DOI: 10.1111/risa.12992

The Value of Performance Weights and Discussion in Aggregated Expert Judgments

Anca M. Hanea,[1,*] Marissa F. McBride,[2,3] Mark A. Burgman,[2,4] and Bonnie C. Wintle[2,5]

In risky situations characterized by imminent decisions, scarce resources, and insufficient data, policymakers rely on experts to estimate model parameters and their associated uncertainties. Different elicitation and aggregation methods can vary substantially in their efficacy and robustness. While it is generally agreed that biases in expert judgments can be mitigated using structured elicitations involving groups rather than individuals, there is still some disagreement about how to best elicit and aggregate judgments. This mostly concerns the merits of using performance-based weighting schemes to combine judgments of different individuals (rather than assigning equal weights to individual experts), and the way that interaction between experts should be handled. This article aims to contribute to, and complement, the ongoing discussion on these topics.

KEY WORDS: Aggregation; confidence; elicitation protocol; performance-based weighting schemes; structured expert judgment

1. INTRODUCTION

Risk assessment and management are often characterized by imminent decisions and scarce resources. Usually, at least some requisite data for pressing decisions are unreliable, rudimentary, or entirely absent. In these situations, policymakers rely on expert judgments to fill the knowledge gaps. Experts estimate quantities (parameters of models), decide the forms of cause-effect relationships, and predict the outcomes of management interventions, all of which are uncertain. We restrict our attention to the elicitation of parameters and their associated uncertainties from experts. How these are best elicited and combined across experts is critical to a decision process, as differences in the efficacy and robustness of elicitation and aggregation methods can be substantial (e.g., Clemen & Winkler, 1999; Cooke, 1991; Morgan, 2015; O'Hagan et al., 2006).

It is generally agreed that in scientific disciplines, experts are very sensitive to contextual and psychological frailties, including anchoring, dominance effects, and overconfidence, which can severely distort technical judgments (e.g., Burgman, 2005; Kahneman & Tversky, 1984; Montibeller & von Winterfeldt, 2015; Slovic, 1999), and that many of these frailties can be mitigated using structured elicitations that involve groups rather than individuals (e.g., Burgman, McBride, Ashton, Speirs-Bridge, & Flander, 2011; Cooke, 1991; Morgan, 2015; Sunstein, 2004). Since the 1990s, and building on developments in decision theory, mathematics, and behavioral science, some researchers have argued that substantial improvement in the quality and reliability of

[1] Centre of Excellence for Biosecurity Risk Analysis, University of Melbourne, Parkville, VIC, Australia.

[2] School of BioSciences, University of Melbourne, Parkville, VIC, Australia.

[3] Harvard Forest, Harvard University, Petersham, MA, USA.

[4] Centre for Environmental Policy, Imperial College London, London, UK.

[5] Centre for the Study of Existential Risk, University of Cambridge, Cambridge, UK.

[*] Address correspondence to Anca Maria Hanea, Centre of Excellence for Biosecurity Risk Analysis, University of Melbourne, Parkville, VIC, Australia; ahanea@unimelb.edu.au.

0272-4332/18/0100-0001$22.00/1

aggregated expert judgments can be achieved by using performance-based weighting schemes of aggregation (Aspinall, 2010; Burgman et al., 2011; Cooke & Goossens, 2008). However, heated debate continues in the literature regarding the practical merits of performance-based weighting, with opponents arguing that stable individual differences between experts on which to base weights are too difficult to measure (Bolger & Rowe, 2015a).

Several (different) elicitation protocols developed over the last decades have been deployed successfully in political science, infrastructure planning, and volcanology (e.g., Aspinall, 2010; Aspinall & Cooke, 2013; Cooke & Goossens, 2008; Bolger et al., 2014b; O'Hagan et al., 2006). Most follow thoroughly documented methodological rules, but they differ in several aspects, including the way interaction between experts is handled, and the way an aggregated opinion is obtained from individual experts. There is no single, best structured expert judgment(SEJ) protocol; each has strengths and weaknesses (e.g., Bolger et al., 2014b; O'Hagan et al., 2006). An SEJ for uncertainty quantification can be roughly defined as a formalized procedure that:

1. Asks questions that have clear operational meanings;
2. Follows transparent methodological rules, i.e., is traceable, repeatable, and open to review;
3. Anticipates and mitigates some of the most important psychological and motivational biases;
4. Is thoroughly documented; and
5. Provides opportunities for empirical evaluation and validation.

These elements make SEJ accountable, transparent, and repeatable.[6] If expert opinion is used as scientific data, then it should be subject to the same kind of methodological rules for quality assurance that are applied to other types of empirical data (Cooke, 1991; Cooke & Goossens, 2000).

This article aims to contribute to, and complement, the ongoing discussion about the reliability and validity of different aggregation schemes for quantitative expert judgments. The most recent debate presented in Bolger and Rowe (2015a) and its accompanying commentaries (Cooke, 2015; Winkler, 2015; Morgan, 2015; Bolger & Rowe, 2015b) acted as a catalyst for writing this article. We will discuss the following three fundamental conjectures:

1. Prior performance can be a useful guide to future performance.
2. There is no clear relationship between the various measurable qualities of experts' judgments.
3. An extensive, facilitated discussion among experts, prior to eliciting their individual final estimates, improves experts' performance.

Various papers support the above conjectures and we will refer to them in due course; however, this article is not intended to be a literature review. We will complement the discussion on these three conjectures by analyzing the data collected from a large-scale elicitation exercise. The elicitation exercise used the IDEA protocol for SEJ (Hanea et al., 2016; Wintle et al., 2012). A brief overview of IDEA, its development, and the data can be found in Section 2. Section 3.1 discusses the first conjecture. It introduces ideas about differential weighting that are somewhat in contrast to those espoused in Bolger and Rowe (2015a,b). It also touches on the paramount importance of validating expert opinion. Section 3.2 discusses the second topic. Measuring the quality of expert judgments, and the experts' performance as uncertainty quantifiers, seems essential if one wants to treat expert data in the same way as scientific data. However, the various measures available for evaluating expert performance in making quantitative estimates reward different dimensions of performance, and may lead to differences in relative performance. Given that the question of whether and how to value the different dimensions of expert performance that exist arise regularly in discussions of expert performance evaluation, we believe an investigation into the question of whether any relationship tends to exist between these different measures in practice is of interest and worth discussing further. The discussion from Bolger and Rowe (2015a), Winkler (2015), Morgan (2015), and Bolger and Rowe (2015b) offers little background on this topic, aside from a clear criticism of the relative use of these measures in the weighting scheme proposed in Cooke (1991). Because of this lack of context, some claims made in Bolger and Rowe (2015a) and Tindale and Kluwe (2015) may be misinterpreted. Section 3.2 makes statements that reflect the authors' position on this topic and may seem in disagreement with certain claims made in Bolger and Rowe

[6] Note that this definition includes the one formulated by Roger Cooke: "The qualifier *structured* means that expert judgment is treated as scientific data, albeit scientific data of a new type" Cooke (2008).

(2015a) and Winkler (2015). Section 3.3 relates to the third topic. It follows and supports (to some extent) ideas presented in Bolger and Rowe (2015a,b). Arguments from Section 3 will be supported by and discussed in the light of data analysis. Section 4 concludes the article.

2. IDEA: A DELPHI-LIKE PROTOCOL WITH A TWIST

There are two main ways in which experts' judgments are pooled (Clemen & Winkler, 1999). One method is usually referred to as *behavioral aggregation*, and involves striving for consensus via discussion (O'Hagan et al., 2006). When experts (initially) disagree, the advocates of behavioral aggregation will advise on a discussion between the experts with divergent opinions, resulting in a "self-weighting" through consensus.[7] But this comes at the cost of verifiability and reproducibility. Moreover, while well-functioning behavioral groups offer participants the opportunity to share knowledge and correct misunderstandings, such interaction is frequently prone to biases, including overconfidence, polarization of judgments, and groupthink (Kerr & Tindale, 2011).

Mathematical rules used in *mathematical aggregation* provide a more explicit, auditable, and objective approach. A weighted linear combination of opinions is one example of such a rule. Equal weighting is often used mostly because of its simplicity (no justification for weights is required). Evidence also shows that the equal weighting scheme frequently performs quite well relative to more sophisticated aggregation methods (e.g., Clemen & Winkler, 1999), but not always (Cooke, 2015; Cooke & Goossens, 2008). There are reasons to believe that differential weighting based on anything (e.g., self-ratings, peer ratings, citation indices) other than performance should be avoided (Burgman et al., 2011; Cooke et al., 2008; Woudenberg, 1991). Probably the most well known and widely used version of a differential weighting scheme is the classical model or Cooke's model (CM) for SEJ (Cooke, 1991). Perhaps its most distinguishing feature is the use of calibration variables[8] to derive performance-based weights.

Mixed SEJ protocols combine behavioral and mathematical aggregation techniques (Ferrell, 1994). The most common mixed protocol is the Delphi method (Rowe & Wright, 2001), in which experts receive feedback over successive questionnaire rounds, in the form of other group members' judgments. Experts remain anonymous and do not interact with one another directly. Instead, a facilitator provides feedback between rounds. As originally conceived, the Delphi method strives to reach consensus after a relatively small number of rounds (Dalkey, 1969), though in modern usages, achieving consensus is not necessarily the primary aim (e.g., von der Gracht, 2012). While research supports a general conclusion that the Delphi method can improve accuracy over successive rounds, this is by no means guaranteed. Critical reviews suggest that even though individual judgments may converge after a number of rounds (von der Gracht, 2012), this convergence does not necessarily lead to greater accuracy (e.g., Murphy, Black, Lamping, Mckee, & Sanderson, 1998; Bolger, Stranieri, Wright, & Yearwood, 2011). Increasingly, it appears that for genuine improvement in judgment quality to take place between rounds, it is important that feedback includes the reasoning behind the expert's opinion (i.e., Bolger et al., 2011; Bolger & Wright, 2011). Moreover, while the Delphi permits sharing of rationales, it still places limitations on the degree to which information sharing can freely take place between experts, with the result that misunderstandings and linguistic ambiguity are more difficult to remove. For example, a recent study that attempted to employ the Delphi approach for use in eliciting probability distributions experienced considerable delays and difficulties, possibly arising due to the lack of expert–expert interaction and expert–facilitator interaction, leading to a decreased ability to clarify reasoning, assumptions, and the details of the survey instrument (Bolger et al., 2014a).

2.1. IDEA Protocol

The IDEA protocol described in this section synthesizes specific elements from all the approaches described above. In doing so, it aims to minimize

[7] However, where a group consensus judgment cannot be reached, individual expert distributions can be elicited and combined using a mathematical aggregation technique. Or, alternatively, where consensus is not the aim, the resulting spread of expert viewpoints following discussion can be maintained and presented to decisionmakers (Morgan, 2015).

[8] Calibration variables are variables taken from the experts' domain for which the true values are known, or will become known, within the time frame of the study (Aspinall, 2010).

the disadvantages of existing approaches and optimize their advantages. The majority of elements that characterize IDEA are not new; its most important contribution is in the structured approach to the combination of these elements (a similar approach is also proposed in Bolger & Rowe, 2015b, yet informally). The elements of the IDEA protocol and the reasons behind these choices are discussed in great detail in Burgman (2005) and Hanea et al. (2016). In this article, we will reiterate the specific steps, without repeating the discussion in the abovementioned references. IDEA is so-called because it encourages experts to investigate, discuss, and estimate, and concludes with a mathematical aggregation of judgments. It is a Delphi-like protocol in that experts give individual judgments over subsequent rounds, and facilitators provide feedback. In contrast to the traditional Delphi, IDEA does not seek consensus and cannot always ensure full anonymity. A diverse group of experts first answers questions without engaging in discussion. Experts are then provided with the judgments of their peers and have the opportunity to endorse agreements and discuss differences of opinion (unlike Delphi), allowing people to reconcile the meanings of words and context (e.g., Carey & Burgman, 2008). The discussion may be remote (on an online platform or over e-mail) or face to face. Facilitators encourage and moderate discussion between rounds. In typical face-to-face discussions, people are often forced to declare the identity of their estimates to others, exposing them to dominant individuals, halo effects, and potential pressure to conform. Conversely, people may refrain from defending their estimate in order to avoid identification. In IDEA, while the complete anonymity that characterizes a traditional Delphi process is lost, making the second estimate strictly anonymous largely mitigates these phenomena. This promotes the benefits of behavioral aggregation, while guarding against some of the most debilitating elements of group elicitation (Burgman, 2015; Montibeller & von Winterfeldt, 2015). Expert estimates obtained through the second round (postdiscussion) are mathematically aggregated. The method for mathematical aggregation can be chosen by the practitioners and it is not dictated by the IDEA protocol. Because IDEA does not strive for consensus, on the contrary, encourages independent anonymous final estimates, the dangers of reaching an artificial consensus are mitigated against. Differences of opinion are recorded and can be easily communicated to decisionmakers.

IDEA uses a structured procedure for questioning experts about uncertain variables, described in detail in Burgman (2015). When used to elicit continuous *quantities,* this procedure uses four questions to elicit the values of variables (corresponding to different quantiles), termed as a four-step format. A slightly different version of this procedure (for fixed quantiles) corresponds to the way questions are asked in the CM, so the mathematics used in the CM for scoring and aggregating expert judgments can be easily used in the last step of IDEA.

When eliciting *probabilities* of event occurrences, IDEA uses three questions, termed as a three-step format, one for the expert's *best estimate* about the probability of occurrence, and the other two for an interval that captures uncertainty around it. Other approaches, including CM, ask the experts to assign events to probability bins $b_i = (p_i; 1 - p_i)$, where p_i corresponds to the probability of occurrence. Bins can have the following form: $b_1 = (0.1; 0.9)$, $b_2 = (0.2; 0.8)$, $b_3 = (0.3; 0.7)$, etc., if the probability of occurrence scale is discretized into 10 intervals. An expert assigns an event to the b_2 bin if his or her best estimate (about the probability of occurrence) is anywhere between 0.1 and 0.2. So, these approaches only ask for best estimates, acknowledging the imprecision in the experts' judgments by allowing a fixed interval around them (equal to the respective bin's length). IDEA allows the expert to chose his or her own interval. The probabilities of binary variables can sometimes be interpreted in terms of relative frequencies. It is then legitimate to ask experts to quantify their degree of belief using a subjective distribution. In this case, the upper and lower bounds asked for in the three-step format may be thought of as quantiles of this subjective probability distribution. However, when the relative frequency interpretation is not appropriate (i.e., when the probability of a unique event is elicited), the three-step format may be criticized for lacking operational definitions for the upper and lower bounds in a probabilistic framework. Then, the elicited bounds can be thought of as an analog of the fixed bounds of the bins, when the probability of occurrence scale is discretized into bins (see above). The imprecision in the experts' judgments is decided and quantified by the experts themselves, rather than chosen by the analyst.

The main reason to elicit bounds in such cases is to improve thinking about the best estimates. Another reason is to provide an indication of relative participant uncertainty. Because the data analyzed in this article consist of elicited unique events' probabilities, most measures of performance that we consider and use to compare expert performances are calculated using the elicited best estimates, rather than the bounds.

Irrespective of the way the probabilities are interpreted, the mathematical apparatus of the CM and its specific feature of using calibration variables can be used in the last step of IDEA.

2.2. Data

The IDEA protocol was refined and tested as part of a forecasting "tournament" that started in 2011 as an initiative of the U.S. Intelligence Advanced Research Projects Activity (IARPA).[9] 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 judgments. 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 (Mellers et al., 2015; Ungar et al., 2012). The Good Judgment Project (GJP) team who won the tournament (Mellers et al., 2015) encouraged similar *think again* and *estimate lower and upper bounds* style practices (e.g., Mellers et al., 2015), though, to our knowledge, IDEA was unique in eliciting multiple, uncertainty judgments for each estimate. As opposed to other protocols, the IDEA protocol combines both feedback and (facilitated) interaction. The benefits of mathematical aggregation for combining forecasters' opinions were also investigated by other teams (e.g., Baron, Mellers, Tetlock, Stone, & Ungar, 2014), but none considered performance weighting.

The results shown in this article are based on the data gathered through this tournament. The data elicited with the IDEA protocol represent the answers to a subset of the questions developed by IARPA. All questions considered correspond to Bernoulli variables of the following sort, "Will the Turkish government release imprisoned Kurdish rebel leader Abdullah Ocalan before 1 April 2013?", which were answered using the three-step format outlined above. All questions usually resolved within 12 months; hence, they were suited for empirical validation studies. The elicitation took place remotely, initially via e-mail, and from the second year of the tournament through a dedicated website[10] that was set up for the participants to answer the questions and upload/download necessary materials. This website facilitated discussion via a discussion board that could be used to share relevant resources from outside the platform (e.g., web links), and to comment on and rate the quality of the information shared by others.

In answering the questions, experts followed the three-step format (see Fig. 1). The best estimate of each expert represents his or her subjective degree of belief in the probability of the occurrence of the event. Accepting that the judgments of experts are inherently imprecise and trying to formalize this imprecision, the bounds (about which experts are asked prior to asking the best estimate in order to avoid anchoring effects) are thought of as an uncertainty interval. The width of this interval provides an indication of how sure the expert is of his or her estimate.

The tournament operated on a yearly basis over the course of four years. Each year, new participants joined the IDEA groups, and other participants dropped out. There were 150 participants (over the four years) who answered at least one question (both rounds). Eight of these participants returned each year. The level of participants' expertise covered a very wide range from self-taught individuals with specialist knowledge to intelligence analysts.[11] A total of 155 questions were answered by at least one participant. However, no participant answered more than 96 questions. The participants were divided into groups and the number of groups varied across years to keep the number of participants per group fairly constant (typically 10). Starting from the third year, *supergroups* were formed,[12] composed of the best performing participants from the previous year.[13] The participants composing the supergroup were unaware of the fact that their group was in any way special. The number of participants

[9] http://www.iarpa.gov/index.php/research-programs/ace

[10] http://intelgame.acera.unimelb.edu.au/

[11] Because the study was conducted with people who may be considered in some contexts to be nonexpert participants, in the remainder of this article, whenever data from this study are concerned, we will use the term *participant* instead of *expert*.

[12] The GJP team tested an idea similar to the supergroup of top forecasters, but on a considerably greater scale, i.e., with the benefit of many more participants.

[13] Performance was measured using the average Brier score. This measure (defined later in this article) was imposed by the forecasting tournament rules and all participating teams had to use it.

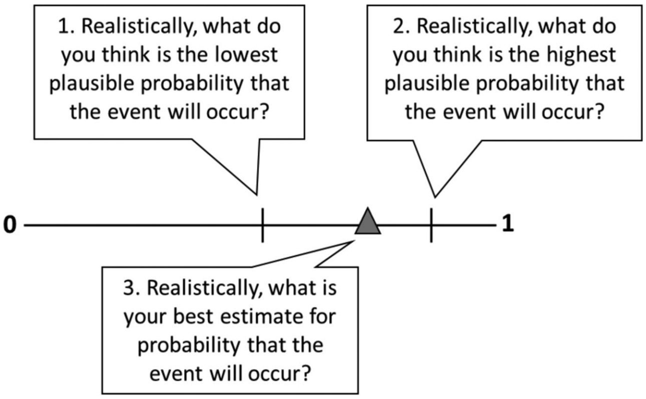
1. Realistically, what do you think is the lowest plausible probability that the event will occur?
2. Realistically, what do you think is the highest plausible probability that the event will occur?
0
1
3. Realistically, what is your best estimate for probability that the event will occur?

Fig. 1. A three-step format question.

composing the supergroup was equal to the number of participants from any other group. This construction was meant to help us accumulate evidence around whether prior performance is a useful guide to future performance on forecasting tasks.

3. PERFORMANCE MEASURES AND THE VALUE OF DISCUSSION

In some protocols using mathematical aggregation, the interaction between experts is limited to training and briefing (e.g., Cooke, 1991). It is generally believed that interaction between experts may harm aggregation because it induces correlation between judgments (e.g., O'Hagan et al., 2006). Nevertheless, when experts share and debate their knowledge during interaction, they establish a common and better understanding of what is being asked of them (e.g., Kerr & Tindale, 2011). We postulate that, in general, any additional dependence between judgments introduced through the discussion between rounds is justified by the increase in information resulting from discussion and by the reduction of misunderstandings and linguistic uncertainty. Results based on the data set described above support this postulate and are presented in Hanea, McBride, Burgman, and Wintle (2018). The dependence structures within and between the groups (before and after discussion) are investigated in Hanea et al. (2018) and found to be effectively the same.[14] Due to these findings, we feel comfortable in pooling the expert data from all groups and across all years to form a larger data set and hence permit more powerful statistical testing.

In evaluating the performance of probability forecasters, we assess their accuracy and their confidence. While alternative definitions of *accuracy* exist in the literature on evaluating subjective probabilities (e.g., Lichtenstein, Fischhoff, & Phillips, 1982; Moore, Tenney, & Haran, 2015; Yaniv & Foster, 1997), we will hereafter consider accuracy to be a measure of the degree of correspondence between the participants' predictions (which are the probabilities of event occurrences) and the observed outcomes. The average Brier score (Brier, 1950) is used to evaluate and compare participants' long-term accuracy. The Brier score for binary variables ranges from 0 for a perfect forecast to 2 for the worst possible forecast. Although the score can be computed based on a single forecast, performance on many questions is necessary to provide a reliable guide to accuracy. In our case, long-term performance is calculated as an average Brier score, using the participants' best estimates of the event probabilities (i.e., Fig. 1, box 3). The average Brier score can be further decomposed into two additive components called calibration and refinement (Murphy, 1973). The calibration component of the Brier score is one of the measures thoroughly investigated in Hanea et al. (2018) as a measure of the performance. In this article, we will also use it in Section 3.2. We treat the length of the uncertainty intervals provided by participants around their best estimates as a measure of their confidence. Under this interpretation, a wider interval denotes decreased confidence, while a narrower interval denotes an increased level of confidence. The length of these intervals can also be thought of as a measure of informativeness: wide intervals are uninformative.

[14] The dependence between experts' answers is measured using the average correlation coefficient of experts within each group and conditional on the outcomes. For details of the calculations, we refer to Hanea et al. (2018).

A more extensive discussion about different concepts and measures of accuracy, calibration, and confidence, and different uses of these terms, can be found in Hanea et al. (2016). A description of the calibration term of the Brier score and other measures of confidence and informativeness calculated for the participants' assessments are discussed and analyzed in Hanea et al. (2018). For the sake of conciseness, we shall not repeat these discussions here, but merely refer the reader to the two mentioned references.

3.1. Prior and Future Performance

Each year we compared the performance of the different groups. When we consider groups of participants, we are interested in the group accuracy as measured by the average Brier score of the group-aggregated estimates. Specifically, for each group, the best estimates of the participants for a given question are averaged, and a Brier score per question, per group is calculated. The average Brier score (across all questions answered by at least one participant from that group) is reported as the accuracy measure of the group. Each year we compared the accuracy of the equally weighted opinions of the groups after discussion, using a within-subject design.

In one year alone (2013–2014), we had sufficient data to calculate differential weights using the CM (Cooke, 1991). We used the best estimate responses of the participants in the IDEA supergroup 1 to calculate calibration and information scores for discrete variables defined in Cooke (1991) and reiterated in Hanea et al. (2018). Performance-based weights were then derived based on these calibrations scores, and the best estimates of the participants were weighted accordingly to obtain the differentially weighted response per question, per group. Brier scores were then calculated per question, per group, for the new performance-weighted responses. Fig. 2 shows the average Brier scores of the equally weighted combination of judgments for each group after discussion, together with standard error bars, for 2013–2014. The performance-based differentially weighted combination of the supergroup participants' judgments is shown in the same figure.

Although not statistically significant (as indicated by the t-test), the supergroup G1 outperformed the other groups of participants, suggesting that prior performance is a useful guide to future performance on similar estimation tasks (out of sample). The same was observed for all the other years of the tournament (except the first year, where no supergroup was used). This finding is in agreement with the findings of Mellers et al. (2014).

Employing the CM, a slight improvement in the performance is achieved using this weighting scheme (when compared to equal weighting). Even though not significant, this improvement provides additional evidence to support prior performance as a useful guide to future performance (in sample).

Asking questions that resolve in the near future and using performance measures allows for validation of the utility of predictions from the particular group of experts being consulted, which is an essential step for any model. Several methods for validating expert opinion have now been used and proposed (Cooke et al., 2014; Eggstaff, Mazzuchi, & Sarkani, 2014), mainly by the advocates of the CM.[15] These methods include in-sample validation and cross-validation, if variables of interest (the variables for which expert judgment is required) are unobservable on relevant time scales. We conjecture that well-designed cross-validation studies offer a credible approach to validating expert judgment and the utility of performance-based weighting (Cooke et al., 2014; Eggstaff et al., 2014), certainly preferable to no validation at all.[16] The most recent cross-validation study performed by Colson and Cooke (2017) offers a high level of support for the idea that prior performance is a useful guide to future performance: for 54 out of 73 professional expert elicitations, the performance-based combination of experts' judgments outperformed the equally weighted combination.

We acknowledge and share the concerns of Bolger and Rowe (2015a,b) about true validation studies (i.e., where the value of expert performance-based weighting for predicting the target variables of interest in a genuine application of expert knowledge is able to be evaluated post-hoc). To this end, we hope that the out-of-sample validation results obtained with IDEA and presented here and in Hanea et al. (2016) and Hanea et al. (2018) strengthen the arguments discussed in Eggstaff et al. (2014), Cooke et al. (2014), Cooke (2015), and Colson and Cooke (2017).

Moreover, when judgments conflict, one will (usually) turn out to be "better" than the other. How

[15] The data used in many of these studies are freely available online at http://rogermcooke.net/.

[16] In other words (using the terminology in Bolger & Rowe, 2015b), we argue that starting from *data* in our journey toward data will get us further than starting from a void.

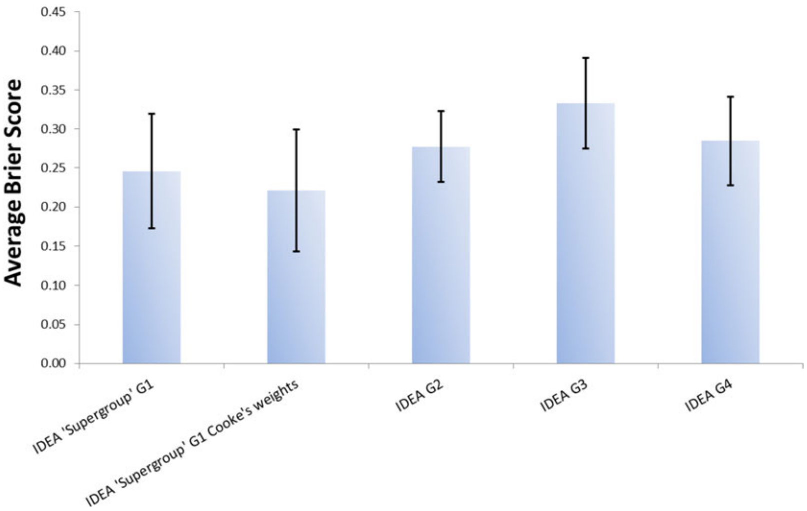

Fig. 2. Forcasting tournament 2013–2014.

can we know, *a priori*, whose opinion is valid (or more reliable) and should be assigned more weight in an aggregation? Previous studies (e.g., Burgman et al., 2011; Cooke et al., 2008) assure us that we cannot, and the only way of knowing is by assessing prior performance on similar tasks. Morgan (2015) argues that the "outlier's views should not get masked by combining multiple experts." While we agree with this statement in terms of the importance of retaining experts' unaggregated opinions along with any group aggregate judgment in order to avoid masking individual differences of opinion, we argue that aggregation is often necessary, and that where information on past performance is available, it can provide a valuable way to discriminate between (and aggregate accordingly) situations when an outlier's views are likely to be worthwhile, and those where they are not.

The topics of measuring performance and differential weighting are obviously interrelated and the above discussion suggests that there are good validated reasons to base differential weighting schemes on measures of prior performance on similar tasks. The question that follows naturally is which measures of performance should one use. If experts' answers have one quality (e.g., accuracy), would this guarantee other qualities (e.g., informativeness)? Does this relationship depend on the way we define and measure these qualities? These questions are not new, and such relationships between various measures of performance have been investigated before in the context of subjective probability (distributions) elicitations. The next section adds yet another case study to the existing literature and the ongoing debate.

3.2. The Correlation Between Performance Measures

Ideally, any increase in an expert's confidence would be matched with an increase in his or her accuracy or calibration. However, research across many different measures of confidence, accuracy, and calibration has found that these measures are often poorly correlated with one another (e.g., Aspinall & Cooke, 2013; Griffin & Tversky, 1992; Lichtenstein et al., 1982; Moore et al., 2015).

In asking experts to quantify parameters and their uncertainty around these estimates, SEJ methods rely on each expert's ability to realistically quantify the limitations of his or her knowledge. When experts underestimate the uncertainty in their judgments, apparent overconfidence occurs, leading to faulty assessments on which hazardous decisions may be based. Even though less uncertainty around a judgment (more confidence) often translates to greater informativeness, we believe that certainty should only be valued when associated with well-calibrated assessments. This contrasts with our interpretation of the views expressed in Winkler (2015), who seemingly views narrow (*informative*) distributions as more important than calibrated ones.[17]

This being said, we would like to emphasize that both measures are necessary in evaluating the performance, and they should be evaluated in tandem whenever possible to avoid rewarding high performance in one measure at the expense of the other.

[17] "I feel that in most cases, informativeness is more important than calibration."

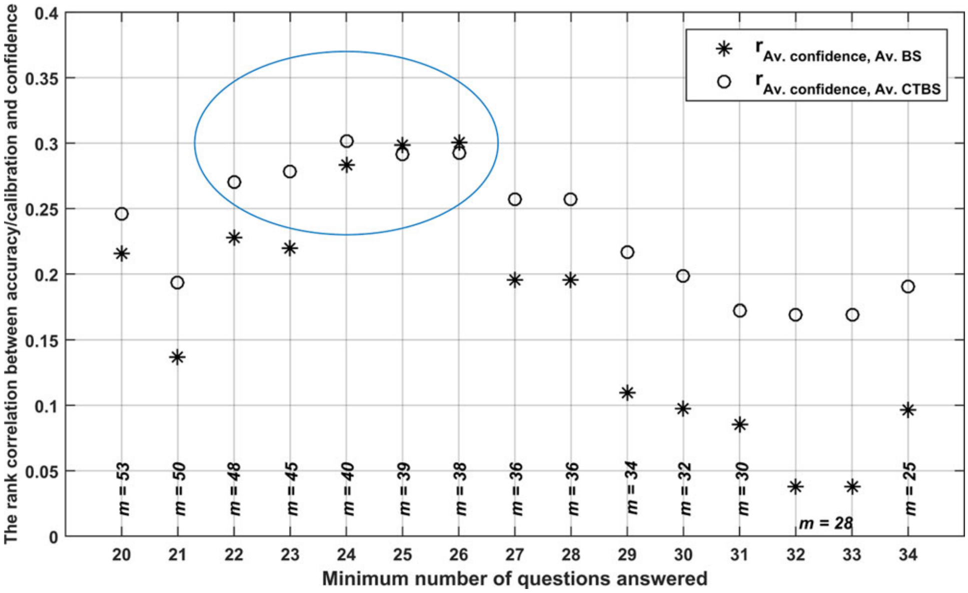

Fig. 3. Confidence versus accuracy and calibration.

In this section, we investigate the relationship between accuracy (measured by the average Brier score) and confidence (as measured by uncertainty interval length) using the data set described in Section 2.2. Because the average Brier score can be decomposed such that we can measure calibration as well, and the calibration component of the Brier score is one of the measures thoroughly investigated in Hanea et al. (2018), we will extend the analysis and investigate the relationship between calibration and confidence using the same data set.

Fig. 3 looks at these two relationships in terms of rank correlations. We have selected the participants who answered at least 20 questions to reduce the variance of the average measures of performance estimators.

The rank correlation between the average Brier score and the average confidence is represented by stars. The rank correlation between the calibration term of the average Brier score and the average confidence is represented by circles. These correlations are shown as a function of the minimum number of questions (n) answered by participants. We denote m the number of participants who answered at least n questions. We added n's corresponding m just above the X-axis (e.g., for $n = 20$, $m = 53$). Each rank correlation is calculated based on m samples, and the value of m determines if the correlation is statistically different than zero. The only values found statistically different than zero at the 0.1 significance level, when performing a two-tailed statistical test (Ramsey, 1989), are the ones enclosed by the ellipse from Fig. 3. However, *none* of the correlations shown in Fig. 3 are statistically different than zero at the 0.05 significance level.

Based on the above analysis, there is very little reason to believe that an increase in an expert's confidence would correspond to an increase in his or her accuracy or calibration. This warns us against relying too heavily on any one dimension of expert performance when evaluating experts under the assumption that adequate performance according to one metric of expert performance (e.g., informativeness) would suggest it for others (e.g., calibration). Moreover, if these evaluations are used to form weights, the weights should reflect this dual importance and evaluate for adequate performance across multiple dimensions of expert performance.

3.3. Does Discussion Help?

Bolger and Rowe (2015b) recommend behavioral aggregation as an alternative to the CM. We agree with only one aspect of this recommendation, the possibility of facilitated discussion. IDEA assimilated the extensive discussion aspect of the behavioral aggregation, but strongly advises against striving for consensus and the "self-weighting" associated with it. The reasons for this advice are briefly addressed in this article (see Section 2.1) and thoroughly discussed in Burgman (2015).

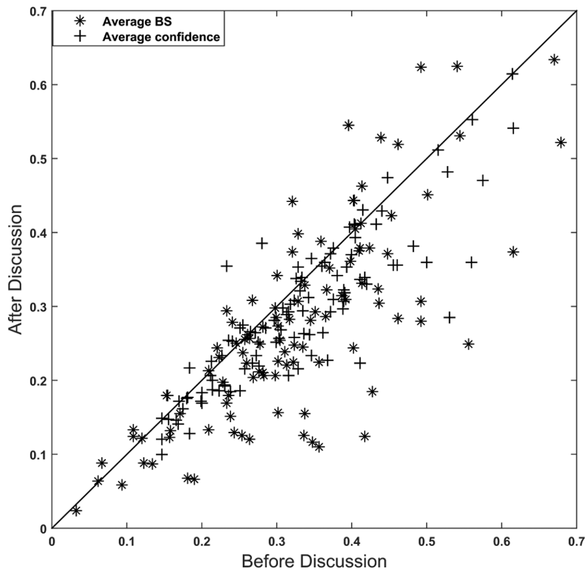

Fig. 4. Average Brier scores and average confidence before and after discussion. Both measures have similar possible ranges, allowing for presentation on the same plot. Averages are calculated for each participant across all questions.

We consistently find (throughout our studies) that sharing information between rounds appears to improve, on average, both accuracy and confidence. Fig. 4 is one more example of this phenomenon. The average Brier scores of all participants before discussion are plotted against their average Brier scores after discussion. These are represented by stars. If participants do not change their minds (about any question) in the second round, their average Brier score does not change (so the point representing this pair of scores will fall on the main diagonal). The majority of points in this scatter plot fall under the main diagonal, indicating that the majority of participants who change their mind, changed in the direction of the realized outcome. The average confidence of the participants before discussion is plotted against their average confidence after discussion in the same plot. These are represented by plus (+) signs. The same general pattern can be observed. Confidence was greater after discussion.

Because our study does not include a control group that provides both first- and second-round judgments but does not undertake discussion in-between, we cannot rule out the possibility that any changes in performance observed are due to iteration alone, rather than the discussion that took place during each iteration. However, this within-subject quasi-experimental approach is not unusual in studying the performance of Delphi-related methods (e.g., Rowe & Wright, 1999). Our result complements other work that reveals the potential importance that any feedback provided includes the reasoning behind the expert's opinion (e.g., Bolger et al., 2011; Bolger & Wright, 2011) in order to contribute to improvements in the performance between different rounds on estimation.

In the last three years of the tournament, the data were collected through a dedicated website. This gave us the opportunity to collect information about the participants' level of engagement in the discussion phase. During the discussion phase, the participants contributed comments and shared resources (i.e., additional relevant information such as links to news articles, scientific papers, blog posts, and additional data) with the group. After resources were added on the group discussion board, they could be viewed and rated. We recorded the number of these actions. These activities define to some extent the engagement of participants in the discussion between the rounds, and we can look at the relationship between measures of performance and these activity measures. We have (again) selected the participants who answered at least 20 questions to reduce the variance of the average measures of performance and activity estimators. We first look at the rank correlation between the average Brier score and: (1) the average number of comments (star), (2) the average number of added resources (circle), (3) the average number of viewed resources (plus), and (4) the average number of rated resources (diamond). These correlations are shown as a function of the minimum number of questions (n) answered by (m) participants. Just like in the previous section, each rank correlation is calculated based on m samples, and the value of m determines if the correlation is statistically different than zero. Fig. 5 shows these correlations. All correlations below the horizontal line are significantly different than zero at the 0.05 significance level.

There is strong negative correlation between the average activity level (as defined by the number of comments made, and by the number of added and viewed resources) and the average Brier score, which means that higher levels of activity correspond to lower, hence better, scores. This result supports our conjecture that discussion improves (at least one aspect) the performance.

However, when looking at the relationships between activity and the length of the uncertainty intervals, the rank correlations are much smaller

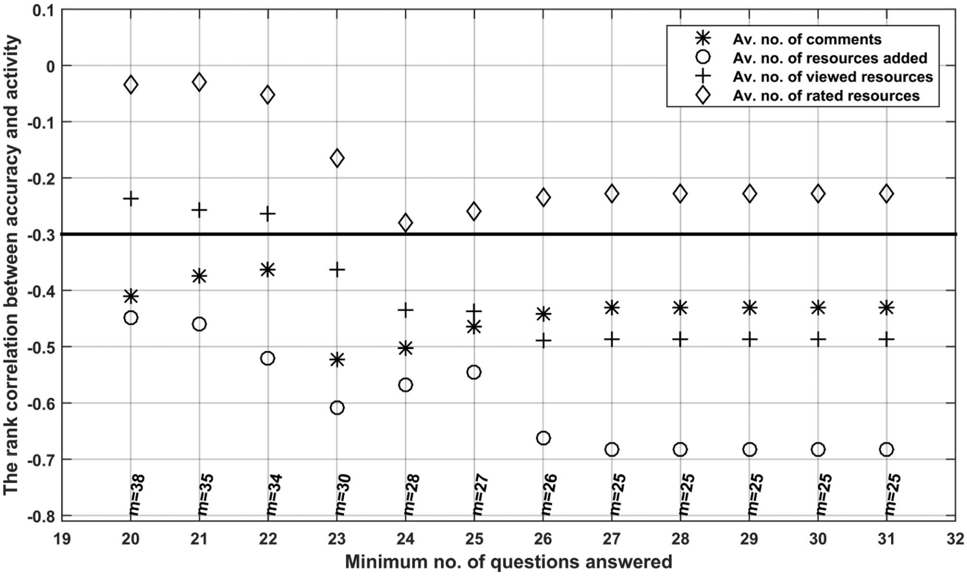

Fig. 5. Average Brier scores and average level of activity.

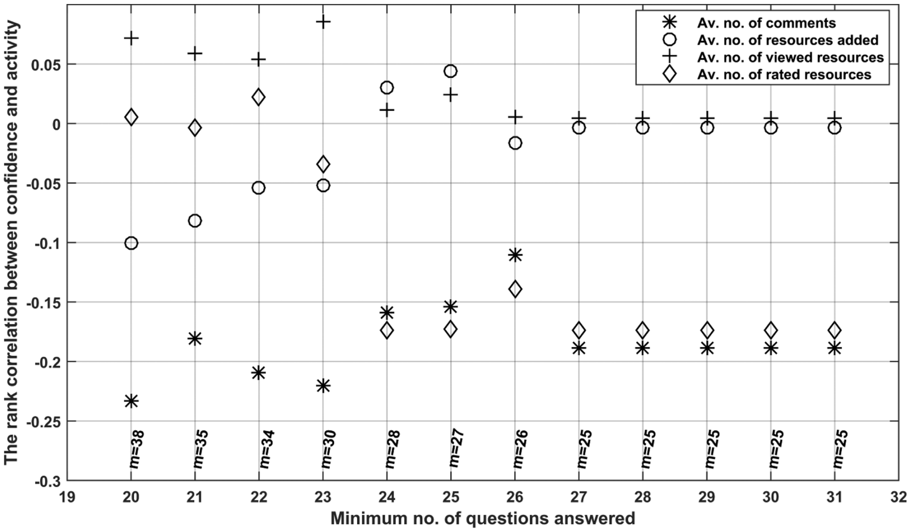

Fig. 6. Average confidence and average level of activity.

and none is significantly different than zero (see Fig. 6). Given this apparent weaker relationship observed between activity and confidence, we speculate that the shift (increase) in average participant confidence following discussion (Fig. 4) may be the result of receiving feedback (i.e., exposure to the intervals of the other group members) rather than discussion, as this is the element of the discussion phase that would not necessarily relate to level of online activity and interaction with other group members.

4. DISCUSSION

The aggregation of expert judgments into a single, transparently constructed estimate is one of the most important steps in eliciting and using expert judgments. Equal weights are intuitively appealing because there is no need to defend differential weights, they are easy to understand, and all experts feel that they contribute equally. However, one of the most important lessons of empirical studies over the last decade is that an expert's performance on technical questions may be predicted to some extent by the history of his or her performance on similar questions previously (Budescu & Chen, 2015; Colson & Cooke, 2017; Mellers et al., 2015). Taking advantage of this phenomenon, Cooke's approach to differential weighting circumvents the difficult issue of the different scales associated with different kinds of questions, and assimilates each expert's confidence and statistical accuracy into a single weight. The result is the potential for an improvement in group performance over equal-weighted-based aggregation methods. Our results suggest that even in the relatively difficult conditions imposed in answering binary questions on the outcomes of geopolitical events, performance-based differential weights calculated using Cooke's method may improve upon the performance of groups, even those composed of comparatively higher performing forecasters, such as those in to the supergroups.

We agree with Bolger and Rowe (2015a) that the way forward for expert knowledge research involves additional experiments and analyses to further test the benefits of performance-weighting schemes against equal weighting and behavioral aggregation. Our article provides one such contribution by drawing on recent experimental findings. The discussion papers (Bolger & Rowe, 2015a,b; Winkler, 2015; Morgan, 2015) were published in 2015, and arguably drew more heavily on an older body of literature, findings, and established positions around the relative merits of performance weighting. We have endeavored to support this dialog by contributing more recent findings and methodologies for drawing on the benefits of both behavioral aggregation (here via facilitated discussion) and performance weighting. Both our study and that of the GJP present empirical evidence incorporating out-of-sample testing, indicating that prior performance predicts future performance. We argue that this provides grounds for the use and the continued investigation of performance-based weighting.

Our article also explored the contested question of whether social interaction between group members erodes judgment quality (Lorenz et al., 2011; Tindale & Kluwe, 2015). As with IDEA, the GJP also made use of structured discussion between forecasters and found that it improved performance (Mellers et al., 2014). In their case, group members were known to each other via their online identifiers only. They could offer rationales and critiques and share information, including their predictions (although, unlike in IDEA, formal feedback on initial group member predictions was not provided). The results of both approaches illuminate the value of facilitated conversations between group participants in reconciling language-based misunderstandings and interpretations of evidence. They also both suggest that controlled interaction and feedback can enhance performance, without overwhelming these benefits with dysfunctional group dynamics. Whether or not these findings hold up in different settings, such as those where group members are motivated to seek status or guard information, remains to be seen.

We conclude by reiterating the recommendation for risk analysts and decisionmakers to use consistent, transparent, and repeatable approaches to SEJ. As data accumulate over time, the strengths and weaknesses of these new methodologies for aggregating expert judgments will become clearer and help to ensure better judgments in the future.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their thorough and numerous comments.

The authors declare no conflict of interest.

REFERENCES

Aspinall, W. (2010). A route to more tractable expert advice. *Nature*, *463*, 294–295.

Aspinall, W., & Cooke, R. M. (2013). Quantifying scientific uncertainty from expert judgement elicitation. In J. Rougier, S. Sparks, & L. Hill (Eds.), *Risk and uncertainty assessment for natural hazards*, Chapter 10 (pp. 64–99). Cambridge: Cambridge University Press.

Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., & Ungar, L. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*, *11*, 451–468.

Bolger, F., Hanea, A., O'Hagan, A., Mosbach-Schulz, O., Oakley, J., Rowe, G., & Wenholt, M. (2014a). Case study in plant health. In EFSA Journal 2014: Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment. Parma, Italy: European Food Safety Authority (EFSA).

Bolger, F., Hanea, A., O'Hagan, A., Mosbach-Schulz, O., Oakley, J., Rowe, G., & Wenholt, M. (2014b). Guidance on expert knowledge elicitation in food and feed safety risk assessment. *EFSA Journal*, *12*(6), 3734. https://doi.org/10.2903/j.efsa.2014.3734.

Bolger, F., & Rowe, G. (2015a). The aggregation of expert judgment: Do good things come to those who weight? *Risk Analysis*, *35*, 5–11.

Bolger, F., & Rowe, G. (2015b). 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.

Bolger, F., Stranieri, A., Wright, G., & Yearwood, J. (2011). Does the Delphi process lead to increased accuracy in group-based judgmental forecasts or does it simply induce consensus amongst judgmental forecasters? *Technological Forecasting and Social Change*, *78*(9), 1671–1680.

Bolger, F., & Wright, G. (2011). Improving the Delphi process: Lessons from social psychological research. *Technological Forecasting and Social Change*, *78*, 1500–1513.

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

Budescu, D. V., & Chen, E. (2015). Identifying expertise to extract the wisdom of crowds. *Management Science*, *61*(2), 267–280.

Burgman, M. A. (2005). *Risks and decisions for conservation and environmental management*. Cambridge: Cambridge University press.

Burgman, M. A. (2015). *Trusting judgements: How to get the best out of experts*. Cambridge: Cambridge University Press.

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.

Burgman, M. A., McBride, M., Ashton, R., Speirs-Bridge, A., & Flander, L., Wintle, B., … Twardy, C. (2011). Expert status and performance. *PLoS One*, *6*, e22998.

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.

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

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

Cooke, R. M. (2008). Special issue on expert judgment. *Reliability Engineering & System Safety*, *93*(5), 655–656.

Cooke, R. M. (2015). The aggregation of expert judgment: Do good things come to those who weight? *Metron*, *35*, 12–15.

Cooke, R. M., ElSaadany, S., & Huanga, X. (2008). On the performance of social network and likelihood-based expert weighting schemes. *Reliability Engineering and System Safety*, *93*(5), 745–756.

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

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

Cooke, R. M., Wittmann, M. E., Lodge, D. M., Rothlisberger, J. D., Rutherford, E. S., Zhang, H., & Mason, D. M. (2014). Out-of-sample validation for structured expert judgment of Asian carp establishment in Lake Erie. *Integrated Environmental Assessment and Management*, *10*(4), 522–528.

Dalkey, N. (1969). An experimental study of group opinion: The Delphi method. *Futures*, *1*, 408–426.

Eggstaff, J. W., Mazzuchi, T. A., & Sarkani, S. (2014). The effect of the number of seed variables on the performance of Cooke's classical model. *Reliability Engineering and System Safety*, *121*, 72–82.

Ferrell, W. (1994). Discrete subjective probabilities and decision analysis: Elicitation, calibration and combination. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 410–451). New York: Cambridge University Press.

Griffin, D. W., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, *24*, 411–435.

Hanea, A., McBride, M., Burgman, M., & Wintle, B. (2018). Classical meets modern in the idea protocol for structured expert judgement. *Journal of Risk Research*, *21*, 417–433.

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

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, *39*(4), 341–350.

Kerr, N. L., & Tindale, R. S. (2011). Group-based forecasting?: A social psychological analysis. *International Journal of Forecasting*, *27*, 14–40.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). Cambridge: Cambridge University Press.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 9020–9025.

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., … Tetlock, P. (2015). 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. (2015). Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science*, *10*(3), 267–281.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., & Tetlock, P,. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, *25*, 1106–1115.

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

Moore, D. A., Swift, S. A., Minster, A., Mellers, B., Ungar, L., Tetlock, P., & Tenney, E. R. (2015). *Confidence calibration in a multi-year geopolitical forecasting competition*. Working Paper.

Moore, D. A., Tenney, E. R., & Haran, U. (2015). Overprecision in judgment. In G. Keren & W. Wu (Eds.), *The Wiley Blackwell handbook of judgment and decision making* (pp. 182–212). Chichester: John Wiley & Sons, Ltd.

Morgan, M. G. (2015). Our knowledge of the world is often not simple: Policymakers should not duck that fact, but should deal with it. *Risk Analysis*, *35*, 19–20.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, *12*(4), 595–600.

Murphy, M. K., 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., Buck, C. E., Daneshkhah, A., Eiser, J., Garthwaite, P., Jenkinson, D., … Rakow, T. (2006). *Uncertain judgements: Eliciting experts' probabilities*. London: Wiley.

Ramsey, P. H. (1989). Critical values for Spearman's rank order correlation. *Journal of Educational Statistics*, *14*(3), 245–253.

Rowe, G., & Wright, G. (1999). The Delphi technique as a forecasting tool: Issues and analysis. *International Journal of Forecasting*, *15*(4), 353–375.

Rowe, G., & Wright, G. (2001). Expert opinions in forecasting: The role of the Delphi technique. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 125–144). Norwell: Kluwer Academic Publishers.

Slovic, P. (1999). Trust, emotion, sex, politics, and science: Surveying the risk-assessment battle field. *Risk Analysis*, *19*, 689–701.

Sunstein, C. R. (2004). *Group judgments: Deliberation, statistical means, and information markets*. John M. Olin Law & Economics Working Paper No. 219, University of Chicago Law School.

Tindale, R. S., & Kluwe, K. (2015). Decision making in groups and organizations. In G. Wu & G. Keren (Eds.), *The Wiley Blackwell handbook of judgment and decision making* (pp. 849–874). New York/Chichester, UK: Wiley.

Ungar, L. H., Mellers, B., Satopaa, V. A., Baron, J., Tetlock, P. E., 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).

von der Gracht, H. (2012). Consensus measurement in Delphi studies: Review and implications for future quality assurance. *Technological Forecasting & Social Change*, *79*, 1525–1536.

Winkler, R. L. (2015). Equal versus differential weighting in combining forecasts. *Risk Analysis*, *35*, 16–18.

Wintle, B., Mascaro, M., Fidler, F., McBride, M., Burgman, M., Flander, L., … Manning, B. (2012). The intelligence game: Assessing Delphi groups and structured question formats. In *Proceedings of the 5th Australian Security and Intelligence Conference.*

Woudenberg, F. (1991). An evaluation of Delphi. *Technological Forecasting and Social Change*, *40*, 131–150.

Yaniv, I., & Foster, D. P. (1997). Precision and accuracy of judgmental estimation. *Journal of Behavioral Decision Making*, *10*(1), 21–32.