2012

The Intelligence Game: Assessing Delphi Groups and Structured Question Formats

Bonnie Wintle
University of Melbourne

Steven Mascaro
Bayesian Intelligence,Melbourne

Fiona Fidler
University of Melbourne

Marissa McBride
University of Melbourne

Mark Burgman
University of Melbourne

See next page for additional authors

Authors

Bonnie Wintle, Steven Mascaro, Fiona Fidler, Marissa McBride, Mark Burgman, Louisa Flander, Geoff Saw, Charles Twardy, Aidan Lyon, and Brian Manning

THE INTELLIGENCE GAME: ASSESSING DELPHI GROUPS AND STRUCTURED QUESTION FORMATS

Bonnie Wintle[1], Steven Mascaro[2], Fiona Fidler[1], Marissa McBride[1], Mark Burgman[1]
Louisa Flander[3], Geoff Saw[4], Charles Twardy[5], Aidan Lyon[6] & Brian Manning[1]

[1] Australian Centre of Excellence for Risk Analysis (ACERA), School of Botany,
University of Melbourne, Australia
bonnie.wintle@unimelb.edu.au, fidlerfm@unimelb.edu.au, m.mcbride@student.unimelb.edu.au,
markab@unimelb.edu.au, brian.manning@unimelb.edu.au

[2] Bayesian Intelligence,
Melbourne, Australia
sm@voracity.org

[3] ACERA, School of Population Health,
University of Melbourne, Australia
l.flander@unimelb.edu.au

[4] Knowledge, Information & Learning Lab, School of Psychological Sciences,
University of Melbourne, Australia
gsaw@unimelb.edu.au

[5] C4I Center,
George Mason University, USA
ctwardy@c4i.gmu.edu

[6] Department of Philosophy,
University of Maryland, USA
alyon@umd.edu

Abstract

In 2010, the US Intelligence Advanced Research Projects Activity (IARPA) announced a 4-year forecasting "tournament". Five collaborative research teams are attempting to outperform a baseline opinion pool in predicting hundreds of geopolitical, economic and military events. We are contributing to one of these teams by eliciting forecasts from Delphi-style groups in the US and Australia. We elicit probabilities of outcomes for 3-5 monthly questions, such as: Will Australia formally transfer uranium to India by 1 June 2012? Participants submit probabilities in a 3-step interval format, view those of others in their group, share, rate and discuss information, and then make a second private judgement. Performance is assessed using Brier scores.

After Year 1, we ranked second of five teams in the competition. The Brier scores from the US Delphi groups improved on the baseline scores by 10%, the prediction market operated by our team in the US beat the baseline by 47%, and the Australian Delphi groups outperformed the baseline by 51% (answering different, matched questions to the US groups). The Australian groups were more socially and demographically diverse than the US groups. Group diversity may be an important factor determining the forecasting performance of the aggregated predictions.

Keywords
Delphi method, forecasting, judgement, subjective probability, uncertainty

INTRODUCTION

What is the Intelligence Game?

The Intelligence Advanced Research Projects Activity (IARPA) is an initiative of the US Office of the Director of National Intelligence. In 2010, IARPA announced a program called ACE (Aggregative Contingent Estimation), which aims "to dramatically enhance the accuracy, precision, and timeliness of forecasts for a

broad range of event types, through the development of advanced techniques that elicit, weight, and combine the judgments of many intelligence analysts." The project takes the form of a "competition" involving five groups, each of which must outperform a baseline opinion pool in predicting hundreds of geopolitical, economic and military events, over a four year period. Failing to beat the baseline could potentially result in elimination.

Through the Australian Centre of Excellence for Risk Analysis (ACERA) at the University of Melbourne, we are contributing to one of these teams, led by colleagues at George Mason University in the US. Our joint team is called DAGGRE—or *D*ecomposition-Based Elicitation & *Aggre*gation. ACERA's role is to elicit forecasts from groups in the US and Australia using a structured Delphi-style iterative elicitation process. In the first year of the competition, we ran four groups in Australia and three in the US, each containing 6-10 participants. In Year 2, we are running ten groups (five in each country), with 15-20 participants in each.

Why is this important?

Unfortunately, empirical evidence suggests that political experts are not particularly good at predicting the future–performing only marginally better than random chance (Tetlock, 2005). Until now, there has been no systematic check on the accuracy of Intelligence forecasts, and there is little evidence that forecasts from the Intelligence community would be any better than those of the political experts. Poor forecasts might arise from 'overpredicting', leading to false positives (or falsely anticipating an outcome, such as finding Weapons of Mass Destruction), and 'underpredicting', leading to false negatives (or failing to anticipate an event, such as 9/11). Together, these errors may result in "accountability ping pong" (Tetlock & Mellers, 2011), where Intelligence analysts are blamed for an error in one direction and overcompensate in the other direction. The absence of a reliable, structured approach to elicitation and aggregation allows for these biases to emerge in forecasts.

How can we improve forecasts?

Mitigating biases with well-structured elicitation

Three well-studied cognitive biases typically lead our thinking astray. They include *overconfidence*—where people think they know more than they actually do—*confirmation bias*—where people seek evidence that confirms a pre-existing belief—and *anchoring*—where people rely too heavily on some implicitly suggested reference point (such as a number contained in the question description) (Kahneman, Slovic, & Tversky, 1982). We structure our elicitation to minimise these biases in two ways: (a) by using a 3-step question format that asks for the (i) highest probability, (ii) lowest probability, and (iii) best guess, and (b) by engaging participants in a Delphi-style judgement iteration process (see Methods).

Empirical findings from cognitive psychology underpin our 3-step question format (Speirs-Bridge et al., 2010). Dividing the question into multiple steps improves the chances that people will think about different kinds of evidence (Soll & Klayman, 2004). This helps avoid answers that are too precise, that is, intervals that are too narrow. Focussing on reasons that make an event likely when answering the 'highest' probability question and conversely, focussing on reasons that make an event unlikely when answering the 'lowest' probability helps to overcome confirmation bias and reduces overconfidence (Koriat, Lichtenstein, & Fischhoff, 1980). Question order also affects judgments. Starting with a single best guess leads to anchoring, where participants tend to simply add or subtract 10% (for example) to answer (i) and (ii). This produces intervals that are overly narrow, compared with when the interval is elicited first (Soll & Klayman, 2004).

We implement elicitation via a modified Delphi method, a procedure developed in the mid 1940s to improve forecasting about technology during the Cold War. A standard Delphi process involves a small group of experts who provide forecasts over two or more rounds. Between rounds, a facilitator provides an anonymous summary of the experts' forecasts together with reasons behind their judgments. Experts can revise their forecasts in subsequent rounds (Linstone & Turoff, 1975). Our groups are not anonymous, to facilitate more direct discussion.

In viewing the summary of individual group members' forecasts, participants are receiving feedback about group variability and the group average. Comparing their own judgments with the group average in itself can improve estimation performance (Wintle, Fidler, Vesk, & Moore, 2012). Observing variability might also lead to interesting discussion points: Why might one individual's probability be so much higher or lower than the

rest of the group? Why are some individuals' intervals so precise and others so wide? Do those individuals have special knowledge that could benefit the group?

Using group wisdom

An important component of our approach is to focus on groups, rather than individuals. Under the right conditions, groups lend the powerful quality of 'collective intelligence' (Surowiecki, 2005). However, groups frequently fail to outperform the average of individual judgments (for a review, see Hastie, 1986; Kerr & Tindale, 2011). Group judgments may be prone to 'group think' (Janis, 1982), where groups make more extreme and risky judgments than would individuals, and they also tend to become overconfident (Sniezek, 1992). Groups can also be dominated by extroverts, rather than experts (Bonner, Sillito, & Baumann, 2007) and led astray by 'halo-effects' (Thorndike, 1920), where the most experienced or charismatic person is uncritically followed by the other group members.

To our surprise, the prognosis in recent research is not so bleak. The positive effect of group interaction is especially clear in *quantitative* judgement tasks (Schultze, Mojzisch, & Schulz-Hardt, 2012), such as estimating the outcome of sporting events and elections—as opposed to brainstorming and creative tasks. Under certain conditions, negotiated group judgments (behavioural aggregation) can even outperform averaged individual judgments (mathematical aggregation) (Bonner & Baumann, 2012).

So what conditions might best utilise the wisdom of the crowd? First, elicitation should be structured to minimise perverse outcomes of group dynamics. For example, anonymity of initial assessments in Delphi groups may reduce 'halo-effects'. Second, the question must be difficult enough so that individual judgments err. Otherwise, an individual judgement would be sufficient. Third, the group should be sufficiently diverse (Page, 2008) so that biases are roughly evenly distributed either side of the 'true value'. That is, the errors of individuals cancel each other out. Within the specific context of intelligence and crisis management, Hackman (2011) outlined six factors that led to improved group decision making: (i) Being a 'real team' (the importance of using an existing intact social system), (ii) Having a compelling purpose, (iii) The right people, (iv) Clear norms of conduct, (v) A supportive organisational environment and (vi) Team-focussed coaching—the importance of group facilitation.

Studies of group forecasts suggest that group structuring (Ang & O'Connor, 1991), group technique used (Sniezek, 1989), and method of combining different views (Kerr & Tindale, 2011; Önkal, Lawrence, & Sayım, 2011) all influence group performance. While the benefits of mathematical versus behavioural aggregation will vary with the task, we believe that there is good evidence to support the use of group interaction under the right conditions.

METHODS

Participants

In Year 1 of the competition, we recruited and established four groups in Australia and three in the US, each containing 6-10 participants. The only requirement for participation was an interest in world affairs. Most participants (85%) held university degrees. We balanced gender, age, discipline and occupation across the Australian groups, in an attempt to maximise diversity in each. In Year 2, we are running ten groups with a total of 172 participants. In addition to balancing groups according to the demographics of participants, they are also composed to capture a range of political ideologies (using a Worldview questionnaire, Tetlock, 2005) and cognitive styles (using a Styles of Reasoning questionnaire, also developed by Tetlock, 2005). The introduction of an online tool has also broadened the geographic scope of participants. In Year 2, three groups will be run as homogenous control groups (same age, education, geographic location) to test against the performance of the diverse groups.

Procedure

Each month, IARPA releases a list of questions about current global affairs—for example, Will the Taliban begin official in-person negotiations with either the US or Afghan government by 1 April 2012? Questions are phrased so they will resolve, one way or another, within a given time frame, usually anywhere from a few

months to a couple of years. This allows for empirical testing of forecasting performance against actual events within a reasonable time period.

Each group answers 3–5 questions each month. In Year 1, the US groups answered a selection of IARPA questions, while the Australian groups answered non-IARPA questions matched to the IARPA questions in terms of time, structure and context. In Year 2, all Delphi groups are receiving questions from the same pool. Participants typically spend about 2-3 hours per month on their forecasts, and are not required to answer every question.

All questions assessed by the Delphi groups in Year 1 were binary, phrased to resolve as 'yes' (1), if the event in question occurred, or 'no' (0), if the event did not occur. IARPA also releases conditional and ordinal questions, some of which we are trialling in Year 2, though for the purposes of this paper we will focus on the standard binary questions.

The question format

For each question, we elicit interval probabilities of the event occurring, using a structured 3-step question format:

Example Question: Will Australia formally transfer Uranium to India by 1 June 2013?

(i)	When you think of reasons that make this <u>likely</u> to happen, how sure do you feel that Australia will formally transfer Uranium to India by this date?
	The highest probability of this occurring is: _____
	(Please answer with a percentage 0-100%)
(ii)	When you think of reasons that make this <u>unlikely</u> to happen, how sure do you feel that Australia will formally transfer Uranium to India by this date?
	The lowest probability of this occurring is: _____
	(Please answer with a percentage 0-100%)
(iii)	Finally, consider the balance of evidence. If you had to put a single figure on your opinion of this outcome, what would it be?
	My best guess is: _____
	(Again, please answer with a percentage 0-100%)

Participants are informed that their upper and lower bounds need not be symmetric around their best guess. They are also informed about the trade-off between accuracy and informativeness (Yaniv & Foster, 1997). Narrow intervals give us more information, but require greater skill in order to capture the truth.

The elicitation procedure

Our elicitation framework involves two rounds of judgments, separated by feedback and discussion (
Figure *1*). It differs from the original Delphi in that we are *not* seeking a group consensus. Instead, we simply average across individual answers from Round 2. Also, the original Delphi procedure maintained complete anonymity of group members. In our method, people openly share information and may disclose their background, education and other identifying details if they so choose. The Year 2 software allows greater anonymity, as participants can choose an unrecognisable username (
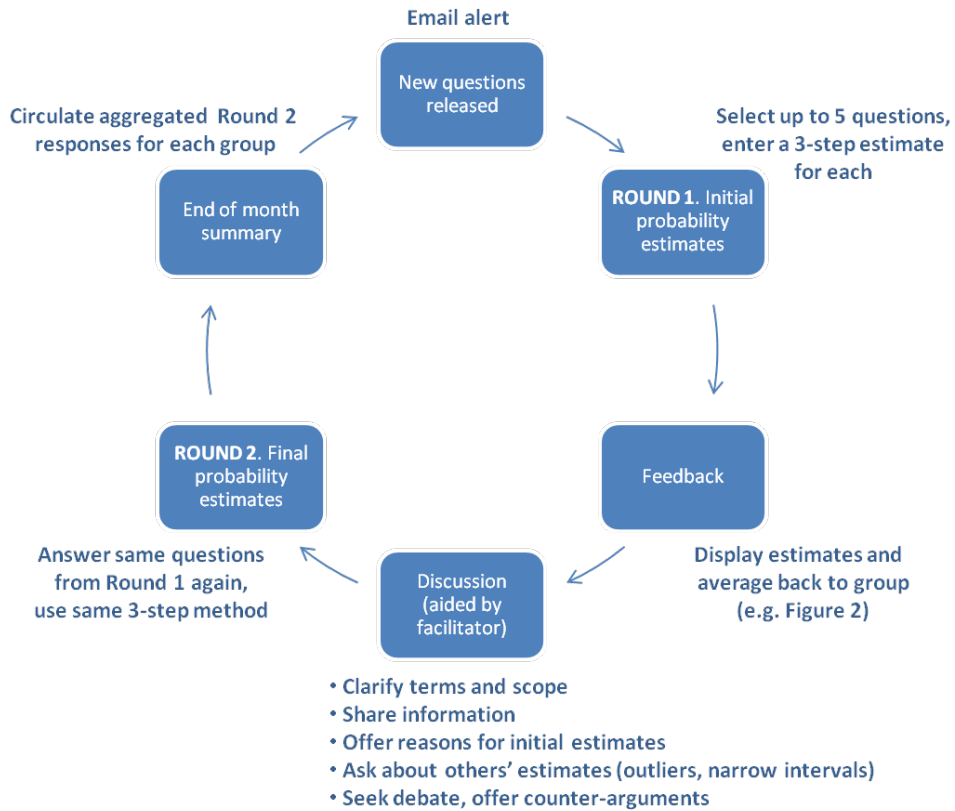Figure *2*).

Figure 1. Flowchart of the monthly question process

Although the basic phases of the approach were the same for Years 1 and 2, the interface for elicitation and discussion was different. In Year 1, the process was managed by a group facilitator, and discussion occurred over a communal email list. In Year 2, participants enter their forecasts into web-based software developed for the project. At the same website, they share and organise information and have online discussions (Figure 2).

The website automatically searches for links related to each question using search terms entered and managed by the facilitator based on question detail and resolution information—what constitutes 'an outcome'—provided by IARPA. Participants also add their own links, and rate and comment on the usefulness of all links.
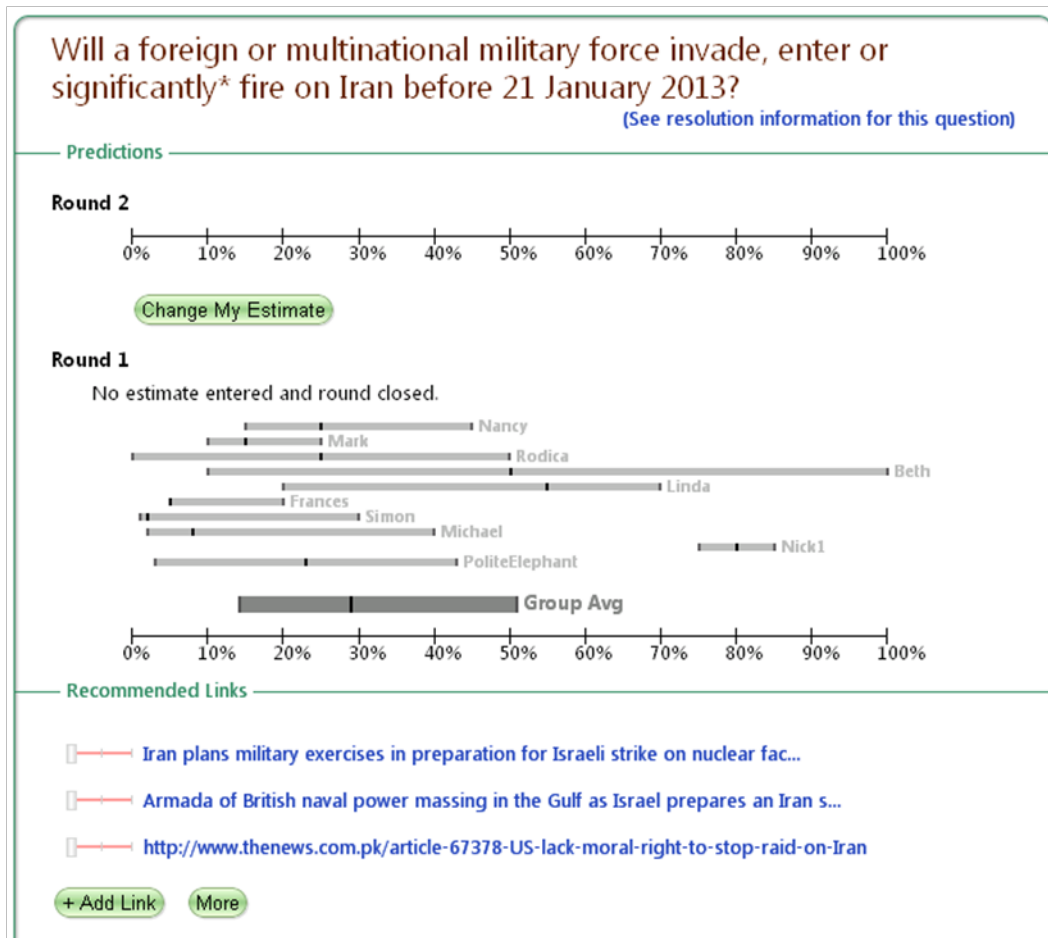
Figure 2. Interval probability judgments from Round 1 are displayed back to each group, links to useful websites are recommended, and discussion takes place below it (not shown).

Measuring performance

Performance is evaluated by comparing the time-matched Brier Scores (Brier, 1950) of the Delphi groups with those of the baseline "ULinOP" (unweighted linear opinion pool of analyst estimates) and with the DAGGRE prediction market, operated by our team in the US.

A Brier Score is a measure of the long term accuracy of estimated probabilities that a given event will occur. Specifically, it measures the mean squared error of a set of probability judgments relative to the observed outcomes, and ranges from 0 (best) to 2 (worst).

For a set of N forecasting instances, each with C possible outcomes (in the case of binary events, $C = 2$), the Brier Score (BS) is calculated as:

$$BS = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} \left(f_{ij} - x_{ij} \right)^2$$

where f_{ij} is the probability that was forecast for outcome j for event i, and x_{ij} is 1 if outcome j eventuates and 0 otherwise.

Predictions in the ACE competition were submitted daily by each competing team over the period for which a question was unresolved. Average Brier Scores for the daily predictions were used to assess performance on individual questions. As the Delphi forecasts are given at a snapshot in time, performance relative to the

ULinOP and DAGGRE market was assessed as the mean daily Brier scores over the 10 day period ending at the close of Round 2 for the month in which a particular question was assessed (Figure 3).

Overall competitor performance on all questions was assessed as the mean of the individual question Brier Scores, and as the Percentage Difference of Mean (PDM) Brier Scores relative to the ULinOP. In addition, to assess the progression in relative performance, the cumulative differences in Brier Scores were also calculated:

$$\text{CumDiff}(t) = \frac{1}{T} \sum_{q=1}^{Q_T} \left(BS_q^{Competitor} - BS_q^{ULinOp} \right)$$

where $q = 1, \dots Q_T$ is the set of questions answered by time t, and BS_q is the mean Brier Score for the set of predictions made for question q.
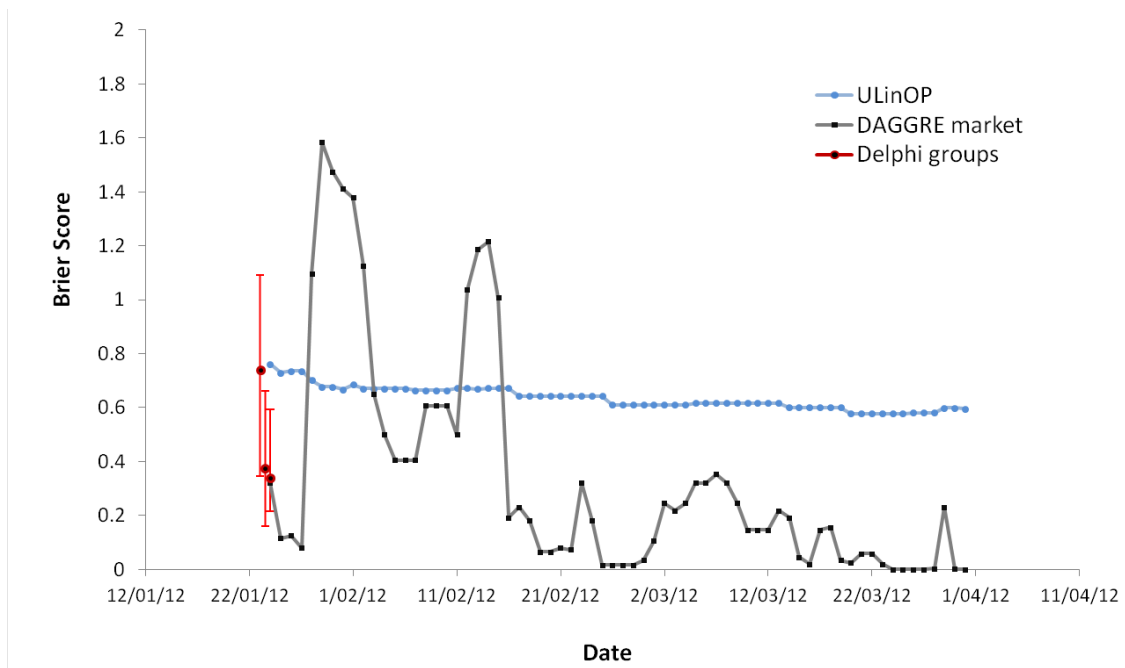

Figure 3. 'Instantaneous' US Delphi group estimates reflect a snapshot in time, compared to the continuous DAGGRE prediction market and ULinOP estimates for a single question.

RESULTS

Delphi groups performed well, evaluated against DAGGRE prediction market estimates and the linear opinion pool (ULinOP) (*Figure 3* & 4). At the April 2011 close for Year 1, the US Delphi groups had outperformed the ULinOP Brier scores by 10%. The Australian Delphi groups outperformed the ULinOP Brier scores by 51%. The cumulative Brier scores for the Australian Delphi groups were initially outperforming the DAGGRE market, but dropped behind it in March.
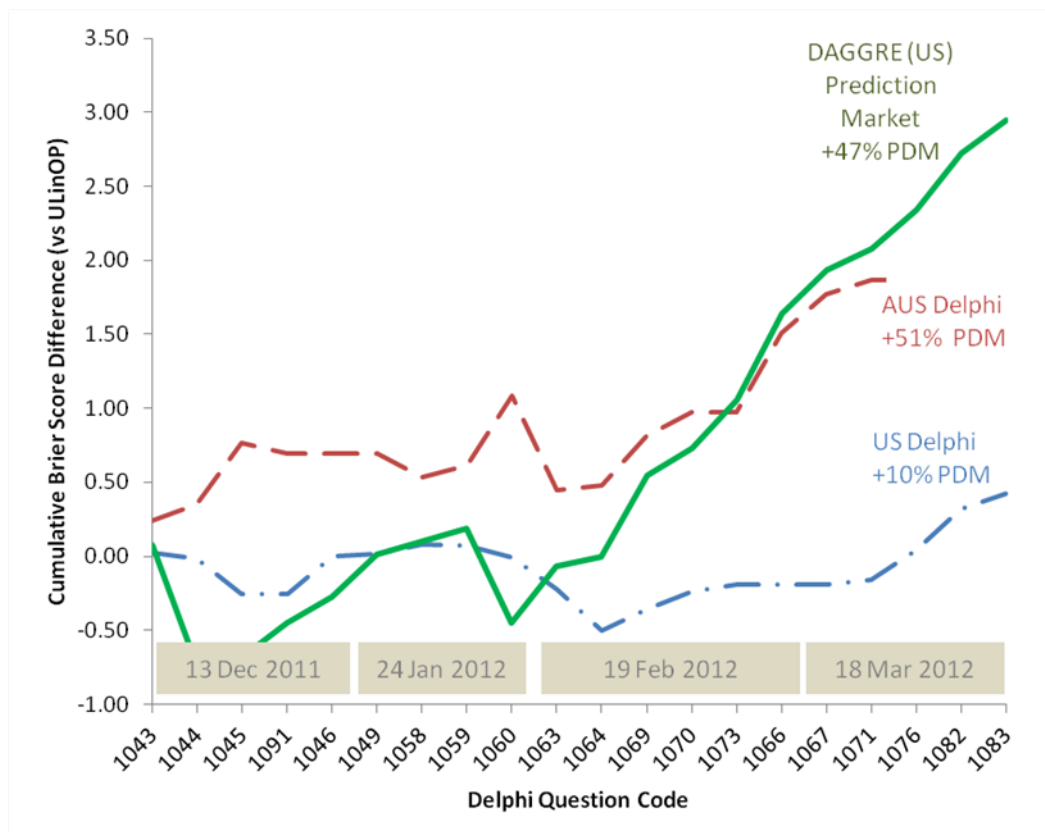

Figure 4. Performance of Delphi groups and DAGGRE prediction market (against ULinOP Brier scores), by question assessment date. Measured as both cumulative Brier Score difference (lines) and Percentage Difference of Mean (PDM) Brier Scores relative to the ULinOP (higher difference means better performance). The early termination of the AUS line results from fewer questions resolving for those groups in the March (final) rounds.

Average responses among Australian groups and among those in the US were quite similar, and hence similarly accurate (Figure 5), despite very different patterns of individual responses within groups (Figure 6).
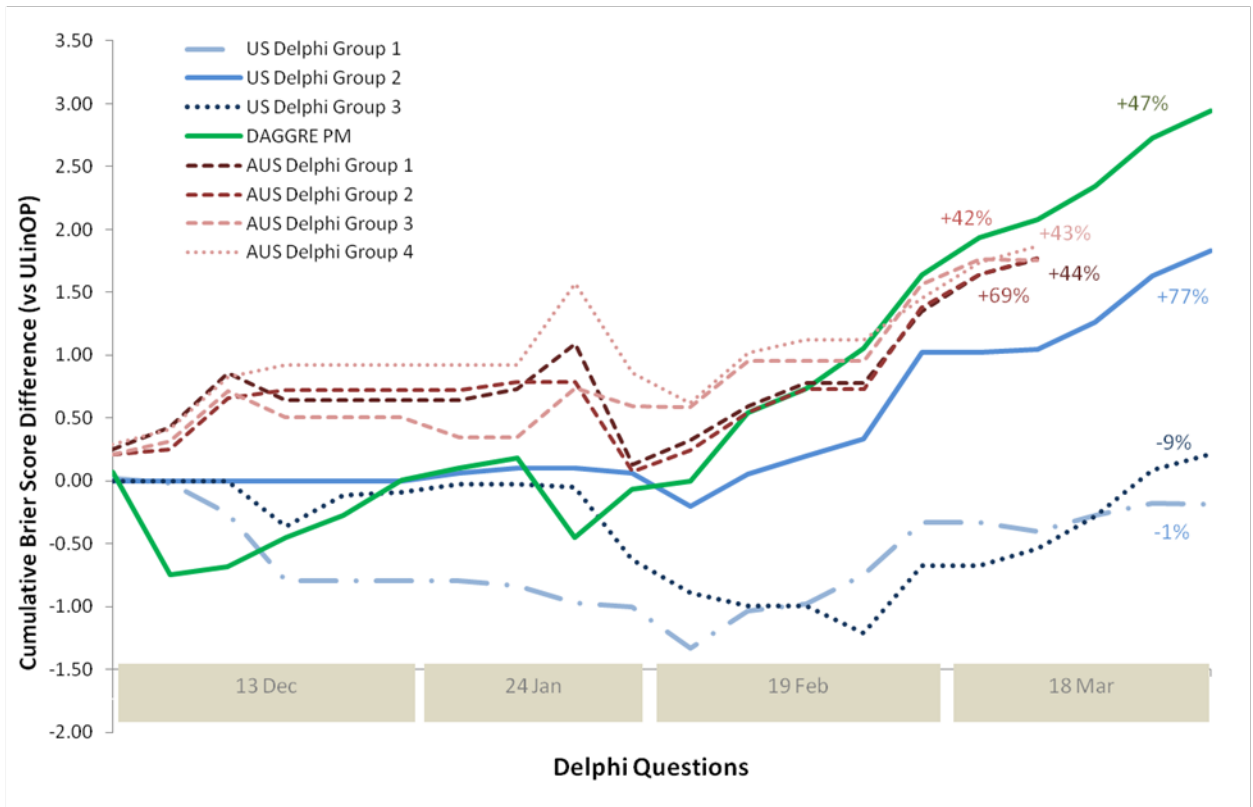
Figure 5. Performance of individual Delphi groups and prediction market (against ULinOP Brier scores). Measured as cumulative Brier Score difference (lines) and Percentage Difference of Mean (PDM) Brier Scores. Higher difference means better performance. The early termination of the AUS line results from fewer questions resolving for those groups in the March (final) rounds.
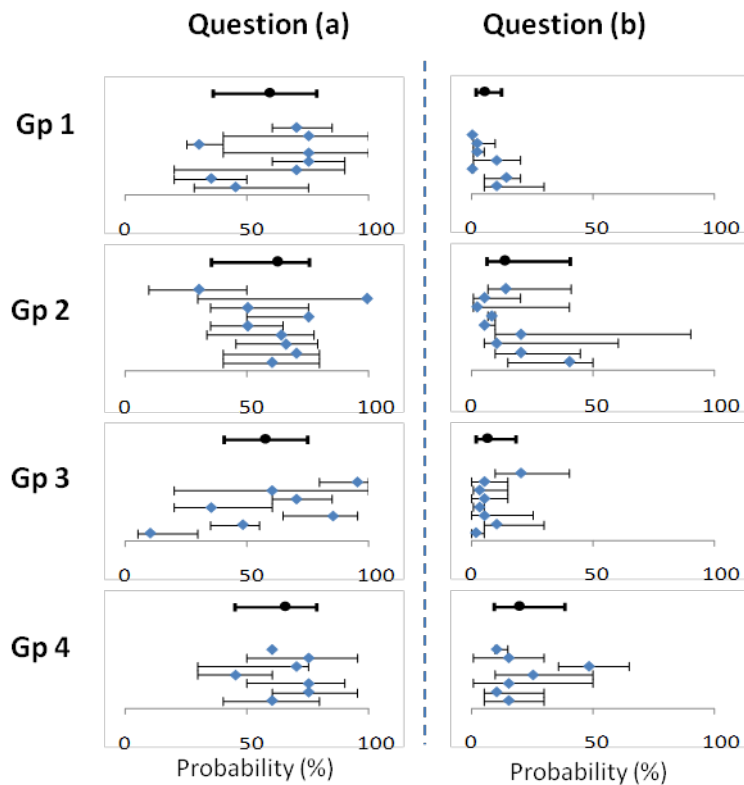

Figure 6. Patterns of individual responses within groups are quite different, despite very similar group averages (bold) between groups.
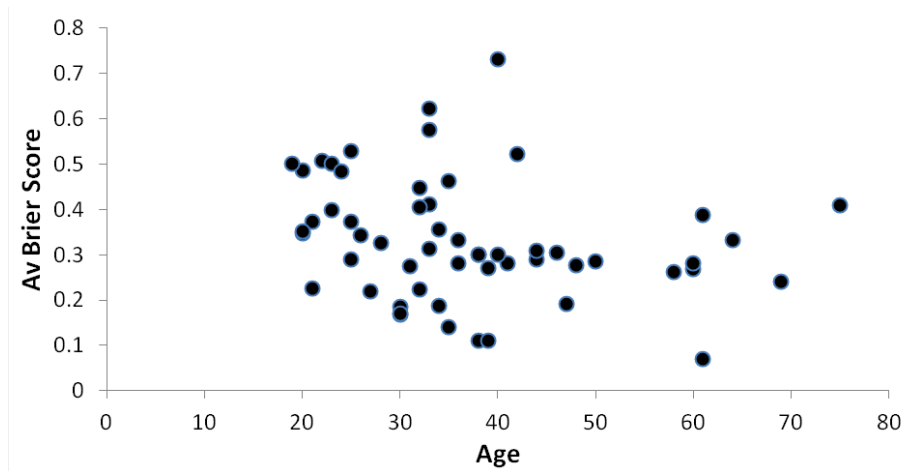
Figure 7. No relationship was detected between age of participants and their average Brier scores

Delphi group results from Year 1 indicate that individual performance does not correlate with cognitive style, age, experience or any other demographic factor that we recorded (e.g. Figure 7).

For those who revised their judgments, second round estimates outperformed first round estimates by 23% (Figure 8).
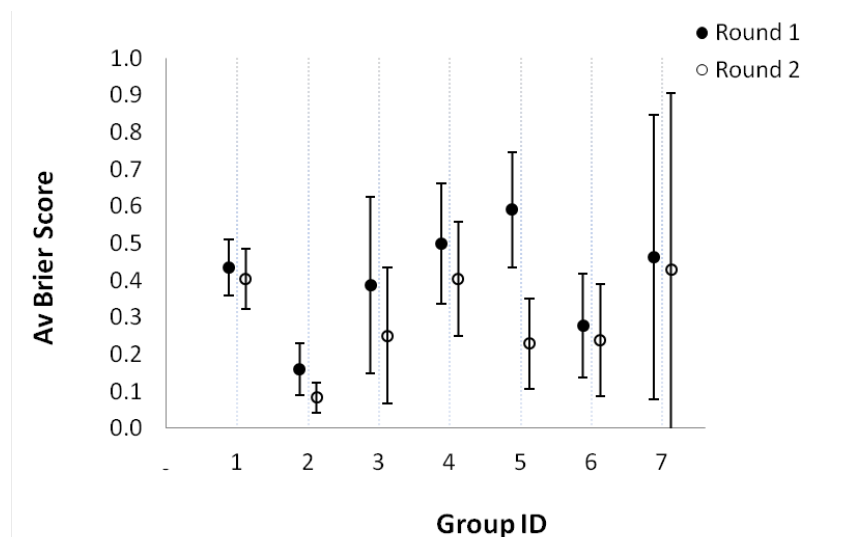

Figure 8. For those who revised their judgments, Round 2 forecasts improved on Round 1 forecasts (95% CIs, higher Brier Scores denote greater error).

DISCUSSION

Group interaction

Despite a large literature about groups as a decision making instrument, the results are inconclusive on whether discussion is good or bad for formulating group judgments (Wright & Rowe, 2011). The popular view since the 80s has been that mathematical aggregation of individual estimates in a group is preferable to behavioural aggregation (Armstrong, 2001; Clemen & Winkler, 1999; Hastie, 1986), because social interaction during consensus-seeking exposes the group to biases that erode the quality of the overall judgement, such as group think (Janis, 1982) and information cascades (Sunstein, 2011). A recent high profile paper (Lorenz, Rauhut, Schweitzer, & Helbing, 2011) suggests that even minimal social influence can undermine the wisdom of crowd effect. Yet, discussion also offers the potential to improve group performance. It can resolve misunderstanding of the question, and provides an opportunity for people to introduce new information and learn from each other. Our results indicate that discussion improves forecasting performance of group averages, given that second round estimates outperformed the first. Discussions may improve group performance by drawing out hidden

information (Mojzisch & Schulz-Hardt, 2010; Stasser & Titus, 2003), encouraging critical thinking (Postmes, Spears, & Cihangir, 2001) and counter-factual reasoning (Galinsky & Kray, 2003), and displaying and resolving differential motivational contingencies (Önkal, et al., 2011).

It would also be interesting to test whether second round forecasts improve beyond that which would be expected from averaging two judgments from a single individual, without interaction (4% points in Herzog & Hertwig, 2009). The improvement with discussion in our study certainly appears to be larger, at least for participants who revised their Round 2 judgments (23%, Figure 8).

Diversity

Previous studies, together with results from this research, have found that single variables (e.g. age or gender) don't tend to correlate with forecasting performance of individuals, except for a moderate correlation between cognitive style and calibration (Tetlock, 2005). The best predictor of good forecasting may be performance on previous questions (e.g. Cooke, 1991). In other words, we have no reliable way to distinguish—before the fact—good judges from bad ones. However, diversity *across* these variables may lead to better forecasting performance of groups (Page, 2008), on the one hand because biases in different directions cancel each other out in the group average, but also because individual members of a diverse group bring different perspectives and information, and ignite interesting debates.

Although we didn't empirically test diversity in Year 1, the US Delphi groups were less demographically and socially diverse than the Australian groups, which may have been a factor in the higher performance of the Australian groups. By comparing diverse groups with homogenous control groups for both countries in Year 2, we will directly test whether the composition of groups affects forecast accuracy.

Theoretically, it has been shown that the benefit of groups is greatest where the overlap between the knowledge bases of individual members is least (i.e. members possess independent knowledge) (Clemen & Winkler, 1985). Selecting for member diversity using information on demographics, experience, worldview and cognitive reasoning style may be one way to reduce dependency between members. Our results showed different patterns of individual responses within groups (Figure 6), yet averages were still very consistent between groups, even when judgments diverged from the actual outcome. It could be that groups accessed similar information, resulting in correlated judgments. We will explore the effect of correlated information sources in Year 2 by tracking information use in the dedicated search software for each group.

Improving Intelligence

Results suggest that our Delphi-style groups have the capacity to considerably outperform the simple aggregations (e.g. the ULinOP), offering an alternative framework for predicting the sorts of events that interest governments and policy makers. A noticeable feature with our Year 1 design was that final forecasts were generally submitted early in the time period allotted for prediction. While early predictions have value, the lack of later predictions prevented the assessment of performance at dates closer to the resolution point. Generally, forecasts become easier as resolution approaches, reflected in the high performance of the continuous prediction market just prior to the question resolving (Figure 3). It is not clear why the continuous ULinOP forecasts did not perform similarly well at this time, although it is possible that their judgments were not continually updated, similar to the Delphi groups. To address this, Year 2 is introducing 'continuous' questions which remain open for an extended time. These may be compared more directly to the performance of the ULinOP and prediction market.

CONCLUSION

The elicitation format and Delphi-style group procedure that we have developed for the Intelligence Game are performing well, compared with methods tested by other teams in the forecasting tournament, and compared with the baseline. Our results also add to growing evidence that group interaction under the right conditions improves forecasts, and group diversity may be an important factor determining the forecasting performance of aggregated predictions. Forthcoming results from Year 2 will provide more insights into the role of group diversity and test the robustness of the 3-step method for eliciting conditional probabilities and ordinal judgments.

DISCLAIMER

REFERENCES

Ang, S., & O'Connor, M. (1991). The effect of group interaction strategies on performance in time series extrapolation. *International Journal of Forecasting, 7*, 141-149.

Armstrong, J. S. (2001). *Principles of forecasting: A handbook for researchers and practitioners*. Boston ; Dordrecht, The Netherlands: Kluwer Academic.

Bonner, B. L., & Baumann, M. R. (2012). Leveraging member expertise to improve knowledge transfer and demonstrability in groups. *Journal of Personality & Social Psychology, 102*(2), 337-350. doi: 10.1037/a0025566

Bonner, B. L., Sillito, S. D., & Baumann, M. R. (2007). Collective estimation: Accuracy, expertise, and extroversion as sources of intra-group influence. *Organizational Behavior & Human Decision Processes, 103*, 121-133. doi: 10.1016/j.obhdp.2006.05.001

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

Clemen, R. T., & Winkler, R. L. (1985). Limits for the precision and value of information from dependent sources. *Operations Research, 33*(2), 427-442.

Clemen, R. T., & Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis, 19*(2), 187-203.

Cooke, R. M. (1991). *Experts in Uncertainty: Opinion and Subjective Probability in Science*. New York: Oxford University Press.

Galinsky, A., D., & Kray, L., J. (2003). From thinking about what might have been to sharing what we know: The effects of counterfactual mind-sets on information sharing in groups. *Journal of Experimental Social Psychology, 40*(5), 606-618. doi: 10.1016/j.jesp.2003.11.005

Hackman, J. R. (2011). *Collaborative Intelligence : Using Teams To Solve Hard Problems*. San Francisco: Berrett-Koehler Publishers.

Hastie, R. (1986). Experimental evidence on group accuracy. In B. Grofman & G. Owen (Eds.), *Decision Research (Vol. 2)*. Greenwich, CT: JAI Press.

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: improving individual judgments with dialectical bootstrapping. *Psychological Science, 20*(2), 231-237.

Janis, I. L. (1982). *Groupthink: Psychological Studies of Policy Decisions and Fiascoes* (2nd ed.). Boston: Houghton Mifflin.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. New York: Cambridge University Press.

Kerr, N., L., & Tindale, R. S. (2011). Group-based forecasting?: A social psychological analysis. *International Journal of Forecasting, 27*, 14-40. doi: 10.1016/j.ijforecast.2010.02.001

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology-Human Learning and Memory, 6*(2), 107-118.

Linstone, H. A., & Turoff, M. (1975). *The Delphi Method: Techniques and Applications*. Reading, Mass.: Addison-Wesley Pub. Co.

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*(22), 9020-9025.

Mojzisch, A., & Schulz-Hardt, S. (2010). Knowing others' preferences degrades the quality of group decisions. *Journal of Personality and Social Psychology, 98*(5), 794-808.

Önkal, D., Lawrence, M., & Sayım, K. Z. (2011). Influence of differentiated roles on group forecasting accuracy. *International Journal of Forecasting, 27*, 50-68. doi: 10.1016/j.ijforecast.2010.03.001

Page, S. E. (2008). *The Difference: How The Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton: Princeton University Press.

Postmes, T., Spears, R., & Cihangir, S. (2001). Quality of decision making and group norms. *Journal of Personality & Social Psychology, 80*(6), 918-930. doi: 10.1037//0022-3514.80.6.918

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. doi: 10.1016/j.obhdp.2011.12.006

Sniezek, J. A. (1989). An examination of group process in judmental forecasting. *International Journal of Forecasting, 5*(2), 171-178.

Sniezek, J. A. (1992). Groups under Uncertainty: An examination of confidence in group decision making. *Organizational Behavior & Human Decision Processes, 52*(1), 124-155.

Soll, J. B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning Memory and Cognition, 30*(2), 299-314. doi: 10.1037/0278-7393.30.2.299

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*(3), 512-523.

Stasser, G., & Titus, W. (2003). Hidden Profiles: A brief history. *Psychological Inquiry, 14*(3/4), 304-313.

Sunstein, C. (2011). Deliberating Groups vs. Prediction Markets (or Hayek's Challenge to Habermas). In A. Goldman & D. Whitcomb (Eds.), *Social Epistemology: Essential Readings*. Oxford: Oxford University Press.

Surowiecki, J. (2005). *The Wisdom of Crowds: Why the Many are Smarter than the Few*. London: Abacus.

Tetlock, P. E. (2005). *Expert Political Judgment : How good is it? How can we know?* Princeton: Princeton University Press.

Tetlock, P. E., & Mellers, B. A. (2011). Intelligent management of Intelligence agencies: Beyond accountability ping-pong. *American Psychologist, 66*(6), 542-554.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*(1), 25-29. doi: 10.1037/h0071663

Wintle, B. C., Fidler, F., Vesk, P., & Moore, J. (2012). Improving visual estimation in the field through active feedback. *Methods in Ecology and Evolution*. doi: 10.1111/j.2041-210x.2012.00254.x

Wright, G., & Rowe, G. (2011). Group-based judgmental forecasting: An integration of extant knowledge and the development of priorities for a new research agenda. *International Journal of Forecasting, 27*(1), 1-13.

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