Group-based judgmental forecasting: an integration of extant knowledge and the development of priorities for a new research agenda

George Wright

Durham Business School

Mill Hill Lane

Durham

DH1 3LB

UK

e-mail: george.wright@durham.ac.uk

Gene Rowe

Institute of Food Research

Norwich Research Park

Colney

Norwich

NR4 7UA

UK

e-mail: gene.rowe@bbsrc.ac.uk

Abstract

We review and integrate extant knowledge on group-based forecasting, paying special attention to the papers included in this special issue of the International Journal of Forecasting. We focus on the relative merits of different methods for aggregating individual forecasts, the advantages of heterogeneity in group membership, the impact of others' opinions on group members, and the importance of perceptions of trust. We conclude that opinion change after group-based deliberation is most likely to be *appropriate* where group membership is heterogeneous, minority opinion is protected from pressure to conform, information exchange between group members has been facilitated, and the recipient of advice is able – by reasoning processes – to evaluate the reasoning justifying proffered advice. Proffered advice is *least* likely to be accepted where the advisor is not trusted – indicated by having different perceived values to the recipient of the advice and being thought to be self-interested. In contrast, the outcome of a group-based deliberation is *most* likely to be accepted when there is perceived procedural fairness and the participants in the process are perceived as trustworthy. Finally, we broaden our discussion of group-based forecasting to include consideration of other group-based methodologies aimed at enhancing judgment and decision making. In particular, we discuss the relevance of research on small-group decision making, the nature and quality of advice, group-based scenario planning, and public engagement processes. From this analysis, we conclude that, in medium- to long-term judgemental forecasting, a variety of non-outcome criteria need to be considered in the evaluation of alternative group-based methods.

Group-based judgmental forecasting: an analysis of extant knowledge and the development of priorities for a new research agenda

Group-based forecasting can be achieved in many ways. The simplest is to average individual opinions and take the achieved figure as the forecast. Alternatively, individuals can meet to discuss the issue in groups – either *with* or *without* some formal structure to the process. Unstructured group processes are historically typical, and seen as a benchmark for judging the performance of structured methods. Research has demonstrated how various social factors can undermine good forecasting (and decision making in general) in unstructured groups, and hence identified the need to somehow control human interaction to pre-empt or ameliorate these factors. Within one such structured-interaction technique - the Nominal Group Technique (NGT) - individuals first make personal forecasts, then the group members meet to discuss the forecast problem, and finally the individuals are given the opportunity to revise their earlier forecasts (Van de Ven & Delbecq, 1974). The average of these revised opinions can be taken as the group forecast. The structure of the Delphi technique is similar to that of the NGT – with the exception that the group members exchange their initial forecasts anonymously with other group members and are given the opportunity to revise their individual forecasts over several Delphi rounds, with the final round average taken as the group forecast (Rowe, Wright, & Bolger, 1991). Other group–based methods include role-playing – in which individuals simulate the interaction of groups in conflict situations in order to understand/predict how the conflict is likely to resolve itself (Green, 2002). Recently, prediction markets have

been studied to see if the evolving monetary-derived predictions - produced by self-interested participants - are accurate (Berg, Nelson, & Rietz, 2008).

The structured methods vary in terms of the *amount* and *type* of information that is exchanged between group members and the *process* by which the information exchange is managed. The interaction of these aspects ultimately impacts upon the degree to which the "advice" of others is integrated with the group members' own opinions. Another important feature affecting the quality of forecasts produced by the different methods is the extent to which group membership is relatively homogeneous or heterogeneous. Finally, some group-based methods are more *acceptable* to group members than others – a practical aspect that is often overlooked in choosing which forecasting method to use in any particular situation.

In this article, we integrate the contributions of the authors in this special issue of *International Journal of Forecasting* toward addressing these key themes in group-based forecasting research, before summarising some general issues for future research to consider. Finally, we contextualise the current forecasting research with respect to other research areas of relevance. We first consider the issue of aggregating individual forecasts.

Aggregating individual forecasts

In the first article, Kerr & Tindale (2011) focus on how to aggregate individual opinions to achieve an accurate group-based judgment. They distinguish between judgment in intellective tasks, in which deduction of the (already-existing) truth is the focus of attention, and judgmental forecasting, in which the forecasters can only

explain and defend their judgments, since the outcome has not yet occurred. This review suggests that pre-existing majority opinions generally determine group consensus decisions in judgmental forecasting tasks, arguing that only in intellective tasks, where those group members who favour the correct answer can explain or demonstrate *why* they are correct, is a correct minority likely to be persuasive. Because of the disproportionate power of majority opinion to determine the subsequent group position in judgmental forecasting tasks, Kerr and Tindale argue that only those aggregation methods that *facilitate information exchange* between group members are likely to be beneficial over-and-above a statistical averaging of prior opinions. Such information exchange in both face-to-face groups and in structured group interaction – such as in Delphi and nominal groups – provides the *enabling conditions* for group members to recognise errors in justifications of judgments. Delphi has the advantage of attenuating any social anxiety about publically identifying other group members' errors. However, face-to-face groups may enhance individual task motivation if individual effort and accuracy is identifiable. Both face-to-face groups and Delphi can, under the right circumstances, provide perceived procedural fairness to all group members who participate, whilst Delphi can protect a minority from group pressure to conform to the majority viewpoint to preserve group harmony – since, within unstructured groups, social goals, such as maintaining group harmony, may conflict with generating the best possible judgmental forecast.

In the next section, we focus on the usefulness of having heterogeneous group membership. As we shall see, hetrogeneity can be enhanced by role-playing.

The advantages of heterogeneity in group membership

The article by Yaniv (2011) reports an empirical study on *susceptibility to framing effects* as a measure of judgment quality. Yaniv labelled as homogeneous those groups made up of individuals that had been assigned to the same framing manipulation within a version of the Kahneman and Tversky's classic Asian Disease problem (Kahneman & Tversky, 1984). By contrast, the heterogeneous groups were made up of individuals with a mix of prior frames for a formally identical decision problem. Yaniv demonstrated that subsequent discussion within the two types of group led to either (i) intensification of framing bias in the homogeneous groupings and (ii) attenuation of framing bias in the heterogeneous groupings. He argues that the invalid consensus in homogeneous groupings may serve to increase group-based confidence but decrease judgmental accuracy. From this perspective, assignment of different role-playing responsibilities to individuals can artificially create useful heterogeneity and thus attenuate conformity to spurious majority opinion.

Onkal, Lawrence, & Sayim (2011) follow up such a suggestion in their own study of "modified consensus groups" – where individual group members role-play functional specialists in marketing, production, and forecasting. In their study, these specialist role-holders provided independent individual forecasts that were then statistically averaged. Participants were instructed to act out their assigned roles as they believed these would be realised in real-life. The experimental materials were sets of time-series data for product sales. A pure model-based statistical forecast was also provided to the group for discussion and, finally, the role-players were asked to accept or adjust this model-based prediction to reach a finally-agreed group-based forecast. Onkal, Lawrence, and Sayim reached the overall conclusion that this group-based

process tended to improve on the accuracy of both (i) the statistical average of the individual-member forecasts and (ii) the pure model-based forecast.

Green & Armstrong's (2011) article studies the worth of inviting individuals to "stand in the other person's shoes". Would this invective give those individuals useful insights into the quality of their initial intuitive judgments? Green and Armstrong's focus was on forecasting the outcomes of conflict situations - with participants instructed to indicate "which decision you think that each party in the situation would prefer to be made and how likely is it that each party's decision will actually occur". But such invective to engage in "role-thinking" proved no more accurate than guessing the outcomes of the conflict situations – even for a range of non-undergraduate ('expert') respondents. In contrast, when students were required to become more engaged with the conflict situations - by "role-playing" simulated interactions between participants in the conflicts – the predictive accuracy of the post-role participants' judgements reached 90%.

All three of these studies therefore speak in one way or another to the need for heterogeneity in groups to aid forecasting – and of the value of using role-playing to create artificial heterogeneity when this does not exist.

In the next section, we turn our focus to the impact of others' opinions (i.e., advice) on opinion revision. Most of the research has been conducted within the Judge-Advisor System (hereafter, JAS) paradigm (Sniezek, 1992) which simulates *Advisors* who give information to others acting as *Judges* (who are ultimately accountable for a

judgment or forecast). Hence, the decision to use or discard the opinions of others is that of the Judge.

The impact of others' opinions

Soll & Mannes (2011) note the well-documented finding that judges often overweight their own opinions relative to the advice of another (when the single advisor offers simple numerical advice) – so called *egocentric discounting*. In such instances, the averaging of one's own opinion and that of the advisor would often have led to greater accuracy. However, as Soll and Mannes note, when the experimental task instead focuses on combining the opinions of others, the relative weights attached in such combination tasks tend to be less biased. One explanation of the overweighting of one's own opinion is that the reasons for one's own opinions are, perhaps, richer and more salient than those of the advisor – where the reasons for proffered advice are seldom part of the experimental paradigm. For example, in the few JAS studies where non-numerical justification of numerical advice has been made available to participants, such justification of advice has been impoverished and superficial. In Van Swol & Sniezek (2001), advisors gave their recommendations as to which answers to multiple-choice items were correct. The advisors were free to elaborate on their recommendations by writing comments that would be seen by the judge. In the minority of instances where advisors elaborated upon their recommendations, such elaborations amounted to little more than comments such as 'I'm definite about this', or 'This is a guess' (p.297). In their own empirical study, Soll and Mannes manipulated both the task of revising one's own opinion and the task of combining the opinions of others - within a single experiment. As these authors note, statistical averaging performs best when both the probability of detecting someone with true

expertise is low, where differences in expertise are also low, and where errors in judgments are randomly distributed. Further, in many (laboratory and real) situations experts are not easily identified and so the advantage of averaging advice - over picking a single advisor - is increased. In their empirical study, Soll and Mannes operationalized "advice" by presenting numerical values on four cues that were purportedly used by an advisor in making judgments of a focal numerical variable. However, note that Soll and Mannes did not utilize verbal justifications or rationales underpinning these predictions. The authors found that participants approached the dual tasks of revision of their own opinions and the combination of the opinions of others in different ways. Their analysis revealed that the obtained egocentric discounting in opinion revision *was not* caused by respondents giving more credibility to their own opinions than the opinions of others. Additionally, respondents did not rate themselves as more accurate in their judgments than their advisors.

Van Swol (2011) focuses on the distinction between intellective and judgmental tasks, discussed earlier, and argues that people are more likely to accept advice on intellective tasks where the true answer is already known. This is because, she argues, people seek out accuracy in answering intellective problems and an advisor may be able to demonstrate the correctness of their advice. By contrast, in judgmental tasks – where the true outcome or answer is not known at the time of the judgment - the degree of trust placed in the advisor is more important for acceptance of proffered advice. If the advisor is perceived as sharing similar values to oneself, then trust increases - and so does acceptance of advice. By contrast, in intellective tasks, high advisor confidence tends to lead to increasing trust in the proffered advice. Van Swol argues that, in judgmental forecasting tasks, advisors are likely to differ in their

advice and so a potentially useful cue that may be used to differentiate advisors is the degree to which an advisor shares the values of the client. In her empirical study, she found that advisors spontaneously provided much additional information in a task where the advice was focussed on which movie the client would find enjoyable. She argues that this unstructured material would likely help establish common values and thus increase trust in the proffered advice.

Jodlbauer & Jonas (2011) investigate the influence of perceptions of an advisor's self-interested intentions on the client's (i.e., advisee's) own intention to accept proffered advice. As these authors note, the most-studied characteristics of the advisor have been perceived expertise, reputation, and confidence associated with proffered advice. But clients, these authors argue, know that the advisor has an advantage in knowledge but are also aware that advisors may be self-interested and so untrustworthy. Using a simple experimental manipulation, Jodlbauer and Jonas varied whether an advisor introduced himself as either a representative from a not-for-profit organization or a for-profit organization. The student clients placed less trust in both the ability and the integrity of the for-profit advisor.

In structured group processes, such as Delphi, the degree to which advisors (i.e. other group members) share common values and are to be trusted in their opinions has not been a focus of research. Recall that forecasts and opinions are shared anonymously in Delphi applications. However, recently, issues to do with procedural justice and trust have been raised (and see also the studies above) and it is to this issue that we turn next.

The importance of perceptions of trust

In many real-world settings, there are additional complexities to conducting research – and to answering the fundamental question of how to obtain a 'good' forecast or decision. Two studies in this Special Issue are noteworthy in their reporting of large-scale real-world interventions in organizations with a future-orientated focus, and these exemplify some of the research difficulties, as well as the definitional difficulty of identifying a 'good' forecast. That is, when one cannot easily control the experimental environment and use forecast *accuracy* as a criterion measure – then how can one judge the merits of a group-based process?

Landeta & Barrutia (2011) utilize a version of the Delphi process within a policy development setting that incorporated specific enhancements to the Delphi technique to help achieve a *process* that was of *high-quality* and *acceptable* to the varied members of a professional bureaucracy. In particular, the authors wished to attenuate any potential conflict between nominal group members. Their case study documents a process intervention that (i) maximises the *perceived importance* of participation, (ii) minimises the possibilities of *manipulation of outcomes by powerful individuals*, (iii) *facilitates the exchange of reasoned justifications* for the divergent opinions existing in the professional bureaucracy. In this way, the process variation created *trust* and a sense of *procedural justice* – such that participants were willing to accept the consequences of the Delphi yield. Interestingly, Landeta and Barrutia's method for the selection of interest-group representatives for membership of the Delphi panel also establishes that those representatives share similar values to their electors and thus, by implication, share the trust of their electors to represent the electors' interests.

Klenk & Hickey (2011) present another variation on the Delphi technique to enhance group-based deliberations. Their focus is on a method development to aid integration of group-based knowledge and to map areas of both consensus and dissent – whilst minimizing any negative group dynamics during deliberations of viewpoints. Using "concept mapping" techniques, their case study illustrates a practical tool to document both shared and individual viewpoints in both knowledge and values. Such a tool may be particularly useful, given the importance – previously discussed – of group members being able to provide rationales/justifications of their forecasts in order to aid other group members in assessing the quality of proffered advice. As with Landeta and Barrutia, the merits of this method were largely assessed through the use of post-event questionnaires seeking the opinions of participants about the process. After all, a process that is unacceptable to its participants is unlikely to have much impact – hence *acceptance* would seem an important additional criterion for judging forecast process quality, and acceptance appears to be closely linked to perceived 'trust' in the purveyors of advice.

We next turn to study the adequacy of group-based processes in another "naturalistic" decision setting – that of selecting the best research papers and research applications.

Naturalistic decision making in Academia

Benda & Engles (2011) take an unusual – yet actually highly salient – perspective on group-based forecasting. They focus on the operation of the peer review process in both the selection of academic manuscripts for journal publication and in the selection of grant applications for research funding – arguing that in each case, what editors/reviewers are doing is attempting to forecast the success of a paper or project

(as might be indicated by, for example, a paper's future citations). They argue that inter-referee agreement is *not at all* important for valid selection – again, seeming to indicate the importance of heterogenous group membership. The key, to Benda and Engles, is that each knowledgeable referee should produce a credible review. Low inter-referee agreement can underpin subsequent strong internal validity of the reviewing process because a lack of agreement by knowledgeable referees can act to discourage superficial vote-counting. Dissensus should be resolved by equally knowledgeable journal editors and grant-awarding panels, who should seek to understand *why* disagreements between credible referees exist. Benda and Engles identify and describe tension in valid peer review processes and go on to show that, without such tension, genuinely innovative research papers and proposals may be rejected. Scientific revolutions and paradigm shifts may encounter resistance when research papers or research proposals are evaluated by groups - compared to research that presents additional increments to existing paradigm-based knowledge. Thus both vote counting and averaging of opinions may attenuate the insights provided by an in-the-minority reviewer. As Benda and Engles point out, groups can be less-than-optimal users of information that is not already generally shared amongst the group membership. As a remedy, these authors advocate that individual referees – who are credible and knowledgeable - should have the occasional power to declare the unilateral acceptance of a research-based paper or research proposal.

But, which of the methods of unstructured and structured group-based judgmental forecasting is best? It is to this issue that we next turn.

Comparison of group-based forecasting methods

Graefe & Armstrong (2011) compare the accuracy of unstructured face-to-face groups with three structured methods: (i) nominal groups, (ii) Delphi, and (iii) prediction markets. Their task was the quantitative estimation of ten almanac quantities, such as the percentage of the US population aged over 65 years in 2000. Overall, they found few differences between the four methods, but all of the three structured group interaction methods improved over the group members' individual prior estimates. Importantly, only the Delphi method improved over the statistical averaging of the group members' prior opinions. However, as Graefe and Armstrong note, their "intellective" estimation task was impoverished in that participants would, likely, be unable to share information that would be of use to aid improvements in the judgments of other group members. Interestingly, the study's participants expressed an attitudinal *preference* for group processes that involved face-to-face synchronous contact with other group members over participation in either Delphi or prediction markets.

Summary of key findings in the Special Issue papers

From the papers discussed in the preceding paragraphs we suggest that an individual's opinion change after group deliberation is most likely to be *appropriate* where:

1. Group membership is heterogeneous. Artificial heterogeneity can and should be achieved by role-playing rather than role-thinking.
2. Minority opinion is protected from majority pressure to conform – which might best be achieved through anonymity of participants' judgments.
3. Information exchange between group members has been facilitated such that errors in opinions can be recognised as such. The addition of novel approaches

such as concept mapping (Klenk & Hickey, 2011) might help with this explication.

4. The advisee is able – by reasoning processes - to evaluate the reasoning underpinning the proffered advice (again highlighting the possible value of reason-decomposition approaches - such as concept mapping).

Proffered advice is *least* likely to be accepted (whether that advice is appropriate or not) when:

1. The advisor is perceived to have different values to the advisee.
2. The advisor is thought to be self-interested.
3. The advisor is not trusted (which is liable to be related to a degree to having different values and being self-interested).
4. The advisor(s) are in the minority.
5. The advisor(s) are not able to justify recommendations made – such as when advice is given in numerical form only.
6. Advisors express little confidence in their opinions.

In contrast, the outcome of a group-based deliberation is most likely to be *accepted* when:

1. There is perceived procedural fairness to the group-based process.
2. The participants in a group-based process are perceived to be trustworthy – as indicated by a commonality of values with the advisees and a lack of self-interest in the advice proffered.

3. There is a sizable majority favouring the prediction or outcome.

In the next sections, we broaden our discussion of group-based forecasting to evaluate other group-based methodologies to enhance judgment and decision making. We draw out implications for improving group-based forecasting.

Group decision making

Schweiger, Sandberg, & Ragan (1986) discuss approaches to engender debate and evaluation of decisions in management teams. They differentiate (i) dialectical inquiry and (ii) devil's advocacy. Both methods systematically introduce *conflict* and *debate* by using sub-groups that role-play. In dialectical inquiry, the subgroups develop opposing alternatives and then come together to debate their assumptions and recommendations. In devil's advocacy, one subgroup offers a proposal, while the other plays devil's advocate, critically probing all elements and recommendations in the proposal. Both methods encourage groups to generate alternative courses of action and minimise tendencies towards premature agreement or convergence on a single alternative. Both methods also lead to a more *critical evaluation* of assumptions by providing mechanisms for *encouraging dissent* whilst at the same time fostering a high-level of understanding of the final group decision. Nevertheless, these role-played, conflict-enhancing, interventions for improving decision making need to be focussed on factual information because personalities can, inappropriately, become the focus of discussion. Schweiger, Sandberg, & Rechner (1989) compared both techniques to a non-adversarial approach where decisions were simply discussed with the aim of achieving a consensus amongst group members. Questionnaire ratings by group participants found that the two conflict-based approaches were rated higher in

terms of producing better recommendations and better questioning of assumptions. Formalizing and legitimizing conflict can thus enhance perceptions of the quality of the outcome of group decision making. However, whilst conflict can improve perceived decision quality, it may weaken the ability of the group to work together in the future if the role-playing is not sensitively managed. Also, as Nemeth, Brown, & Rogers (2001) document, authentic minority dissent, when correctly managed, is superior to role-playing interventions in stimulating a greater search for information on all sides of an issue. But, generally, the authentic dissenter is disliked even when she/he has been shown to stimulate better thought processes. However, Nemeth & Chiles (1998) showed that the persistent authentic dissenter, while not liked, can be admired and respected. Also, it must be recognised that implementation of decisions rests on securing the subsequent cooperation of involved parties (as highlighted by several of the papers in this special issue – e.g. Landeta & Barrutia, 2011) and so affective personal criticism invoked in the prior critical debate will be dysfunctional.

As we have discussed, in a forecasting context, an understanding of the likely outcomes of conflicts can be invoked by the use of role-playing the interactions of the conflicted parties to a dispute (Green & Armstrong, 2011). However, the focus on critique and debate that are entailed in dialectical inquiry and devil's advocacy has not been a prior topic of research in group-based forecasting. For example, the work of Yaniv (2011) and Onkal *et al* (2011) promotes the inclusion of heterogeneity of opinions rather than the structuring of dissent. We advocate that methods which invoke critique and dissent should now become such a research focus.

The next section of this paper reviews research on the acceptability of advice in contexts other than JAS or Delphi. As we shall see, justifications for proffered advice have been shown to be a crucial component of the advice's subsequent acceptability.

Acceptability of advice

Expert systems capture the reasoning of experts within computer systems that can then act to replicate the expert's decision making. Often such systems are used by less-expert decision makers as an aid to decision making. Arnold, Clark, Collier, Leech, & Sutton (2006) found that novice users of expert systems tend to accept systems' recommendations, while more-expert users have a stronger interest in examining the explanations that the systems generate for particular recommendations. As such, expert users are interested in comparing the recommendations and underpinning reasoning of the systems with their own judgment. In fact, this focus on evaluation and verification may be a precondition for acceptance of systems by more-expert users.

Apart from the use of expert systems, statistical models can be used to automate decision making, or aid decision makers to make decisions. In the USA, a quarter of a million people are admitted unnecessarily to hospital each year with suspected heart failure. Yet, using seven predictive indicators (four based on quantifications of a patient's medical history and three being summary measures of electrocardiogram tests), Corey & Merenstein (1987) developed a quantitative linear regression model that was correct 85% of the time. Because of its predictive success, use of this decision aid was made mandatory for physicians in one major hospital. However, after some time, its use was made a voluntary choice. From then on, the linear model

was used to aid diagnosis of only 3% of patients with suspected heart failure. Why did this very effective decision aid not get consulted more often?

Seick & Arkes (2005), in a study of decision aid neglect, found that decision-makers who had access to the statistical equation underpinning an effective decision aid often didn't bother to examine the workings of the method. The tendency was for the decision-makers to rely on their own judgment and, later, report that they performed better on the prediction task than the advice offered by the decision aid – although the decision aid actually outperformed their intuitive judgments. One participant in the study commented: '… the statistical equation gave me more confidence if it was similar to my original guess. If it was different, I went along with my gut instinct rather than use the equation. If I had absolutely no idea, I went with what the equation gave me'.

In fact, people are much more likely to follow a recommendation that comes from an expert, for example a physician, rather than one that comes from a statistical model. Expert systems that provide the user with explanations of the advice given are more likely to be heeded than the unexplained, although accurate, predictions of linear models. This result bears comparison with the previously discussed research in forecasting (Kerr & Tindale, 2011) that emphasizes that the justifications for proffered advice from other group members needs to be made as explicit as possible, in order to have any influence on opinion change. We develop discussion of this issue in the next section.

Extending study of the reasons underpinning proffered advice

We have had a longstanding concern with the nature of information exchange between participants in nominal groups doing forecasting tasks (see Rowe & Wright, 1996), and in particular, the need to understand the *processes* leading to judgment/forecast *change* in individuals within such groups (and interacting groups more widely) (Rowe, Wright & Bolger, 1991). That is, what factors are responsible for leading participants to accept others' judgments or forecasts and amend their own to some degree? Some of the papers in this special issue have provided further evidence of a need for the information being exchanged to possess certain qualities – including indicating shared values between information provider and recipient. Information that leads a recipient to trust the information provider, or that indicates their expertise (at least in certain tasks – perhaps intellective moreso than judgmental) is also important. However, what is it about the content and type of forecast justifications and challenges that induces participants to change their positions, and thus potentially improve forecasts? We had hoped that this special issue would flush out a number of empirical papers addressing this topic, but that has not been the case - and this research area remains under-explored.

However, some early research spoke to this issue, and, we propose, ought to be revisited by modern researchers. For example, Brockriede & Ehninger (1960) have shown that only a limited number of argument types are, in principle, available to people advocating specific propositions or claims – arguments of *parallel case*, *analogy*, *motivation*, and *authority*:

In *Analogous reasoning*, the reason given makes use of our general knowledge of relationships between two events in dissimilar situations. For example, if someone is trying to estimate the time it will take to drive to a nearby airport, an advisor may reason that, "the airport is roughly the same distance away as the shopping mall. Therefore, the time it will take to get to the airport will be approximately the same as it is to travel to the shopping mall – about 30 minutes".

Parallel case reasoning involves making use of our knowledge of a previous experience of a near identical situation. For example, if someone is trying to estimate the time it will take to drive to a nearby airport an advisor may reason that, "it will take about 30 minutes to drive to the airport because it took me 30 minutes at the same time of day last month".

Authoritative reasoning involves making use of substantive knowledge. For example, "the radio announcer has said that traffic to the airport is heavy today and so I estimate that you should add 20 minutes to your journey time".

Motivational reasoning involves making use of specific insights about people's motivations or desires. For example, "since you will be in a hurry, then I reckon that you can cut five minutes off your usual journey time".

Importantly, whilst reasoning by analogy, parallel case, authority, and motivation are available justifications for advice in judgmental tasks, only justification by authority,

or expertise, is available as a justification for advice in intellective tasks. Thus, many crucial questions remain to be explored and answered. For example, what components of advice-giving cause opinion change in the judgmental forecasting of individual experts? How is advice evaluated and under what conditions will advice be assimilated or discounted? When one expert defers in his or her own opinion to the well-argued opinion, or challenge, of another, is this an indicator of the presence of valid advice that will improve validity in (the revised) judgmental prediction?

In the next section, we focus on another technique that is used by decision makers to anticipate the future – scenario planning. We show that stakeholder analysis – akin to role-thinking (as discussed earlier - see Green & Armstrong, 2011) - is a component of current scenario development practice.

Scenario planning and stakeholder perspectives

The scenario method explores the complex relationship between social, economic, technological, environmental and political factors from multiple perspectives, enables sense making of their interactions, and provides a vehicle for the development of plausible futures that may impact on the focal organization.

The approach entails some consideration of stakeholder values and actions to add realism to already-constructed scenarios. In practice, stakeholder analysis is an optional addition to the 'mix' of ingredients; as 'a tool to be used in parallel with the scenario process, as and when members of the scenario team find it useful' [van der

Heijden, Bradfield, Burt, Cairns, & Wright, 2002, p.219]. Stakeholders include the focal organization's competitors, customers, regulators, etc.

Wright & Goodwin (2009) have argued for a more intense focus on stakeholder analysis within the scenario development process - as the likely actions of stakeholders to enhance and preserve their own interests in a particular unfolding scenario are thought-through. Such *role-thinking* assumes that scenario participants will be able to put themselves in the shoes of each particular stakeholder grouping when this does not involve actual interactions with representatives of such groupings. At the same time, stakeholder interests and values may be more subtle than those that are obvious on the surface. However, as we have discussed, there is, by contrast, evidence that *role-playing* unfamiliar roles can lead to insights. Green (2002) first showed that when university students were asked to role-play the participants in six heterogeneous conflict situations, their subsequent group-based resolutions of the conflict – or the group-based forecasts of the outcomes of these conflicts – were accurate. Intuitively, it would seem that one's own experiences of the past resolution of conflicts – perhaps as recalled or previously experienced, and including both personal and non-personal conflicts – offer a strong guide to the prediction/resolution of the outcomes of novel conflicts. In other words, if the resolutions of conflicts are, generally, the result of the operation of basic human motivations and value systems, then the conditions for reasoning by analogy – and acting by analogy - are favourable (Wright, 2002). Our current analysis now leads us to advocate that scenario development should incorporate role–playing of stakeholders - rather than use less-experiential role-thinking activities.

Further, Cairns, Sliwa, & Wright (in press) advocate the interrogation of the scenario stories from the perspective of the full range of involved and affected actors through application of Flyvbjerg's question framework for phonetic inquiry. They advocate structuring this analysis using a matrix that lists the developed scenarios along the top row of a matrix (*Where are we going?*) and the range of identified actor groups down the left-hand column. Within each box that marks an intersection of a scenario and an actor group, they suggest considering two issues: (i) the impact of the unfolding future on the actor group's interests and values (*Is this development desirable?*) and (ii) the likely action/reaction of the group to the particular unfolding future (*What, if anything, should we do about it?* where 'we' is the particular actor group). Use of role-playing would enable participants to become more sensitised to the plight of each of the groups of actors and become aware of the degree of power of action that each of them has to preserve or enhance their own interests as a particular scenario unfolds.

We next turn to the presence of potential framing effects in scenario development and, following Yaniv (2011), argue for the preservation of heterogeneity in scenario development teams.

Scenario planning and heterogeneity of participants

In most scenario planning exercises, the scenarios are developed by participants from within a single organization. It follows that these participants are likely to have a homogeneous frame on the nature of the future. One way, in practice, used to counter this potential bias is to employ outsiders - so-called 'remarkable people' – who hold minority viewpoints about the future. Such deliberately-invoked diversity is likely to reduce frame blindness in the context of a facilitated process intervention within an

organization. However, at present, the incorporation of such potential insights is unstructured and unevaluated.

Also, in practice, scenario development sometimes involves a scenario team composed of representatives from multiple agencies – i.e., the scenario team is initially formed from a heterogeneous constituency. Cairns, Wright, Bradfield, van der Heijden, & Burt (2006) have argued that the process of scenario planning can provide a non-adversarial common viewpoint to unite, what may be, initially-fragmented groupings. By contrast, in terms of our analysis, the fragmentation should instead be conserved – at least until the point when any action response to the constructed set of scenarios is debated (see Goodwin & Wright, 2010). In the more usual scenario development activity, conducted within a single organization, the conventional process results in the initial development of four skeleton scenarios that are then each fleshed-out by one of four sub-groups. But differences, in world-views, between these sub-groups are likely to be small. On our analysis, once a particular scenario is fully developed it should then be subjected to adversarial critique by one or more of the other subgroups. In this way, also, the systematic introduction of conflict and challenge is likely to enhance the quality of the finally-developed scenarios.

Finally, we now turn to a discussion of public engagement processes. As we will see, evaluation of public engagement methods has direct implications for the evaluation of alternative methods for group-based forecasting – in situations where outcome data, that can be used for forecast verification, is not available.

Public engagement processes

A contemporary domain in which many of the issues in this special issue are currently being played out is that of public engagement in agenda setting and policy making. (Though a caveat is needed here, as copious alternative terms have also been used to describe this general domain – for example, replacing the prefix 'public' with 'stakeholder' or 'citizen' and the suffix 'engagement' with 'communication' or 'consultation' or participation' e.g. see Rowe & Frewer, 2005, for a discussion of definitional nightmares). In essence, public (or stakeholder... etc etc) engagement is the process of involving a wider range of perspectives into some decisional or agenda-setting (or even forecasting) process than would traditionally be the case. The origins of this zeitgeist are difficult to pin down, but it has been associated with a number of major failures in the traditional decide-announce-defend approach to policy making, in which the public and other stakeholders would simply be the recipients of communications about the derived policy (developed by governmental agencies, legislators, etc.).

Associated with this has been a decline in public trust in government, politicians, and scientists, in many democratic societies over the last few decades (e.g. De Marchi & Ravetz, 1999; Laird, 1989). Whereas in the past a compliant public might have accepted what governments claim as best policy, or scientists as 'the truth', nowadays there appears to be more dispute, disbelief and distrust, with the public not behaving as their traditional advisors would recommend. The ideal of engagement is that, by involving the public (or its representatives, or other excluded stakeholders) that, somehow, *trust* will be regained and, also, that decisions may be improved - because of the addition of lay knowledge and perspectives.

The way in which input is gained from this wider constituency is invariably through group-based approaches (Rowe & Frewer, 2005, list over 100 'methods' for doing this). The parallels to the topic of this special issue are thus a) the assumed benefits of having heterogenous input, and b) the importance of a process being perceived as acceptable to its participants for it to have utility (e.g. Webler, 1995). Indeed, the value of public engagement is seen by many as self-evident, which has hindered attempts at asking critical evaluative questions, such as does public engagement work? Are policies derived through this approach actually 'better' than those achieved through traditional approaches? Which of the various methods of engagement work best in which situation? And indeed, what do we mean by 'work'?

It is in this latter point that alternative methods of public engagement are confronted with similar practical evaluative problems to those involving medium- to long-term forecasting – where there is the absence of the possibility of using accuracy to assess forecast quality. More recently, the issue of evaluation has risen up many agendas – particularly as authorities have realised a need to justify their expenditures on expensive engagement processes. And here, possibly, might be some lessons for the forecasting domain. Evaluation research has generally sought to assess two main aspects of engagement processes for quality – the first being *process acceptability* to the participants involved, and the second being the *quality of the process* used in enacting engagement. Various criteria have been forwarded to measure acceptability (see Rowe & Frewer, 2000), such as process *transparency*, *independence* of the facilitation of a process from the event sponsors' potentially vested interests, and the appropriate *representativeness* of participants. Process quality has been assessed

according to criteria such as the availability of *appropriate resources* to complete the task (including time resources), the utilisation of relevant *decision-structuring* processes, and the adequate and full *definition of the task*. It is possible that these criteria (or others from this domain) might be useful in the evaluation of medium- to long-term forecasting processes – using acceptability and process quality measures as surrogates for unobtainable validity measures. Of particular relevance is the consideration of what is an adequately 'representative' set of participants. The findings from several studies in this issue have highlighted the importance of heterogenous group membership – but research in the public engagement domain would prompt forecasters to think further in terms of exactly who those heterogenous members should be, and if they are to be experts in certain domains, would ask the question – what do we mean by an 'expert', and how might we measure and confirm this (c.f., Rowe & Wright, 2001)? All these issues now deserve a place on our future research agenda for group-based judgemental forecasting.

Conclusion

The papers in this issue reveal that group-based forecasting is a complex, multi-faceted issue. One of the distinctions made in several of the papers is between intellective and judgmental (forecasting) tasks – with some evidence put forward that different factors may be more or less relevant in determining the output of groups considering each task type. In group-based forecasting practice, there are likely to be aspects of *both* of these types of tasks, with experts bringing factual knowledge to bear to support their forecast and trying to persuade others of the correctness of their special knowledge (and thereby of their own expertise).

But there is a third significant component to real-world forecasting that is also touched upon here, particularly in the papers of Landeta & Barrutia (2011) and Klenk & Hickey (2011), and that is the policy making angle (after all, the 'Policy Delphi' is a distinctly named variant of one of the main group-based forecasting techniques). We must ever remember that forecasts are rarely, in themselves, disinterested and innocent products of the group process in which they are produced. Forecasts – in non-laboratory settings – often serve a purpose in policy making. If forecasts are for hoped-for outcomes, they may encourage policy making to foster their occurrence; if forecasts are for feared outcomes, then they may encourage the forecasters (and policy makers – who may or may not comprise the same set of individuals) toward actions that might undermine the forecast. And this reality should cause us to reconsider how we *evaluate* forecasts. In laboratory studies – particularly using almanac questions (intellective tasks) and short-term forecasts – forecasting accuracy can be determined, and therefore the factors responsible for a particular group-based forecasting method producing a better forecast can be investigated. But in medium- to long-term forecasts, other criteria need to be brought to bear in assessing whether a particular group-based process has value.

In summary, we believe that this special issue on group-based forecasting does advance our knowledge of the domain in a number of ways. But there are others paths we might follow, and we urge the research community to consider them.

References

Arnold, V., Clark, N., Collier, P.A., Leech, S. A. & Sutton, S.G. (2006). The differential use and effect of knowledge-based system explanations in novice and expert judgment decisions. *MIS Quarterly*, 30, 79-97.

Benda, W. G. G. & Engels, T. C. E. (2011). The predictive validity of peer review: a selective review of the judgmental forecasting qualities of peers, and implications for innovation in science. *International Journal of Forecasting*...

Berg, J., Nelson, F & Rietz, T.A. (2008). Prediction market accuracy in the long run. *International Journal of Forecasting,* 24, 285-300.

Brockriede, W. & Ehninger, D. (1960). Toulmin on argument: an interpretation and application. *Quarterly Journal of Speech,* 46, 44-53.

Cairns, G., Sliwa, M. & Wright, G. Problematizing international business futures through a 'critical scenario method. *Futures,* in press.

Cairns, G., Wright, G., Bradfield, R., van der Heijden, K. & Burt, G. (2006). Enhancing foresight between multiple agencies: issues in the use of scenario thinking to overcome fragmentation. *Futures,* 38, 1011- 1025.

Corey, G. A. & Merenstein, J. H. (1987). Applying the ischemic heart disease predictive instrument. *The Journal of Family Practice,* 25, 127- 133.

De Marchi, B. & Ravetz, J.R. (1999) Risk management and governance: a post-normal science approach, *Futures*, 31, 743-757.

Graefe, A. & Armstrong, J.S. (2011). Comparing face-to-face meetings, nominal groups, Delphi and prediction markets on an estimation task. *International Journal of Forecasting...*

Green, K.C. (2002). Forecasting decisions in conflict situations: a comparison of game theory, role-playing, and unaided judgment. *International Journal of Forecasting,* 18, 321-433.

Green, K.C. & Armstrong, J. S. (2011). Role thinking: standing in other people's shoes to forecast decisions in conflicts. *International Journal of Forecasting...*

Goodwin, P. & Wright, G. (2010) The limits of forecasting in anticipating rare events. *Technological Forecasting and Social Change*, 77, 355-368.

van der Heijden, K., Bradfield, R., Burt, G., Cairns, G. & Wright, G. (2002). *The Sixth Sense: Accelerating Organisational Learning with Scenarios*, Chichester: Wiley

Jodlbauer, B. & Jonas, E. (2011). How does perception of strategic behaviour influence acceptance of advice? *International Journal of Forecasting...*

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

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

Klenk, N. L. & Hickey, G. M. (2011). A virtual and anonymous, deliberative and analystic participation process for planning and evaluation: the concept mapping policy Delphi. *International Journal of Forecastin*g...

Laird, F.N. (1989). The decline of deference: The political context of risk communication. *Risk Analysis*, 9, 545-550.

Landeta, J. & Barrutia, J. (2011). People consultation to construct the future: a Delphi application. *International Journal of Forecasting*...

Nemeth, C., Brown, K. & Rogers, J. (2001). Devil's advocate versus authentic dissent: stimulating quantity and quality. *European Journal of Social Psychology,* 31, 707-720.

Nemeth,C. & Chiles, C. (1998). Modeling courage: the role of dissent in fostering independence. *European Journal of Social Psychology*, 18, 275-280.

Onkal, D., Lawrence, M. & Sayim, K.Z. (2011). Influence of differentiated roles on group forecasting accuracy. *International Journal of Forecasting*...

Rowe, G. & Frewer, L.J. (2000). Public participation methods: A framework for evaluation. *Science, Technology, & Human Values*, 25, 3-29.

Rowe, G. & Frewer, L.J. (2005). A typology of public engagement mechanisms. *Science, Technology, & Human Values*, 30, 251-290.

Rowe, G. & Wright, G. (1996). The impact of task characteristics on the performance of structured group forecasting techniques. *International Journal of Forecasting*, 12, 73-89.

Rowe, G. & Wright, G. (2001). Differences in expert and lay judgments of risk: Myth or reality? *Risk Analysis*, 21 (2), 341-356.

Rowe, G., Wright, G. & Bolger, F. (1991). The Delphi technique: A re-evaluation of research and theory. *Technological Forecasting and Social Change*, 39, 235-251.

Schweiger D. M, Sandberg, W.R., & Ragan, J.W. (1986). Group approaches for improving strategic decision making: A comparative analysis of dialectical inquiry, devil's advocacy, and consensus. *Academy of Management Journal*, 29, 51-71

Schweiger D. M, Sandberg, W.R., & Rechner, P.A. (1989). Experiential effects of dialectical inquiry, devil's advocacy, and consensus approaches to strategic decision making. *Academy of Management Journal*, 32, 745-772.

Seick, W.R. & Arkes, H.R. (2005). The recalcitrance of overconfidence and its contribution to decision aid neglect. *Journal of Behavioral Decision Making,* 18, 29-53.

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

Soll, J.B. & Mannes, A.E. (2011). Judgmental aggregation strategies depend on whether the self is involved. *International Journal of Forecasting...*

Van de Ven, A. H. & Delbecq, A. L. (1974). The effectiveness of Nominal, Delphi, and Interacting group decision making processes. *Academy of Management Review,* 17, 605-621.

Van Swol, L. (2011). Forecasting another's enjoyment versus giving the right answer: trust, shared values, task effects, and confidence in improving the acceptance of advice. *International Journal of Forecasting...*

Van Swol, L.M. & Sniezek, J.A. (2001). Trust, confidence, and expertise in a Judge-Advisor System. *Organizational Behavior and Human Decision Processes, 84,* 288-307.

Webler, T. (1995). 'Right' discourse in citizen participation: An evaluative yardstick. In *Fairness and competence in citizen participation: Evaluating models for environmental discourse*, edited by O. Renn, T. Webler, and P. Wiedemann, 35-86. Dordrecht, Netherlands: Kluwer Academic Publishers.

Wright, G. (2002). Game theory, game theorists, university students, role-playing and forecasting ability. *International Journal of Forecasting,* 18, 383-387.

Wright, G & Goodwin, P. (2009). Decision making and planning under low levels of predictability: enhancing the scenario method. *International Journal of Forecasting, 25, 813-825.*

Yaniv, I. (2011). Group diversity and decision quality: amplification and attenuation of the framing effect. *International Journal of Forecasting...*