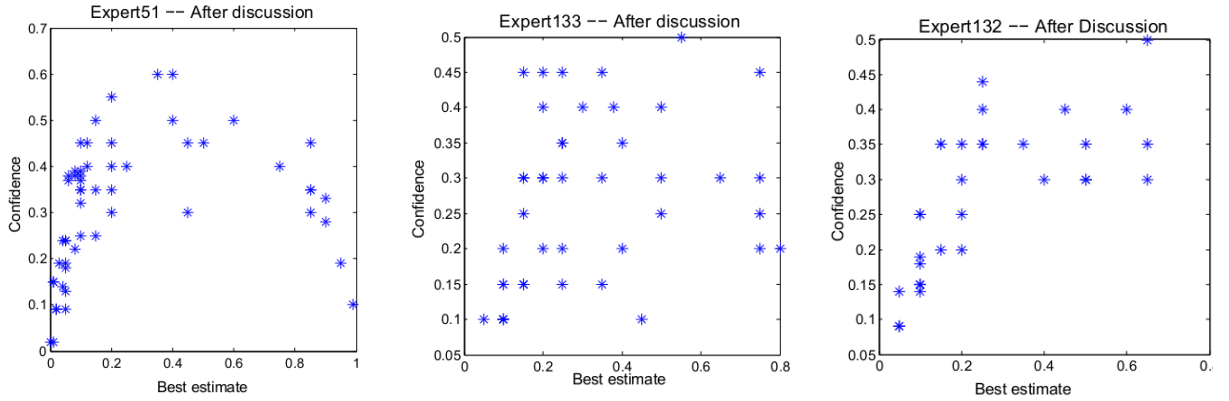


Figure 10: The relationship between the best estimate and confidence

Investigating the relationship between the best estimate and confidence for the returning participants who answered more than 30 questions, we found that the answers of nine (out of 16) appeared to be random, and the answers of seven followed the $\cap$ -shape.

Even though roughly half of the participants exhibited random patterns between the best estimate and confidence, the other half followed the expected $\cap$ -shaped pattern. Based on these findings we can define an informativeness score to measure how extreme (i.e. how different from the uniform case) a participant's assessments are.

4. Discussion

Our results show that accuracy improves when participants share information between rounds. They also indicate that confidence increases between rounds (in terms of narrower intervals) yet our results do not detect a clear relationship between confidence (as measured by interval length) and accuracy overall. Other studies show mixed results on the accuracy-confidence relationship as well. Together, the evidence suggests large individual differences,

making an effect difficult to detect without a large sample. To increase the sample size, we could repeat the same analysis, pooling all experts in a single group, rather than investigating groups separately. Nevertheless, such pooling is only legitimate if the dependence between the experts is weak, and fairly constant with respect to group membership. The latter condition may easily be violated due to different group sizes, group compositions, commitment, and the way discussion is facilitated. The dependence analysis is one of the authors current research themes.

The magnitude of the best guess seems to influence the length of the confidence interval associated with it, such that more extreme estimates are associated with higher confidence. This feature is currently investigated by measuring the departure from the uniform distribution (0.5 probability of occurrence) in participants' assessments. This measure shares characteristics with the informativeness measure defined in Cooke (1991). We expect a negative correlation between informativeness (as measured by the distance to the uniform distribution) and confidence.

In this study, returning, more experienced participants were less confident and relatively less accurate than they were in the first year of their involvement in the study. We expected returning participants to be at least as accurate, if not more so, so this result is quite surprising and requires more detailed research. We speculate that the most likely cause is decreased engagement. We note that this finding is in contrast to that of Mellers et al. (2015a), who observed consistent improvements in participant calibration from year to year during the tournament, though they also found that more years of forecasting experience did not correspond to better performance than that of less experienced performers (Moore et al., 2015b).

Given the relatively small sample sizes in some of our analyses, these results help us formulate conjectures rather than strong conclusions. They should be seen as signals to further explore the relationships described here.

We acknowledge the limitations in interpreting our results, some of which may be due to the variability that arises in human subjects research. As noted, in the initial stages of protocol development, we explored minor variations in recruitment, group composition and training styles. These along with the relatively small sample sizes could have influenced our results in ways we cannot discern at this time. A data set for Year 3 (2014 - 2015) of the forecasting tournament is currently being analysed, so new comparisons will shed further light on our preliminary findings.

Moreover, many of our observations are based on the Brier score measurements. This score is an aggregate measure for individual variables. The aggregated score fluctuations depend on the number of variables assessed. In other words, the score is a random variable and interpreting the values of the score requires knowledge of the score's distribution, which in the Brier score's case is unavailable. Other scores (amongst which the calibration term of the Brier score) are currently being investigated and compared in order to design a robust validation process of the participants' assessments. Validation is perhaps the most important, and at the same time, the most challenging aspect of a reliable expert judgement protocol.

References

- Aspinall, W. (2010). A route to more tractable expert advice. *Nature*, *463*, 294–295.
- Aspinall, W. & Cooke, R. (2013). Quantifying scientific uncertainty from expert judgement elicitation. In: Rougier, J., Sparks, S., & Hill, L. (eds.), *Risk and uncertainty assessment for natural hazards*, chap. 10, pp. 64–99, Cambridge: Cambridge University Press.
- Baron, J., Mellers, B., Tetlock, P., Stone, E., & Ungar, L. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*, *11*, 451–468.
- Bergstrom, D., Lucieer, A., Kiefer, K., Wasley, J., Belbin, L., Pedersen, T., & Chown, S. (2009). Indirect effects of invasive species removal devastate World Heritage Island. *Journal of Applied Ecology*, *46*, 73–81.
- Bolger, F. & Rowe, G. (2015). The aggregation of expert judgment: Do good things come to those who weight? *Risk Analysis*, *35*, 5–11.
- Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*, 1–3.
- Burgman, M., Carr, A., Godden, L., Gregory, R., McBride, M., Flander, L., & Maguire, L. (2011). Redefining expertise and improving ecological judgment. *Conservation Letters*, *4*, 81–87.
- Carey, J. & Burgman, M. (2008). Linguistic uncertainty in qualitative risk analysis and how to minimize it. *Annals of the New York Academy of Sciences*, *1128*, 13–17.
- Clemen, R. & Winkler, R. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis*, *19*, 187–203.
- Cooke, R. (1991). *Experts in uncertainty: Opinion and subjective probability in science*. Environmental Ethics and Science Policy Series, Oxford University Press.
- Cooke, R. & Goossens, L. (2008). TU Delft expert judgment data base. *Reliability Engineering and System Safety*, *93*(5), 657–674.
- French, S. (2012). Expert judgment, meta-analysis and participatory risk analysis. *Decision Analysis*, *9*, 119–127.
- Graefe, A. & Armstrong, J. (2011). Comparing face-to-face meetings, nominal groups, Delphi and prediction markets on an estimation task. *International Journal of Forecasting*, *27*, 183–195.
- Griffin, D. & Brenner, L. (2004). Perspectives on probability judgment calibration. In: Koehler, D. & Harvey, N. (eds.), *In Blackwell handbook of judgment and decision making*, pp. 177–199, Malden: Blackwell Publishing Ltd.
- Griffin, D. & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, *24*, 411–435.
- Gustafson, D., Shukla, R., Delbecq, A., & Walster, G. (1973). A comparative study of differences in subjective likelihood estimates made by individuals, interacting groups, Delphi groups, and nominal groups. *Organizational Behavior and Human performance*, *9*, 280–291.
- Hanson, R. (2003). Combinatorial information market design. *Information Systems Frontiers*, *5*(1), 107–119.
- Hanson, R. (2007). Logarithmic market scoring rules for modular combinatorial information aggregation. *The Journal of Prediction Markets*, *1*(1), 3–15.
- Harvey, N. (1997). Confidence in judgment. *Trends in Cognitive Sciences*, *1*, 78–82.
- Hinsz, V., Tindale, R., & Vollrath, D. (1997). The emerging conceptualization of groups as information processors. *Psychological Bulletin*, *121*(1), 43–64.
- Hong, L. & Page, S. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the USA*, *101*, 16385–16389.

- Kahneman, D. & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, 39(4), 341–350.
- Keeney, R. & von Winterfeldt, D. (1991). Eliciting probabilities from experts in complex technical problems. *IEEE Transactions on Engineering Management*, 38, 191–201.
- Keren, G. (1991). Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica*, 77, 217–273.
- Larrick, P. (2004). Debiasing. In: Koehler, D. & Harvey, N. (eds.), *Blackwell handbook of judgment and decision making*, pp. 316–359, Malden: Blackwell Publishing Ltd.
- Lichtenstein, S., Fischhoff, B., & Phillips, L. (1982). Calibration of probabilities: The state of the art to 1980. In: Kahneman, D., Slovic, P., & Tversky, A. (eds.), *Judgment under uncertainty: Heuristics and biases*, pp. 306–334, Cambridge: Cambridge University Press.
- Lyon, A., Wintle, B., & Burgman, M. (2015). Collective wisdom: Methods of confidence interval aggregation. *Journal of Business Research*, 68, 1759–1767.
- McBride, M., Garnett, S., Szabo, J., Burbidge, A., Butchart, S., Christidis, L., Dutson, G., Ford, H., Loyn, R., Watson, D. M., & Burgman, M. (2012). Structured elicitation of expert judgments for threatened species assessment: A case study on a continental scale using email. *Methods in Ecology and Evolution*, 3, 906–920.
- McKenzie, C., Liersch, M., & Yaniv, I. (2008). Overconfidence in interval estimates: What does expertise buy you? *Organizational Behavior and Human Decision Processes*, 107, 179–191.
- Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S., Ungar, L., Bishop, M., Horowitz, M., Merkle, E., & Tetlock, P. (2015a). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied*, 21, 1–14.
- Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., & Chen, E. (2015b). Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science*, 10(3), 267–281.
- Meyer, M. & Booker, J. (1991). *Eliciting and analyzing expert judgment: a practical guide*. New York: Academic Press.
- Moore, D. & Healy, P. (2008). The trouble with overconfidence. *Psychological review*, 115(2), 502–517.
- Moore, D., Tenney, E., & Haran, U. (2015a). Overprecision in judgment. In: Keren, G. & Wu, W. (eds.), *The Wiley Blackwell handbook of judgment and decision making*, pp. 182–212, Chichester: John Wiley & Sons, Ltd.
- Moore, D. A., Swift, S., Minster, A., Mellers, B., Ungar, L., Tetlock, P., & Tenney, E. (2015b). Confidence calibration in a multi-year geopolitical forecasting competition. *Working Paper*.
- Morgan, M. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. In: *Proceedings of the National Academy of Sciences*, vol. 111, pp. 7176–7184, USA.
- Murphy, A. & Winkler, R. (1992). Diagnostic verification of probability forecasts. *International Journal of Forecasting*, 7, 435–455.
- Murphy, M., Black, N., Lamping, D., Mckee, C., & Sanderson, C. (1998). Consensus development methods and their use in clinical guideline development. *Health Technology Assessment*, 2(3), 1–88.
- O’Hagan, A. (2005). Elicitation. *Significance*, 2, 84–86.
- O’Hagan, A., Buck, C., Daneshkhah, A., Eiser, J., Garthwaite, P., Jenkinson, D., Oakley, J., & Rakow, T. (2006). *Uncertain judgements: Eliciting experts’ probabilities*. London: Wiley.
- Phillips, L. & Phillips, M. (1907). Facilitated work groups: Theory and practice. *Journal of the Operational Research Society*, 44(6), 533–549.
- Rabinowitz, A. (1998). Helping a species to go extinct: the Sumatran rhino in Borneo. *Conservation Biology*, 9, 482–488.
- Rowe, G. & Wright, G. (2001). 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.
- Schultze, T., Mojzisch, A., & Schulz-Hardt, S. (2012). Why groups perform better than individuals at quantitative judgment tasks: Group-to-individual transfer as an alternative to differential weighting. *Organizational Behavior & Human Decision Processes*, 118, 24–36.
- Shine, R. (2010). The ecological impact of invasive cane toads (*Bufo Marinus*) in Australia. *The Quarterly Review of Biology*, 85(3), 253–259.
- Slovic, P. (1999). Trust, emotion, sex, politics, and science: surveying the risk-assessment battle field. *Risk Analysis*, 19, 689–701.
- Soll, J., Milkman, K., & Payne, J. (2015). A user’s guide to debiasing. In: Keren, G. & Wu, W. (eds.), *Wiley-Blackwell handbook of judgment and decision making*, pp. 924–951, Chichester: John Wiley & Sons, Ltd.
- Speirs-Bridge, A., Fidler, F., McBride, M., Flander, L., Cumming, G., & Burgman, M. (2010). Reducing overconfidence in the interval judgments of experts. *Risk Analysis*, 30, 512–523.
- Sun, W., Hanson, R., Laskey, K., & Twardy, C. (2012). Probability and asset updating using Bayesian networks for combinatorial prediction markets. *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI-2012)*.

- Sunstein, C. (2004). Group judgments: Deliberation, statistical means, and information markets. *John M. Olin Law & Economics Working Paper No. 219, University of Chicago Law School*.
- Tetlock, P. (2005). *Expert political judgment: How good is it? How can we know?* Princeton: Princeton University Press.
- Ungar, L., Mellers, B., Satopaa, V., Baron, J., Tetlock, P., Ramos, J., & Swift, S. (2012). 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).
- Valverde, L. (2001). Expert judgment resolution in technically-intensive policy disputes. In: *Assessment and management of environmental risks*, pp. 221–238, Norwell: Kluwer Academic Publishers.
- Weiss, D. & Shanteau, J. (2012). Decloaking the privileged expert. *Journal of Management & Organization*, 18, 300–310.
- Wintle, B., Fidler, F., Vesk, P., & Moore, J. (2013). Improving visual estimation through active feedback. *Methods in Ecology and Evolution*, 4, 53–62.
- Wintle, B., Mascaro, M., Fidler, F., McBride, M., Burgman, M., Flander, L., Saw, G., Twardy, C., Lyon, A., & Manning, B. (2012). The Intelligence Game: Assessing Delphi groups and structured question formats. In: *Proceedings of the 5th Australian Security and Intelligence Conference*.
- Wolfers, J. & Zitzewitz, E. (2004). Prediction markets. *Journal of Economic Perspectives*, 18(2), 107–126.
- Yaniv, I. & Choshen-Hillel, S. (2012). Exploiting the wisdom of others to make better decisions: Suspending judgment reduces egocentrism and increases accuracy. *Journal of Behavioral Decision Making*, 25, 427–434.
- Yaniv, I. & Foster, D. (1997). Precision and accuracy of judgmental estimation. *Journal of Behavioral Decision Making*, 10(1), 21–32.
- Yates, J., Pricea, P., Leeb, J., & Ramireza, J. (1996). Good probabilistic forecasters: The consumer’s perspective. *International Journal of Forecasting*, 12(1), 41–56.