ELSEVIER

International Journal of Forecasting 12 (1996) 73–89

international journal of forecasting

The impact of task characteristics on the performance of structured group forecasting techniques

Gene Rowe[a], George Wright[b,*]

[a]*Department of Psychology, University of Surrey, Guildford GU2 5XH, UK*
[b]*School of Business and Economic Studies, University of Leeds, Leeds, LS2 9JT, UK*

Abstract

A number of approaches (e.g. Delphi) have been developed to structure information exchange in nominal groups to aid judgment under uncertainty. We address the rationale underlying such techniques and point to shortcomings in research on the validity of the techniques per se. We advocate a new direction of research, moving from the holistic appraisal and comparison of techniques, towards consideration of the processes responsible for inducing individual judgment *change* during such interventions, and the identification of key personal and situational factors that may predict such change.

We describe one repeated-measures design in which groups of subjects are asked to forecast economic and political events using Delphi-like procedures that differ according to the presence/absence and nature of feedback. Characteristics elicited from subjects include measures of 'confidence', 'desirability of forecasts', and 'self-rated expertise'. We analyse how individual characteristics on the above dimensions are related to (a) accuracy and (b) willingness to change estimates, for each of the Delphi-like procedures. Results indicate an increase in accuracy over each of the three conditions, including a no-feedback control, although similar outcomes appear to derive from different processes. Measurement of panellists' 'self-rated expertise', 'objective expertise', 'confidence', and 'desirability of outcomes' show differential utility for a-priori panellist selection for structured groups. We discuss results in terms of the interaction of technique and panellist specifications. In general, our results support the "theory of errors" as an explanation of the effectiveness of the Delphi technique. Implications of our findings for the creation of judgment-aiding techniques, panellist selection, and future research are discussed.

Keywords: Judgmental forecasting; Prediction; Forecast evaluation; Forecast error; Expert opinion; Delphi

1. Introduction

Forecasts and decisions are often made informally by groups of individuals. However, such interactions of individual group members can evidence 'process loss' due, for example, to ineffective communication.

One response to the potential 'process loss' in interacting groups (Steiner, 1972) is to structure the communication exchange of group members. The groups involved in such procedures are usually termed 'structured groups', and, of the numerous formalised procedures, perhaps the

* Corresponding author.

SSDI 0169-2070(95)00658-3

best known and most studied is that of the Delphi technique (e.g. Dalkey, 1969). This technique, involving the repeat questioning of individual group members and aggregate feedback of the group's responses to individual (non-interacting) group members has not, however, been free of controversy.

Much of the criticism that has been levelled at Delphi has focused upon the relative lack of empirical evaluation of the technique's validity, and upon the poor conduct of the early Delphi studies (e.g. Sackman, 1975; Hill and Fowles, 1975; Stewart, 1987). Rowe et al. (1991), however, have criticised the empirical methodology used in more-recent Delphi evaluations–a methodology generally advocated by the earlier critics. They classified studies of Delphi into either 'process' or 'technique-comparison' types. The former are concerned with the internal processes and interactions related to the detailed applications of the technique per se. The latter (comprising the majority of studies since the mid-1970s) are concerned with comparing the Delphi technique to other judgmental or forecasting procedures in order to determine which is 'best'. Rowe et al. (1991) noted the variability of findings concerning Delphi effectiveness, and attributed the equivocacy of results, largely, to the highly variable formats and implementations of the technique. For example, experimental exemplars of Delphi have tended to vary in terms of the number of rounds employed, the format of the feedback, the number of panellists employed and the nature of the judgment tasks. Furthermore, most empirical examples of Delphi vary in distinctive ways from the technique ideal as advocated by Delphi's originators and proponents (e.g. see Linstone, 1978; Martino, 1983). For example, the 'classical' Delphi has an unstructured first round, uses expert panellists, and uses feedback comprising medians plus written rationales from panellists whose estimates fall outside the upper and lower quartiles. The typical laboratory Delphi, on the other hand, uses structured rounds (and usually no more than two of these), student subjects, and feedback of simple means or medians.

Rowe et al. (1991) point out that if these dimensions (on which Delphi implementations vary) are shown to have significant influences on technique validity (or indeed, on any other measure of performance) then it is arguable whether Delphic versions can be considered comparable; and indeed, perhaps they should be considered different techniques. Theoretical, empirical, and intuitive bases for this intuition would suggest this to be the case. For example, it has been theoretically demonstrated how the effectiveness of statistical groups vary according to aspects such as the number of subjects, their individual validity, and their inter-judgmental correlations (e.g. see Einhorn et al., 1977; Hogarth, 1978). It is likely that such factors will also be of importance in structured and interacting groups. Research in social psychology also suggests that different feedback formats will influence judgment change in qualitatively different manners, e.g. through social comparative or persuasive argument means (e.g. see Isenberg, 1986). Therefore, to label all technique exemplars as 'Delphis' because they share certain gross features is to invite confusion: if one variant provides more accurate estimates than comparison interacting groups (as an example), and another variant does not, then what does one conclude about the idealised Delphi? Unfortunately, rather than concentrating upon the specificity of results, researchers have tended to draw grand conclusions about the effectiveness of the techniques per se, viz. Delphi is good or Delphi is bad. The evidence for such claims is conflicting, because the various studies have failed to clarify and analyse the precise conditions under which the Delphi was implemented.

2. The present study: a framework

In line with the critique of Rowe et al. (1991), this study eschews the largely ungeneralisable technique-comparison approach, and attempts to consider the Delphi technique in terms of the processes of judgment change that occur within the procedure. Judgment change is here conceptualised as coming about through the inter-

action of internal factors related to the individual panellist (e.g. characteristics related to personality, expertise, confidence), and external factors related to the technique design and judgment scenario (e.g. feedback format, task nature). From this perspective, the role of research should be to identify which factors are most important in explaining how and why an individual changes his/her estimate during the procedure, and which factors bear greatest responsibility for ensuring that such change is in the desired direction (i.e. towards a more accurate judgment or forecast).

3. The task

Past Delphi studies (both empirical and applied) have generally been used for assessing judgment in tasks that may broadly be classified as of one of three types. These are: forecasting tasks, policy formation tasks, and information-poor judgment tasks. Policy formation tasks are those where subjective opinions and views are sought because objective optimal solutions are difficult to specify. Information-poor judgment tasks are typified by studies using almanac-type questions in which unknown values with objectively optimal solutions are assessed. We have focused here on a task involving the forecasting of political and economic events for a period of about 2 weeks in the future. The short time period of the forecasts ensured that we could verify judgment accuracy, and do so in a task that is more ecologically valid than those using easier-to-verify almanac items.

The results of our study will be less relevant for tasks of the policy formation type. This is because our study is able to assess the accuracy of the forecast made. In policy formation problems alternative criteria for validation are needed, such as the satisfaction of panellists with the ultimate outcome. In terms of the factors influencing judgment change, however, it is feasible that the results from our study may be applicable to policy formation scenarios to some degree.

The domain of politics (and relatedly, economics) was chosen for our forecasting task for a number of reasons. First, politics is an ecologically valid topic for forecasting–unlike many of the topics addressed in past studies (particularly in those using almanac questions). Second, politics is a domain in which passions are often roused and factors such as the desirability of an outcome may influence judgment change and accuracy. Again, this trend is perhaps in opposition to the kind of emotionally neutral topics that typify the majority of experimental studies, and hence this arguably makes our task more typical of those confronted by real people making real decisions. Third, the task is of such common interest that it ensures a wide range of expertise may be obtained from student subjects, for many of whom it is undoubtedly a topic of great importance.

4. External factors

The one independent variable we examine is that of 'feedback'–an 'external' factor that Rowe et al. (1991) identified as a potentially important source of the variability of findings in technique-comparison studies involving Delphi, and one on which there have been a small number of 'process' studies. Boje and Murnighan (1982) found that a Delphi-like procedure (with feedback comprising the estimates of the panellists, supported by a single reason from each), resulted in judgments that became less accurate over rounds, while a procedure that simply required subjects to 'think again' about previous estimates (i.e. with no feedback from others) actually resulted in increased accuracy. Similarly, Parenté et al. (1984) separated the contributions of polling and feedback in a Delphi-like approach by using a design that allowed an orthogonal decomposition of their effects on accuracy. They also found that 'iteration' alone resulted in error reduction for 'when' a predicted event would occur, while feedback alone did not. (The feedback in this study comprised the percentage of the subjects who thought that each event would occur within a specified time period and, when a

majority thought that an event would occur, a median predicted time for this.)

In neither of the above studies is it clear why an iterative procedure should outperform a similar procedure in which extra information is provided in the form of feedback from a non-interacting group of others. One possible reason could be the relative superficiality of the feedback typically used in experimental studies. For example, a figure which only indicates the percentage of others agreeing or disagreeing with one's position may not be compelling, and certainly does not provide much in the way of novel ideas or arguments. Similarly, potentially profound feedback of arguments may prove ineffectual if the arguments represent subjects' attempts at justifying quantitative estimates on unusual almanac items. If such explanations have any validity, then we might expect that a more profound type of feedback of reasons (in conjunction with sensible questions) will lead to increased influence and accuracy over rounds in comparison both to a simple statistical indication of the normativeness of one's position, and to a simple no-feedback iterative process (for more discussion of this issue, see Rowe et al., 1991). Indeed, Best (1974) provided some (slight) evidence to support this position, finding that, for one of two task items, a Delphi group which was given feedback of reasons in addition to a median and range of estimates, was significantly more accurate than a Delphi group which was simply provided with the latter information.

Alternatively, the absence of the pressures of a face-to-face group environment may remove all incentive to consider or be persuaded by the arguments of others. This is particularly so when one considers the removal of cues about the confidence, competence, eloquence (etc.), of one's now-nominal colleagues. If this is the case, then perhaps changes in accuracy might prove unrelated to the form of feedback provided. A third alternative could be that subjects, in the absence of overt pressures from others, may be inclined to selectively attend to feedback that agrees with their stated position and ignore feedback that disagrees with it (see, for example, Koriat et al., 1980). As a result, positions become more extreme over rounds–perhaps turning initially poor estimates into even poorer estimates. The role of the form of feedback in this situation is uncertain: perhaps the more profound it is, the more compelling it will be seen by subjects in terms of evidence for their own position.

These are but a few possible explanations for the results of the above studies. This study attempts to examine the role of feedback by seeing how it relates to individuals' accuracy and propensity to change forecasts over rounds. 'Feedback' here comprises one of three types, namely 'iteration' (i.e. essentially a control condition with no feedback), 'statistical' (feedback of statistics representative of nominal group opinion), and 'reasons' (feedback of nominal group rationales). It should be noted here that we are not suggesting that our feedback types (described more fully later) are the most effective possible, or represent the extremes on any particular 'feedback dimension': they are merely exemplars of (a) a statistical type of feedback and (b) a reasons type of feedback. Whether a different combination of information type might alter results is a matter for future study.

In order to limit the number of variables so as not to make results uninterpretable, other 'external' task features (such as 'number of rounds') were not studied, but were kept constant across the experimental conditions. The specifics of the design will be elucidated in the method section.

5. Internal factors and the "theory of errors"

Although the Delphi technique lacks a coherent and rigorous theoretical underpinning, one axiomatic explanation of its potential effectiveness has been put forward, and is known as the theory of errors (Dalkey, 1975). Parenté and Anderson-Parenté (1987) have summarised this 'theory', and have drawn implications from it concerning how iterated polling and feedback may lead to improved accuracy over Delphi rounds. They describe the hypothesised existence of two sets of individuals within the more general population of panellists, known as the

holdouts and the swingers. The former are assumed to comprise the more-knowledgeable panellists, and the latter the less-knowledgeable ones. The theory (according to Parenté and Anderson-Parenté, 1987) suggests that, with each iteration of the questionnaire, the least accurate panellists will, essentially through reflecting on the feedback, realise their relative lack of knowledge and be drawn towards the median value on subsequent rounds (hence the 'swingers'). By contrast, the more-knowledgeable panellists will appreciate their relative level of knowledge and maintain their judgments in the face of feedback (hence the 'holdouts'). From this, it can be theoretically demonstrated that the median response of the entire group, 'M', should move towards the true value, 'T', over rounds. However, empirical support for this explanation of the Delphi process is scant, and a number of other models can be proposed to account for changes in judgments, and in judgment accuracy, over rounds. We shall define here a number of 'internal' factors that may explain why some subjects are 'holdouts' and others are 'swingers'. These internal factors are elucidated below, with the corresponding hypotheses explained later.

(A) Self-rated expertise. An alternative discriminator between 'holdouts' and 'swingers' may be self-rated, or subjective (as opposed to objective) expertise. Intuitively, one would expect an individual's belief in their own expertise to be related to their willingness to alter judgments in the face of feedback from others. This may or may not be related to objective expertise (depending on the accuracy of self-perception), but would seem a potentially more useful concept for understanding judgment change in Delphi-like groups than the latter. It is also a more practical tool for panellist selection because the self-rating may be obtained prior to implementation of the procedure, rather than after it. Concerning the validity of self-rated expertise, evidence is equivocal (e.g. Best, 1974; Brockhoff, 1975; Wright et al., 1994), and little work has been done on examining how likely individuals are to change judgments, on the basis of such ratings, in structured groups.

(B) Confidence in assessment. The importance of confidence in judgment and choice has been widely discussed (e.g. Sniezek, 1992). Sniezek and Henry (1989) note how the confidence that a group has in its judgment may determine whether its decision is implemented, and may therefore be as important to the ultimate outcome as decision quality itself. Indeed, it is clear that high confidence can directly affect performance or outcome, as when a group works hard to implement a decision in which it believes, or when a group that expresses high confidence is given additional support, which enhances its chance of success (e.g. in self-fulfilling prophecy). Group confidence may also serve as an end in itself, as when, for example, groups are used to make a decision predominantly as a strategy to gain the commitment and acceptance of a decision by those who will implement it (Mason and Mitroff, 1981). There are also models of group effectiveness which treat confidence as an important outcome (e.g. Gladstein, 1984).

In many situations, however, the goal of a group meeting (whether interacting or structured) will be to produce an enhanced decision, rather than merely to establish greater group confidence in that decision. Problematically, establishing the objective accuracy of judgments or choices is often not feasible, particularly in the kind of tasks for which groups are typically assigned (e.g. Hart, 1985). In such cases, a group's evaluation of its product may be the only measurable determinant of the group's effectiveness. However, for confidence to be an appropriate performance measure, it must be demonstrated to be related to objective performance quality.

Sniezek (1992) has reviewed research on group confidence and concludes that this generally supports the validity of confidence as an indicator of performance accuracy, i.e. group confidence generally increases in line with increased judgment quality. By 'group confidence' Sniezek means "the confidence of a group in its judgment or choice–an expression that is formed through the collective actions of persons in the group" (p. 127). This definition dissociates the

holistic group confidence judgment from the specific confidence judgments of the group members. However, differences exist between the findings on group and individual confidence. For example, research on the calibration of individual judgments has generally found people to be over-confident (e.g. Lichtenstein et al., 1982), i.e. their confidence estimates are of imperfect appropriateness. Furthermore, effects of confidence in structured groups have not been extensively examined–although research on individual confidence is liable to be of some relevance here.

Our main interest is centred on whether the initial confidence ratings of the individual members of a structured group are relevant predictors of the individuals' accuracy and propensity to change judgments as a result of the structured group process. Any relationship between initial confidence and subsequent accuracy or propensity to change judgments in the face of information from others will have implications for the selection of individuals into structured groups, and implications for whether confidence can, as Sniezek (1992) hopes, be used as a surrogate for elusive objective accuracy criteria.

(C) Subjective desirability. Subjective desirability has been studied in a variety of judgment and forecasting situations and has been shown to influence judgment accuracy and performance. Zakay (1983) found that subjects perceive desirable life events as more likely to occur to themselves than to another person similar to themselves (and vice versa for undesirable events). Evidence for such an 'optimistic bias' has been demonstrated elsewhere (e.g. Weinstein, 1980). Milburn (1978) found that subjects perceived desirable events as becoming increasingly more likely, and undesirable events less likely, in each of four successive future decades.

Wright and Ayton (e.g. 1987, 1989, 1992), in a series of studies, considered the influence of desirability (and other factors) on probabilistic forecasting. Overall, they found that desirability had a greater influence on the values (they increased) and accuracy (this decreased) of probability forecasts of personal as opposed to non-personal events, and of events that were less-imminent than those that were imminent. For example, Wright and Ayton (1989) found that desirability was positively correlated with probability of occurrence and with over-confidence, but negatively correlated with calibration (a measure of appropriateness of probability estimates). These correlations were greater for events that had personal implications for the forecasters, than for peer-related or world events. Thus, in certain circumstances, high desirability leads to over-confidence in the expectation that an event will occur (and the converse for undesirable events), with associated reductions in the accuracy of forecasts. The impact of desirability on judgment change in structured group situations has not, however, been studied, although an effect would seem likely.

Above are just a few of the possible 'internal' factors that *may* prove efficient predictors of whether an individual will change his/her judgment or prediction in a structured group approach like Delphi. In this study, we investigate Parenté and Anderson-Parenté's (1987) interpretation of the theory of errors to see which of these internal factors, and in what combination, best explains the reality of the situation. In essence, the theory of errors provides a *single* mechanism for explaining what happens in a Delphi application. We, however, believe that a number of other possible mechanisms may be involved which depend upon the specifics of one's Delphi design (i.e. an interaction of the *internal* factors with the *external*). Specific hypotheses follow after a consideration of our performance measures.

6. Performance/response measures

All structured group techniques (and indeed any technique aimed at gaining enhanced judgment) are concerned primarily with inducing change in their members, either in the hope of achieving tighter consensus or greater accuracy. It is therefore of interest how influential the specific techniques are at inducing change, particularly change in the appropriate direction. The

influence of internal factors (such as confidence) will also be considered. Therefore, performance is measured using two main criteria, namely:

(1) Accuracy. This measure concerns the accuracy of point forecasts made on Rounds 1 and 2, in terms of mean absolute percentage error (MAPE), and the extent to which accuracy improves over the two rounds.

(2) Change. This is a measure of the extent to which subjects change their point forecasts from the first to the second round, in terms of changes in a z-score measure (see Results section for formulation). It ignores the directionality of change (i.e. whether change leads to better or worse forecasts).

7. Hypotheses

The hypotheses below relate to the main effects we expect to observe in relation to each of the 'external' and 'internal' variables we have identified, in terms of our two performance measures of *change* and *accuracy*. Because the interactions between our predictor variables are potentially complex, we will make no specific hypotheses about these.

Further, *most* of the hypotheses consider the relationships between our predictor variables and (a) Round *1* (initial) accuracy, and (b) *degree of change* over rounds, but *not* (c) Round *2* (ultimate) accuracy. This is because ultimate accuracy will be determined by the (potentially complex) interactions between aspects such as *change* and *initial* accuracy (and mediated by our predictor variables). We will consider relationships between our predictor variables and ultimate accuracy, but again make no specific hypotheses here about these.

In addition, predictions about *differences* in Round 1 accuracy are made irrespective of feedback condition, since first round estimates occur before this experimental manipulation. Those hypotheses which consider the *differential* effect of the feedback conditions are made with the important proviso that subjects under each condition face roughly the same average degree of opinion pull against their position. This latter aspect may be an important factor affecting the judgment change variable and (though the random allocation of subjects, items, and conditions should control for varying influence), we will attempt to confirm this in our analysis.

(1) Effects of feedback. We predict that (a) *improvements* in accuracy and (b) propensity to change judgments, will both depend upon the depth of feedback provided, hence: reasons > statistical > iteration. The more profound the type of feedback, the more its influence on an individual panellist's propensity to change an initially inaccurate forecast.

(2) Objective expertise. In line with the theory of errors (e.g. Dalkey, 1975; Parenté and Anderson-Parenté, 1987), we hypothesise that the more accurate the individual (in terms of least Round 1 predictive error), the less they will change judgments over rounds (i.e. 'holdouts' are experts, and 'swingers' are not). Since only two of our conditions meet the broad requirements of Delphi (involving some form of feedback), we predict that this relationship will hold for the statistical and reasons conditions only, and make no specific hypothesis for the iteration condition.

(3) Self-rated expertise. We predict that (a) high self-rated expertise is related to more accurate Round 1 predictions (i.e. self-rated expertise has some validity, in line with Wright et al., 1994), and (b) high self-rated expertise is related to low propensity to make judgment changes over rounds. The lower propensity to change will be most evident in the iteration, then statistical, then reasons condition, due to the increasing power of the feedback to counter the anticipated intransigence effect of high self-rated expertise.

(4) Confidence in answers. We predict that (a) high Round 1 confidence is related to more accurate Round 1 predictions, and (b) high Round 1 confidence is related to low propensity to make forecast changes over rounds (cf. Sniezek and Henry, 1989; Sniezek, 1992). The lower propensity to change will be most evident in the iteration, then statistical, then reasons condition, due to the increasing power of the feedback to counter the anticipated intransigence effect of high confidence.

We make no predictions about second round confidence, since the relationship between this and 'change' and 'accuracy', and the potential influence of the different feedback types on both variables, is less clear. Nevertheless, it has been argued (e.g. Sniezek, 1992) that confidence estimates may (given the difficulty of assessing validity of responses in the real world) reasonably act as surrogate indicators in those situations where it is difficult to directly assess technique efficiency. We will consider second round confidence ratings in this light.

(5) Desirability of predicted outcomes. We predict that (a) Round 1 high and low desirable items will be less accurately predicted than mid-desirable items (i.e. those viewed with relative indifference) (cf. Wright and Ayton, 1989), and (b) high and low desirable items will be less susceptible to judgment change than mid-desirable items, an effect which will be most pronounced in the iteration condition, less pronounced in the statistical condition, and least pronounced in the reasons condition.

8. Method

8.1. Subjects

Sixty undergraduate students from the Bristol Business School and Bristol Polytechnic were recruited to complete two questionnaires on separate days, each taking approximately 45 minutes. One subject who had completed the first round questionnaire failed to show up for the second questionnaire 2 days later, and his responses were omitted from further analysis. The subjects were randomly allocated to 12 nominal groups comprising five members each.

Subjects were paid a total of £6 sterling on completion of the second questionnaire.

8.2. Materials

The questionnaires comprised 15 political and economic events for which short-term forecasts (i.e. for approximately 2 weeks thence) were required. Most of the 15 items related to the (then) forthcoming local government elections, and concerned how the dominant political parties would fare, in terms of gains or losses in total seats or councils, across the country. Other forecasts concerned events such as whether inflation or unemployment would show a rise or fall on the announcement of the next government statistics. Such items seemed highly pertinent to the elections and, in particular, the performance of the ruling political party. The addition of such items ensured there were sufficient for a symmetrical and balanced design, providing enough items for each condition of the independent variable in a repeated measures design. The 15 items were randomly allocated to three sets of five items that were kept constant in terms of item membership throughout. These three sets were permutated into the six possible arrangements, and ten questionnaires were produced for each of these. Each nominal group of five subjects received questionnaires with the forecast events in the same arrangement, such that two groups received questions ordered according to each of the possible arrangements.

Each event had a number of associated questions. First, subjects were required to select one of four exclusive and exhaustive options as a prediction. These generally specified whether a "large" or "small" gain or loss would occur, for example, whether the Labour party would gain a large number of seats at the elections, gain a small number, lose a small number, or lose a large number, with actual values attached to the definitions of large and small ("no changes" are unlikely, and these were nominally added to the "small loss" category to ensure the options were exhaustive). We expected that such qualitative responses (as opposed to quantitative responses) would make subsequent questions about 'confidence' and 'desirability' more meaningful to subjects. That is, we expected subjects would find it easier to express themselves about categorical responses on which they could anchor (and be seen as 'right' or 'wrong'), rather than about quantitative responses (in which they would largely be 'wrong', but to a greater or lesser degree). Further, the categorical responses enabled us to take a simple measure of degree of

homogeneity of opinion of the subjects in each group on each question, allowing us to assess degree of opinion pull in each case (see Hypotheses and Results sections for further explanation). At this point, *all* subjects were informed by questionnaire instructions to "give one reason to support your prediction", and were reminded that their reasons might be used as feedback to the other members of their nominal group, and hence that they should be precise and considered. These reasons were only actually used as feedback in the appropriate condition, but were initially elicited from all subjects to ensure the first-round comparability of all conditions.

Next, the questionnaires required subjects to give a quantitative figure or point prediction of the exact expected change for each event. This is a more typical form of collecting predictions than our categorical method and provides a higher scaling level of data.

Subjects were then required to rate every response (on seven-point scales) in terms of their 'confidence' in their categorical prediction and the 'desirability' of their predicted outcome. At the beginning of each questionnaire subjects were also required to rate (on a seven-point scale) their knowledgeability of "domestic politics and economics".

Second round questionnaires were identical to those of the first round, in terms of the ordering of items (for subjects in each nominal group), and the responses required from subjects–save for the omission of the self-rated expertise question. Additionally, each questionnaire came supplied with a 'feedback sheet' that indicated to each subject the appropriate estimates of their nominal group. 'Feedback' for the five 'iteration' items comprised a simple reiteration of (a) the point estimate the subject had made for that item in the first round and (b) their category prediction. 'Statistical' feedback comprised the median and the actual responses given by the (four or five) nominal group members for each item (i.e. including the subject's own responses). The feedback had both qualitative and quantitative aspects–for example, the median response might be represented as a 'small gain in seats' with an associated median point value. The median values were emphasised for these items as the representations of the 'group average'. 'Reasons' feedback comprised a single written reason from each member of the nominal group on why they had made their (stated) particular categorical prediction (with no emphasis on the median or group average), and included the particular subject's own estimate and reason. Confidence and desirability values were not used as feedback in any of the conditions.

8.3. Design

The single independent variable 'feedback' had three levels ('iteration', 'statistical' and 'reasons'), which varied according to the nature of the feedback provided to subjects on the second of the two rounds of the study. The design was of a repeated measures nature, with each subject receiving each of the types of feedback for five (of the 15) questions. The order in which subjects received the feedback types was counterbalanced, so that each of the six possible sequences were given to two of the nominal groups of five subjects (each of the 12 sets of questionnaires having already been counterbalanced according to question sets, as noted earlier).

8.4. Procedure

Subjects were randomly allocated to the 12 nominal groups of five. The first round questionnaires were presented, and the subjects were told that they had each been allocated to a nominal group comprising four others, whose membership was known only to the experimenters, such that their own anonymity was ensured. During the next 2 days the collected results were collated so that individually tailored feedback sheets could be constructed for each subject. At the beginning of the second round of study, 3 days later, subjects were given a second round questionnaire that was identical to the one they had completed in the first round (except for the removal of certain now-redundant questions), plus the feedback sheet reporting results from their nominal group. Because of the design, each

subject received feedback of all three types for different sets of five items. Subjects were instructed to complete the questionnaire as before, but this time to refer to their feedback sheet and to consider the opinions and positions of the other members of their nominal group before making any decisions. For items for which subjects got no feedback per se (i.e. only the reiteration of their own first round predictions), subjects were asked to "think again" about the event and to "try to think of any additional information that might affect your forecast".

The outcomes of the 15 forecast events were variously obtained from the Press Association, and from national newspapers such as *The Times* and *The Independent*. These results enabled us to test the accuracy of the predictions of the subjects. The gap of 3 days between rounds could have induced our subjects to pay especial attention to relevant news coverage of our forecast items in the intervening period and this may have induced opinion change. However, we would anticipate such opinion change to have a random effect across our experimental treatments.

9. Results

Recall that in the Hypotheses section we noted that our hypotheses considering the differential effects of the feedback conditions were made with the proviso that subjects under each condition faced roughly the same average degree of opinion pull against their position. To address this issue, we quantified the *degree of homogeneity* of each group on each item by counting the number of different qualitative *categories* that were used by the five (or four) subjects in a group. Thus, a score of 4 for an item would indicate that every one of the four categories was used, with relatively high disagreement in the group (low homogeneity) and a high opinion pull (per subject, on average) away from group members' initial positions. A score of 1 would indicate that all subjects in a group chose a single one of the four categories, and that they were in relative agreement about their prediction (with high homogeneity–only differing in quantitative estimates), and with a low opinion pull (per subject, on average) away from group members' initial positions. Additionally, it could be argued that opinion pull is also a function of the level of agreement *between* those disagreeing with a position, for example, for any single subject the highest pull on his/her opinion would be if he/she held a view diametrically opposed to the rest of the group.

We quantified *homogeneity* for each group for each question under all three conditions, and then calculated mean homogeneity values for each condition. The values obtained were: 2.33 (iteration), 2.18 (statistical), and 2.27 (reasons), which do not differ significantly. This result, together with the similarity of the three distributions, suggests that degree of opinion pull (across conditions) was adequately controlled for by our randomisation procedure, and we will not consider this factor further in our analysis.

9.1. Changes in accuracy over rounds

In order to make the data comparable, the raw forecast estimates were converted into percentage error scores for each of the 15 items. Mean absolute percentage error (MAPE) scores were then calculated for all subjects over each condition and round (there was no evidence of substantial skew in the MAPE distributions, so untransformed MAPE values were used in subsequent analyses). The average MAPE values are tabulated in Table 1. As anticipated, accuracy increased (i.e. average MAPEs decreased) in each of the conditions over rounds, although from Table 1 it can also be seen that there were substantial variations in the first round MAPEs across condition (indeed, differences across conditions on the first round are of similar magnitudes to those *within* conditions across rounds). It is not entirely clear why these across-condition differences emerged, although it may be related to an unexpected interaction between the actual expertise of the subjects and the differential difficulty of the three sets of five items. (If, for example, the least-accurate sub-

Table 1
Mean absolute percentage error scores over rounds for the three experimental conditions

Condition	Round 1 MAPE	Round 2 MAPE	Difference (%)	t (df)	P (1-tailed)
Iteration					
Mean ($n = 59$)	0.764	0.707	0.062	1.80 (58)	0.039*
SD	0.278	0.235	(6.2%)		
Statistical					
Mean ($n = 59$)	0.989	0.828	0.070	1.76 (58)	0.042*
SD	0.387	0.387	(7.0%)		
Reasons					
Mean ($n = 59$)	0.798	0.684	0.114	3.03 (58)	0.0018**
SD	0.400	0.313	(11.4%)		

* $P < 0.05$; ** $P < 0.01$.

jects received the hardest set of five items in one condition, this might lead to a lower average Round 1 MAPE for that condition than for the others). As reported later, in Subsection 9.3, this conjecture receives indirect support in our correlational analyses. However, given the repeated measures nature of the design, the noted first round differences *between* conditions should not affect *cross*-round comparisons and it is in these latter comparisons that we are interested. Therefore, t-tests were conducted on the difference between the means across rounds within each condition, and all differences proved significant ($P < 0.05$ for iteration and statistical; $P < 0.01$ for reasons).

Table 1 indicates that the greatest degree of improvement in accuracy over rounds (in terms of simple magnitude) occurred in the reasons condition, followed by the statistical condition, and with the least improvement occurring in the iteration condition. This order was predicted in Hypothesis 1(a). However, a one-factor repeated measures analysis of variance which used Conditions (3) as the within-subjects factor and MAPE improvement as the dependent variable, failed to show a statistically significant main effect ($F(2, 116) = 0.61$, $P > 0.05$). We cannot, therefore, reject the null hypothesis in this case. Clearly, a simple iterative procedure, followed by aggregation, can increase judgment accuracy (cf. Boje and Murnighan, 1982; Parenté et al., 1984).

9.2. Propensity to change per condition

It was originally intended to calculate the degree of subject forecast change in a similar manner to the calculation of MAPE scores (using the difference in first and second round forecasts, in place of the difference between forecast and true value). However, the occurrence of subject forecasts of 'zero' necessitated the division of values by zero, and hence the occurrence of sporadic missing values/error terms throughout the matrix of change scores. In order to overcome this problem, subjects' forecasts were converted to z-scores. By determining mean differences in z-scores from each subject on each item on the first and second rounds, a quantitative measure of degree of change was obtained. This change measure (Z) is formulated below.

$$z = \frac{F_{i,k,j} - \bar{X}_{1,k}}{S_{1,k}}$$

where

$$\bar{X}_{1,k} = \frac{\sum_{j=1}^{59} F_{1,k,j}}{N}$$

(the Round 1 mean forecast for a given question over all subjects) and

$$S_{1,k} = \frac{\sum_{j=1}^{59} (F_{1,k,j} - \bar{X}_{1,k})^2}{N-1}$$

(the Round 1 standard deviation of the forecasts for the same question over all subjects).

i = the round number (1 or 2); j = the subject number (1 to 59); k = the question number (1 to 15); N = the number of subjects (59).

A one-factor repeated measures analysis of variance, using Conditions (3) as the within-subjects factor and z-score change as the dependent variable, established a significant main effect for condition ($F(2, 116) = 3.27$, $P < 0.05$). To further examine this effect, planned comparisons were conducted in the form of two-tailed repeated measures t-tests between the change measures in each of the three conditions. The results, which are tabulated in Table 2 reveal that change in the iteration condition was significantly greater than that in the reasons condition ($P < 0.01$). This result is in opposition to the predictions made in Hypothesis 1(b), and seems to indicate that subjects were more liable to change their responses on a second round when in the absence of feedback, rather than in the presence of it.

We next correlated the *degree* of opinion change over rounds (in terms of z-scores) with the degree of improvement of forecasts over rounds (in terms of the *change* in MAPE scores) for every subject. The only significant correlation to emerge was for the reasons condition (two-tailed Pearson's $r = +0.448$, $P < 0.01$, df = 57). Although there appeared to be a similar relationship between these two variables in the statistical condition, this trend did not reach significance ($r = +0.215$, $P < 0.05$, df = 57), while there was no evidence of such a relationship in the iteration condition ($r = -0.100$, $P > 0.05$, df = 57). These results indicate that, although subjects were less inclined to change their forecasts as a result of receiving reasons feedback than other types (see Table 2), when they *did* change forecasts this tended to be for the better, such that the changes led to a reduction in percentage error (i.e. change was in the *appropriate* direction). Although subjects tended to make greater changes to their forecasts in the iteration and statistical conditions than in the reasons condition, these changes did not necessarily tend towards better predictions.

9.3. Objective expertise

Correlations between subjects' first round MAPEs and the z-score change measure revealed that, in line with Hypothesis 2 and the theory of errors, the most accurate individuals (initially) were indeed those who were least prepared to change their judgments on the second round, in both the reasons condition (one-tailed $r = +0.323$, $P < 0.01$, df = 57) and the statistical condition (one-tailed $r = +0.283$, $P < 0.05$, df = 57), although not in the iteration condition (two-tailed $r = +0.09$, $P > 0.05$, df = 57). These results also appear to support the finding of the previous section. They indicate that, in the reasons condition, it was subjects who exhibited relatively poor initial accuracy who had the higher propensity to change forecasts on receiving feedback. This relationship

Table 2
Analysis of difference in propensity to change prediction estimates over rounds, between the three feedback conditions

	Mean z-score change over rounds per condition	1-tailed t-test analysis	
		Statistical	Reasons
Iteration	Mean = 0.4754 (SD = 0.3568)	$t = 1.18$ $P = 0.24$ df = 58	$t = 2.74$ $P = 0.0082$** df = 58
Statistical	Mean = 0.4062 (SD = 0.3288)	—	$t = 1.29$ $P = 0.2$ df = 58
Reasons	Mean = 0.3290 (SD = 0.3071)	—	—

** $P < 0.01$.

that was less strong in the statistical condition and non-evident in the iteration condition. Thus, although the reasons feedback stimulated less judgment change in individuals, it did stimulate the poorest predictors to change. Even though they stimulated more change, other feedback types were not as successful in stimulating appropriate change. These results provide evidence to support the theory of errors (e.g. Dalkey, 1975; Parenté and Anderson-Parenté, 1987).

The most accurate predictors on the first round also tended to be the most accurate predictors on the second round, regardless of condition. Correlations between Round 1 MAPE and Round 2 MAPE for iteration, statistical, and reasons, respectively, were: $r = +0.478$, $r = +0.639$, $r = +0.697$, all two-tailed $P < 0.01$, df = 57.

9.4. Self-rated expertise

Subjects' self-rated expertise on 'politics and economics' was significantly correlated to their first round MAPEs over all items and conditions (one-tailed Spearman's rho = −0.366, $P < 0.01$, df = 57). This supports Hypothesis 3(a): the higher the subject's self-rated expertise, the higher their initial actual expertise (i.e. the lower their Round 1 MAPE). As a predictor of ultimate accuracy, however, self-rated expertise proved less efficient: the correlation between Round 2 MAPE and self-rated expertise only reached significance in the statistical condition (two-tailed Spearman's rho = −0.293, $P < 0.05$, df = 57), but not in the reasons (rho = −0.071, $P > 0.05$, df = 57) or iteration conditions (rho = +0.054, $P > 0.05$, df = 57). This suggests that although self-rated expertise provides a predictor of initial accuracy, its relationship to accuracy after implementation of a structured group procedure is less clear, and, indeed, it was only a predictor in the case of the statistical condition.

Correlation between self-rated expertise and propensity to change (according to the z-score measure) did not provide support for our hypothesis on the relationship between these variables (3(b)). No significant correlation emerged between the expertise and change measures on any of the three conditions. There was, however, a significant correlation between subjects' self-rated expertise and their mean confidence on all items in Round 1 (one-tailed Spearman's rho = +0.590, $P < 0.01$, df = 57), and also in Round 2 (one-tailed Spearman's rho = +0.486, $P < 0.01$, df = 57). This suggests a strong relationship between subjects' initial expertise beliefs and their subsequent confidence in their answers on each of the items.

9.5. Confidence

The data from the measures of confidence, obtained on seven-point scales, were used in a number of ways. Mean confidence values were obtained for each subject for all items from their first round questionnaire, and these values were correlated with the subjects' Round 1 MAPEs. Although there was a trend in line with Hypothesis 4(a), this was not significant (one-tailed Pearson's $r = -0.147$, $P > 0.5$, df = 57), and so we cannot conclude that high first round confidence is related to high accuracy. Further, there was no evidence to support Hypothesis 4(b) concerning the relationship between propensity to change judgments and confidence: correlations between Round 1 confidence and the z-score change measure did not reach significance (i.e. $P > 0.05$) for any of the three conditions ($r = +0.087, -0.169, -0.049$, for iteration, statistical, and reasons, respectively). A further analysis was done on the confidence data across items (rather than across subjects). A similar pattern was found to the above, i.e. there were no significant relationships between accuracy or forecast change and Round 1 confidence ratings.

With regard to second round accuracy, however, we did find evidence of a relationship: Round 1 confidence was significantly correlated with Round 2 MAPE for the statistical condition (two-tailed $r = -0.331$, $P < 0.01$, df = 57), although not for either the reasons ($r = +0.007$, $P > 0.05$, df = 57) or iteration conditions ($r = -0.014$, $P > 0.05$, df = 57). So, alike with high self-rated expertise, high first round confidence does appear to be a systematic predictor of high ultimate accuracy.

Of additional interest to those who would use structured group techniques is the actual effect the implementation has on the confidence of the individuals using the technique. Various authors have suggested that this might be an important criterion of technique effectiveness in its own right, given that the assessment of objective performance criteria (such as accuracy) may be infeasible in the 'real world' (e.g. Sniezek, 1992). Two-tailed t-tests revealed that subjects were, on average, more confident in their second round answers than their first, for all items regardless of condition ($t = 9.19$, $P < 0.01$, df = 58). This pattern was repeated for each of the feedback conditions, and there were no significant differences between these in terms of the degree of increased confidence. To a degree, this increased average confidence over rounds can be seen as broadly appropriate, given that accuracy improved significantly across all three of the conditions (see Table 1).

9.6. Desirability

Mean ratings for the desirability of the predicted events were computed across subjects for all items for the first and second rounds. Mean desirability exhibited little variation from the 'average' (midpoint) rating of 4 (Round 1 mean = 3.955, SD = 0.701; Round 2 mean = 4.250, SD = 0.708), although the difference between mean ratings across rounds was significant (two-tailed $t = 4.96$, $P < 0.01$, df = 58). If we assume that subjects are reliable in giving their ratings, this result suggests either a propensity for subjects to change from less- to more-desirable predictions on the second round, or for subjects to actually perceive that their predicted events have become, on average, more desirable (or a combination of the two).

In order to test Hypothesis 5, desirability ratings (from 1 to 7) were re-coded to take into account the strength of desirability (i.e. magnitude) regardless of whether an event was desirable or not (i.e. direction). Values of 1 and 7 were coded as '3', 2 and 6 as '2', 3 and 5 as '1', and 4 as '0'. A Spearman's correlation was conducted across subjects and questions, comparing the re-coded values of desirability with the variables related to 'accuracy' (Round 1 MAPE) and 'change' (the z-score measure). In line with Hypothesis 5(a), we found evidence of a significant trend of lower accuracy (higher percentage error) with higher magnitude item desirability (one-tailed rho = +0.127, $P < 0.01$, df = 883). We also found evidence for a similar relationship between desirability and Round 2 MAPE, for both the statistical (two-tailed rho = +0.124, $P < 0.01$, df = 883) and the reasons condition (rho = +0.148, $P < 0.01$, df = 883), but not for the iteration condition (rho = +0.056, $P > 0.05$, df = 883). It thus appears that magnitude of desirability acts to bias ultimate accuracy in our two feedback conditions, but that the initial biasing effect is reduced in the iteration condition. In general, these results support the findings of Wright and Ayton (1989).

Against Hypothesis 5(b), however, we found no evidence that strength of desirability was related to propensity to change forecasts over rounds (for all items over all conditions), according to the z-score measure ($r = +0.062$, $P > 0.05$, df = 883). Correlations between Round 1 desirability measures and the change measures for the three conditions were similarly non-significant. Additionally, we found no evidence linking change appropriateness (i.e. improvement in accuracy over rounds) with our desirability measure. We cannot conclude that strength of desirability inhibits people from changing forecasts in the light of feedback from others. Nevertheless, desirability does appear to exert an initial negative effect on forecast accuracy. This appears to persist *after* feedback, although not after a simple iteration procedure.

10. Discussion

The first objective of this study has been to advocate a re-orientation of research methodology in the examination of structured groups. In particular, we have sought to provide support for Rowe et al.'s (1991) contention that the alternative prominent methodology (characterised as the 'technique-comparison' approach) adds little

to knowledge. This is because the latter approach leads to inappropriate generalisations about group techniques on the basis of comparisons of subtly different forms, which may be more or less representative of technique ideals. Our results support this contention, and suggest that at least one characteristic of 'Delphi' groups, namely the format of feedback, is important in determining the process and outcome of structured group interaction. Our second major objective, involving the operationalisation of our advocated methodological approach, has been to identify potential predictors of individual *accuracy* and *propensity to change* in structured groups. Such identification, we hope, will lead to greater understanding of the dynamics of structured groups, and more practically, lead to some guide-lines on the criteria for the selection of group members and for the appropriate constitution of such groups. To this end, all our potential predictors were shown to have *some* relationship either to accuracy and/or to propensity to change.

Table 3 summarises our findings of the relationships between potential predictors of *Round 2 accuracy* and our three experimental conditions.

Our findings suggest that only 'objective expertise' is an appropriate variable for Delphi panellist selection and that removing any estimates initially rated as highly desirable/undesirable would be inappropriate, since bias appears to be reduced over the second round with a commensurate increase in accuracy on those items. However, the a priori selection of panellists on the basis of 'objective expertise' is difficult and it is partly because of this 'identification' problem that groups are used in the first place (e.g. Larreché and Moinpour, 1983; Rowe, 1992).

Additionally, we found, against expectation, no significant relationship between *propensity to change* one's opinion and self-rated expertise, confidence or desirability. However, objective expertise *was* significantly correlated with propensity to change, in both the statistical and reasons (but not iteration) conditions. This result is important, in that it provides evidence to support the theory of errors: subjects who were the best predictors on Round 1 were the least prepared to change judgment in the face of feedback, whilst the poorest predictors were most prepared to change.

That this trend did not occur in the absence of feedback (i.e. the iteration condition) indicates the necessity of feedback in order for a 'Delphi' to be explained by the theory of errors. The exact causal mechanism behind feedback influence is unclear. One explanation may be that it enables the more expert subjects to appreciate their expertise (without which they would have to make such decisions in a vacuum), while providing the relatively less-expert with information about the problem, which is used to direct their judgment changes. This mechanism would point to the presence of at least two types of information in the feedback: firstly, information about the quality of the judges/forecasters, and secondly, information about the forecast problem itself.

In summary, this study has considered just a few of the potentially great number of factors

Table 3
Predictors of Round 2 accuracy

Conditions	Potential predictors			
	Objective expertise (Round 1 accuracy)	Self-rated expertise	Confidence	Desirability
Iteration	√	X	X	X
Statistical	√	√	√	√
Reasons	√	X	X	√

that are liable to affect the performance of an individual in a structured group environment, in terms of propensity to change estimates, and in terms of accuracy. Further study is needed to explain the exact mechanisms behind the influence of the identified factors, and to explore the domain further to identify other influencing variables. In particular, the differential effects of the various types of feedback needs further study, as does the role of 'external' factors such as the judgment domain and the number of rounds. A greater understanding of the influences of these 'external' factors on the 'internal' factors related to the group members, is necessary before we can say much more about the supremacy or appropriateness of any judgment-aiding structured-group format, or about the selection criteria for structured group membership. Nevertheless, by accepting this approach to the issue of aiding judgment and forecasting (in place of the premature approach of conducting 'technique-comparison' studies) one inevitably accepts the contingency of procedures–and one does so at the proper expense of the rather naive belief that any one technique (such as Delphi) is universally either 'good', or 'bad'.

References

Best, R.J., 1974, An experiment in Delphi estimation in marketing decision making, *Journal of Marketing Research*, 11, 448–452.

Boje, D.M. and J.K. Murnighan, 1982, Group confidence pressures in iterative decisions, *Management Science*, 28, 1187–1196.

Brockhoff, K., 1975, The performance of forecasting groups in computer dialogue and face to face discussions, in: H. Linstone and M. Turoff, eds., *The Delphi Method: Techniques and Applications* (Addison-Wesley, London).

Dalkey, N.C., 1969, The Delphi method: An experimental study of group opinion, RM-5888-PR (The Rand Corporation, Santa Monica).

Dalkey, N.C., 1975, Towards a theory of group estimation, in: H. Linstone and M. Turoff, eds., *The Delphi Method: Techniques and Applications* (Addison-Wesley, London).

Einhorn, H.J., R.M. Hogarth and E. Klempner, 1977, Quality of group judgment, *Psychological Bulletin*, 84, 158–172.

Gladstein, D., 1984, A model of task group effectiveness, *Administrative Science Quarterly*, 29, 499–517.

Hart, S.L., 1985, Toward quality criteria for collective judgments, *Organizational Behavior and Human Decision Processes*, 36, 209–228.

Hill, K.Q. and J. Fowles, 1987, The methodological worth of the Delphi forecasting technique, *Technological Forecasting and Social Change*, 7, 179–192.

Hogarth, R.M., 1978, A note on aggregating opinions, *Organizational Behavior and Human Performance*, 21, 40–46.

Isenberg, D.J., 1986, Group polarization: A critical review and meta-analysis, *Journal of Personality and Social Psychology*, 50, 1141–1151.

Koriat, A., S. Lichtenstein and B. Fischhoff, 1980, Reasons for confidence, *Journal of Experimental Psychology: Human Learning and Memory*, 6, 107–118.

Larreché, J.C. and R. Moinpour, 1983, Managerial judgment in marketing: The concept of expertise, *Journal of Marketing Research*, 20, 110–121.

Lichtenstein, S., B. Fischhoff and L.D. Phillips, 1982, Calibration of probabilities: the state of the art to 1980, in: D. Kahneman, P. Slovic and A. Tversky, eds., *Judgment Under Uncertainty: Heuristics and Biases* (Cambridge University Press, Cambridge).

Linstone, H.A., 1978, The Delphi technique, in: R.B. Fowles, ed., *Handbook of Futures Research* (Greenwood Press, Westport, CT).

Martino, J., 1983, *Technological Forecasting for Decision-making*, 2nd edn. (American Elsevier, New York).

Mason, R. and I. Mitroff, 1981, *Challenging strategic planning assumptions* (Wiley, New York).

Milburn, M.A., 1978, Sources of bias in the prediction of future events, *Organisational Behavior and Human Performance*, 21, 17–26.

Parenté, F.J. and J.K. Anderson-Parenté, 1987, Delphi inquiry systems, in: G. Wright and P. Ayton, eds., *Judgmental Forecasting* (Wiley, Chichester).

Parenté, F.J., J.K. Anderson, P. Myers and T. O'Brien, 1984, An examination of factors contributing to delphi accuracy, *Journal of Forecasting*, 3, 173–182.

Rowe, G., 1992, Perspectives on expertise in the aggregation of judgments, in: G. Wright and F. Bolger, eds., *Expertise and Decision Support* (Plenum, New York).

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

Sackman, H., 1975, *Delphi Critique* (Lexington Books, Lexington, MA).

Sniezek, J., 1992, Groups under uncertainty: An examination of confidence in group decision making, *Organizational Behavior and Human Decision Processes*, 52, 124–155.

Sniezek, J.A. and R.A. Henry, 1989, Accuracy and confidence in group judgment, *Organizational Behavior and Human Decision Processes*, 43, 1–28.

Steiner, I.D., 1972, *Group Process and Productivity* (Academic Press, New York).

Stewart, T.R., 1987, The Delphi technique and judgmental forecasting, *Climatic Change*, 11, 97–113.

Weinstein, N.D., 1980, Unrealistic optimism about future life events, *Journal of Personality and Social Psychology*, 39, 806–820.

Wright, G. and P. Ayton, 1987, Task influences on judgemental forecasting, *Scandinavian Journal of Psychology*, 28, 115–127.

Wright, G. and P. Ayton, 1989, Judgmental probability forecasts for personal and impersonal events, *International Journal of Forecasting*, 5, 117–126.

Wright, G. and P. Ayton, 1992, Judgmental probability forecasting in the immediate and medium term, *Organizational Behavior and Human Decision Processes*, 51, 344–363.

Wright, G., G. Rowe, F. Bolger and J. Gammack, 1994, Coherence, calibration, and expertise in judgmental probability forecasting, *Organizational Behavior and Human Decision Processes*, 57, 1–25.

Zakay, D., 1983, The relationship between the probability assessor and the outcomes of an event as a determiner of subjective probability, *Acta Psychologica*, 53, 271–280.

Biographies: George WRIGHT gained a Ph.D in Psychology from Brunel University in 1980. He is currently Professor of Business Administration at Strathclyde Graduate Business School. His research interests include behavioural decision making, IT in financial services and scenario planning.

Gene ROWE gained his B.Sc. from the University of Bristol and his Ph.D from the University of the West of England. He is currently a research fellow at the University of Surrey. His research interests are mainly concerned with judgment and decision making, enhancing judgmental performance, and public perceptions of risk.