EXPERT OPINIONS IN FORECASTING: THE ROLE OF THE DELPHI TECHNIQUE

Gene Rowe
Institute of Food Research, Norwich Research Park, UK

George Wright
Strathclyde Graduate Business School, Strathclyde University

ABSTRACT

Expert opinion is often necessary in forecasting tasks because of a lack of appropriate or available information for using statistical procedures. But how does one get the best forecast from experts? One solution is to use a structured group technique, such as Delphi, for eliciting and combining expert judgments. In using the Delphi technique, one controls the exchange of information between anonymous panelists over a number of rounds (iterations), taking the average of the estimates on the final round as the group judgment. A number of principles are developed here to indicate how to conduct structured groups to obtain good expert judgments. These principles, applied to the conduct of Delphi groups, indicate how many and what type of experts to use (five to 20 experts with disparate domain knowledge); how many rounds to use (generally two or three); what type of feedback to employ (average estimates plus justifications from each expert); how to summarize the final forecast (weight all experts' estimates equally); how to word questions (in a balanced way with succinct definitions free of emotive terms and irrelevant information); and what response modes to use (frequencies rather than probabilities or odds, with coherence checks when feasible). Delphi groups are substantially more accurate than individual experts and traditional groups and somewhat more accurate than statistical groups (which are made up of non-interacting individuals whose judgments are aggregated). Studies support the advantage of Delphi groups over traditional groups by five to one with one tie, and their advantage over statistical groups by 12 to two with two ties. We anticipate that by following these principles, forecasters may be able to use structured groups to harness effectively expert opinion.

Keywords: Delphi, expertise, interacting groups, statistical groups.

In many real-world forecasting exercises, statistical techniques may not be viable or practical, and expert judgment may provide the only basis for a forecast. But which experts should one use? How many? And how should one elicit their forecasts? We will try to answer these questions by examining one widespread technique, the Delphi technique, which was developed to help forecasters aggregate expert opinion. By considering best practice for implementing this technique, we can derive general principles for using expert opinion in forecasting.

Since its design at the RAND Corporation during the 1950s, the Delphi technique has been widely used for aiding judgmental forecasting and decision making in a variety of domains and disciplines. Delphi was originally devised as a procedure to help experts achieve better forecasts than they might obtain through a traditional group meeting. Its structure is intended to allow access to the positive attributes of interacting groups (such as knowledge from a variety of sources and creative synthesis), while pre-empting the negative aspects that often lead to suboptimal group performance (attributable to social, personal, and political conflicts).

Four necessary features characterize a Delphi procedure, namely, anonymity, iteration, controlled feedback of the panelists' judgments, and statistical aggregation of group members' responses. Anonymity is achieved through the use of self-administered questionnaires (on either paper or computer). By allowing the group members to express their opinions and judgments privately, one may be able to diminish the effects of social pressures, as from dominant or dogmatic individuals, or from a majority. Ideally, this should allow the individuals to consider each idea based on merit alone, rather than based on potentially invalid criteria (such as the status of an idea's proponent). Furthermore, by iterating the questionnaire over a number of rounds, one gives panelists the opportunity to change their opinions and judgments without fear of losing face in the eyes of the (anonymous) others in the group.

Between each iteration of the questionnaire, the facilitator or monitor team (i.e., the person or persons administering the procedure) informs group members of the opinions of their anonymous colleagues. Often this "feedback" is presented as a simple statistical summary of the group response, usually a mean or median value, such as the average group estimate of the date before which an event will occur. As such, the feedback comprises the opinions and judgments of all group members and not just the most vocal. At the end of the polling of participants (after several rounds of questionnaire iteration), the facilitator takes the group judgment as the statistical average (mean or median) of the panelists' estimates on the final round.

While the above four characteristics define the Delphi procedure, they may be applied in numerous ways. The first round of the classical Delphi procedure (Martino 1983) is unstructured; instead of imposing on the panelists a set of questions derived by the facilitator, the individual panelists are given the opportunity to identify what issues are important regarding the topic of concern. The facilitator then consolidates the identified factors into a single set and produces a structured questionnaire requiring the panelists' quantitative judgments on subsequent rounds. After each round, the facilitator analyzes and statistically summarizes the responses (usually into medians plus upper and lower quartiles), and these summaries are then presented to the panelists for further consideration. Hence, starting with the third round, panelists can alter their prior estimates in response to feedback. Furthermore, if panelists' assessments fall outside the upper or lower quartiles, they may be asked to give (anonymous) reasons why they believe their selections are correct

even though they oppose majority opinion. This procedure continues until the panelists' responses show some stability.

However, variations from this ideal (the standard definition) exist. Most commonly, round one is structured to make applying the procedure simpler for the facilitator and the panelists; the number of rounds is variable, though seldom goes beyond one or two iterations; and panelists are often asked for just a single statistic, such as the date before which an event has a 50 percent likelihood of occurring, rather than for written justifications of extreme estimates. These simplifications are particularly common in laboratory studies of Delphi and have important consequences for the generalizability of research findings. For comprehensive reviews of Delphi, see Linstone and Turoff (1975), Hill and Fowles (1975), Sackman (1975), Lock (1987), Parenté and Anderson-Parenté (1987), Stewart (1987), Rowe, Wright, and Bolger (1991), and Rowe and Wright (1999).

PRINCIPLES IN THE CONDUCT OF DELPHI

One of the problems with the empirical research that uses Delphi is researchers' lack of concern for how they conduct the technique. Because they use simplified versions of Delphi in the laboratory, versions that depart from the ideal on a number of potentially significant factors (i.e., nature of panelists and type of feedback), it is uncertain how generalizable their results are from one study to the next. Some studies show Delphi to be an effective forecasting tool, and some do not. A harsh interpretation is that the separate studies have generally examined different techniques, telling us little about the effectiveness of Delphi per se. A softer interpretation is that the various versions of Delphi used in research are potentially acceptable forms of a rather poorly specified technique, and that we can examine the unintended variations across studies to distill principles regarding best practice. If we accept this latter interpretation, we can go even further and consider alternative techniques, such as the Nominal Group Technique (NGT), as simply more dramatic versions of the same fundamental structured group approach. (NGT is similar to Delphi except that it allows some group discussion, though individuals still make their final judgments in isolation [Van de Ven and Delbecq 1971].) We use this latter interpretation here.

Because empirical Delphi variations are typically unplanned and occur across studies, few pieces of research directly address how variations in the implementation of Delphi affect its effectiveness. Our principles should not, therefore, be accepted as cast-iron certainties, but as the result of our interpretation, which may be overturned by future research based on planned, within-study variations and controls.

- **Use experts with appropriate domain knowledge.**

Delphi was devised as a practical tool for use by experts, but empirical studies of the technique have tended to rely on students as subjects. How panelists respond to Delphi feedback will depend upon the extent of their knowledge about the topic to be forecast; this might, for example, affect their confidence in their own initial estimates and the weight they give to the feedback from anonymous panelists. One would expect experts to resist changing their estimates unless they could appreciate the value of the feedback they received (which, arguably, they could not do if feedback was simply of a statistical nature). On the other hand, consider the response of naïve subjects making judgments or forecasts

about an issue about which they have no knowledge or expertise, such as the diameter of the planet Jupiter (this is an example of an almanac question used in Delphi research). Having little basis for retaining their first-round estimate, subjects might be expected to be drawn toward the feedback statistic on subsequent rounds—arguably, an appropriate strategy, given their lack of knowledge. However, since this average would be composed of the guesses of similarly uninformed individuals, final round accuracy might be no greater than that of the first round.

The equivocal results regarding Delphi effectiveness may be traced to such factors as the varying, uncontrolled expertise of panelists. Indeed, there is some slight evidence from Delphi research that expertise does matter. Jolson and Rossow (1971) used computing corporation staff and naval personnel as subjects for separate panels and found that when these panels estimated values of almanac items in their fields, their accuracy increased over rounds, but when the items were not in their fields, their accuracy decreased. Although Riggs (1983) used students as panelists, he considered the expertise question by assessing the information or knowledge the students had about the different forecast items. He asked them to forecast the point spread of college football games and found that Delphi was a more effective instrument (i.e., it led to a greater improvement in forecasts) for a football game about which they had more information (i.e., were more knowledgeable), than for a game about which they knew relatively little.

The wider utility of expertise has been studied and reviewed elsewhere (e.g., Welty 1974, Armstrong 1985). Evidence suggests that expertise is of limited value for forecasting tasks, and that expert opinion is more useful for assessing current levels ("nowcasting") than for predicting change (forecasting) (Armstrong 1985). Delphi practitioners should take into account this wider research. Because researchers' use of naive panelists may lead them to underestimate the value of Delphi, however, we may not yet appreciate its potential as a forecasting tool.

- **Use heterogeneous experts.**

Combining the judgments of experts increases the reliability of aggregate judgments, and for this reason, statistical groups (in which the judgments of non-interacting individuals are combined) are *generally* more accurate than individuals (although they may be less so in some conditions (Stewart 2001)). When individuals interact, as in a traditional group meeting or in the structured Delphi format, the error or bias in individual judgments, deriving from incomplete knowledge or misunderstanding, may be reduced (along with unreliability). One should therefore choose experts whose combined knowledge and expertise reflects the full scope of the problem domain. Heterogeneous experts are preferable to experts focused in a single speciality.

- **Use between 5 and 20 experts.**

No firm rule governs the number of panelists to use in the Delphi procedure, although panel size clearly will have an impact on the effectiveness of the technique. While larger groups provide more intellectual resources than smaller ones, potentially bringing more knowledge and a wider range of perspectives to bear on a problem, they also make conflict, irrelevant arguments, and information overload more likely. In Delphi groups, information exchange can be controlled, making overload less of a problem than it might be in regular committees of the same size. Also, one can assemble large numbers of individuals

which would be infeasible in a regular committee. Indeed, practical applications reported in journals sometimes use panels comprising scores or even hundreds of members. But are such large panels sensible? With larger panels come greater administrative costs in terms of time and money. To maximize the use of human resources, it is desirable to limit panel sizes. The answer to the question of what is the optimal size, however, is uncertain.

Hogarth (1978) considered how such factors as group size and relative panelist knowledge might affect the validity of judgments of statistical groups. (This has relevance to Delphi, as the mathematical aggregation of panelists' estimates after each round effectively equates to the formation of a statistical group.) The specifics of his models are unimportant here, but his results suggest that groups over a certain size cease to improve in accuracy as they add further members. Armstrong (1985) suggests that groups in general should probably comprise between 5 to 20 members. The number will depend on the number of experts available, although such aspects as the nature and quality of feedback being provided (i.e., more in-depth feedback might suggest a smaller panel) should also be considered, as should cost.

Direct empirical research in the Delphi domain is limited. Brockhoff (1975) compared Delphi groups comprising five, seven, nine, and 11 panelists and found no clear distinctions in panel accuracy. Similarly, Boje and Murnighan (1982) compared the effectiveness of groups of three, seven, and 11, and found no significant differences among them.

- **For Delphi feedback, provide the mean or median estimate of the panel plus the rationales from all panelists for their estimates.**

The use of feedback in the Delphi procedure is an important feature of the technique. However, research that has compared Delphi groups to control groups in which no feedback is given to panelists (i.e., non-interacting individuals are simply asked to re-estimate their judgments or forecasts on successive rounds prior to the aggregation of their estimates) suggests that feedback is either superfluous or, worse, that it may harm judgmental performance relative to the control groups (Boje and Murnighan 1982; Parenté, et al. 1984). The feedback used in empirical studies, however, has tended to be simplistic, generally comprising means or medians alone with no arguments from panelists whose estimates fall outside the quartile ranges (the latter being recommended by the classical definition of Delphi, e.g., Rowe et al. 1991). Although Boje and Murnighan (1982) supplied some written arguments as feedback, the nature of the panelists and the experimental task probably interacted to create a difficult experimental situation in which no feedback format would have been effective.

When one restricts the exchange of information among panelists so severely and denies them the chance to explain the rationales behind their estimates, it is no surprise that feedback loses its potency (indeed, the statistical information may encourage the sort of group pressures that Delphi was designed to pre-empt). We (Rowe and Wright 1996) compared a simple iteration condition (with no feedback) to a condition involving the feedback of statistical information (means and medians) and to a condition involving the feedback of reasons (with no averages) and found that the greatest degree of improvement in accuracy over rounds occurred in the "reasons" condition. Furthermore, we found that, although subjects were less inclined to change their forecasts as a result of receiving reasons feedback than they were if they received either "statistical" feedback or no feedback at all, when "reasons" condition subjects *did* change their forecasts they tended to change towards more accurate responses. Although panelists tended to make greater changes to their

forecasts under the "iteration" and "statistical" conditions than those under the 'reasons' condition, these changes did not tend to be toward more accurate predictions. This suggests that informational influence is a less compelling force for opinion change than normative influence, but that it is a more effective force. Best (1974) has also provided some evidence that feedback of reasons (in addition to averages) can lead to more accurate judgments than feedback of averages (e.g., medians) alone.

What is the best structure for the feedback phase? In Delphi, no interaction between panelists is allowed, but in the NGT (also known as the estimate-talk-estimate procedure), verbal interaction during the assessment or evaluation phase is seen as potentially valuable in allowing panelists to clarify and justify their responses (Van de Ven and Delbecq 1971). This difference may be the only substantive one between Delphi and NGT, and studies comparing the effectiveness of the two techniques may be interpreted as studies examining the best way of allowing feedback or explanation *between* occasions when panels provide anonymous estimates. As with Delphi, the final forecast or judgment in NGT is determined by the equal weighting of the estimates of the panelists at the final round.

One might expect the NGT format to be more effective because it seems to allow a more profound discussion of differences of opinions and a greater richness in feedback quality. Comparisons of Delphi and NGT, however, show equivocal results. Although some studies show that NGT groups make more accurate judgments than comparable Delphi groups (Gustafson, et al. 1973, Van de Ven and Delbecq 1974), other studies have found no notable differences between the two techniques in the accuracy or quality of judgments (Miner 1979, Fischer 1981, Boje and Murnighan 1982), and one study has shown Delphi superiority (Erffmeyer and Lane 1984). It is possible that the act of discussing feedback may lead to an overemphasis on the opinions of those panelists who are most vocal or eloquent, and some of the difficulties associated with interacting groups may be manifest at this stage. Clearly, we need more research on the flow of influence within such structured group variants as NGT. At present, however, no compelling evidence exists that NGT improves accuracy over the standard Delphi format, and Delphi's low cost and ease of implementation (there is no need to gather one's panelists together at a single time and place) give it an advantage over NGT.

In implementing Delphi, we recommend that feedback includes arguments in addition to summary statistics. The classical definition of Delphi suggests that arguments should come only from those whose estimates lie outside the quartiles, although we found that allowing all panelists to express arguments improved the effectiveness of the Delphi technique (Rowe and Wright 1996). Because people who make similar forecasts may have different underlying reasons for doing this, and because expressing these reasons may be informative, we tentatively recommend eliciting anonymous rationales from all panelists. More research is needed to confirm this, for example, to compare the effectiveness of panels whose feedback consists of all members' arguments, to the effectiveness of panels whose feedback consists of the arguments from only the most extreme (outside quartile).

- **Continue Delphi polling until the responses show stability; generally, three structured rounds are enough.**

Researchers have devoted little attention to the value of using an unstructured first round to clarify and define the questions to be used in subsequent structured rounds. This procedure would seem valuable in allowing panelists to help specify the key issues to be addressed, rather than compelling them to answer a set of questions that they might feel

were unbalanced, incomplete, or irrelevant. Empirical studies of Delphi, however, invariably use only structured rounds, and then only two or three. What research does show is that panelists' opinions generally converge over rounds, which is reflected in a reduced variance of estimates. The practical question is, what is the optimal number of structured rounds? There is no definitive answer to this: the accepted criterion is when responses show stability, and it is up to the facilitator to decide when to call the procedure to a halt. Stability does not necessarily equate to complete convergence (zero variance), however, as panelists might, over successive rounds, settle for their own estimates and refuse to shift further toward the average position. Indeed, if panelists have fundamental bases for settling upon their divergent forecasts, it would be a mistake to conduct additional rounds in the hope of forcing consensus.

Erffmeyer, Erffmeyer and Lane (1986) found that the quality of Delphi estimates increased up to the fourth round but not thereafter. Brockhoff (1975) found that the accuracy of estimates increased up to round three, but then decreased. Other studies using two to three structured rounds have also shown accuracy improvement over rounds (Rohrbaugh 1979 Rowe and Wright 1996). Other researchers simply report the final round Delphi aggregate and not the aggregate of prior rounds or else do not specify the number of rounds used (e.g. Miner 1979) and hence provide no insight into this issue.

From this limited evidence, we suggest that three structured rounds is sufficient in Delphi, although practical considerations are relevant. If after the third round responses still show a high degree of variability, the facilitator could hold further rounds to see if unresolved issues might be clarified. Panelists, however, tend to drop out after each round (Bardecki 1984), so a high number of rounds might lead to a high drop-out rate. If those who drop out are the worst panelists, accuracy should improve, but they might simply be the busiest or most impatient. This is an empirical question that needs answering.

- **Obtain the final forecast by weighting all the experts' estimates equally and aggregating them.**

The forecast from a Delphi procedure is taken to be the average of the anonymous forecasts made by all panelists on the final round. (Because extreme values can distort means, it may be best to use median or a trimmed mean that excludes these extreme values. Selecting appropriate experts should, however, reduce the occurrence of extreme values.) This is equivalent to the average of the equally weighted estimates of the members of a statistical group. It is possible, however, to weight panelists' estimates differentially, and this would make sense if one knew which panelists were best at the task. The issue of unequal-weighting has not been directly researched in Delphi studies, although Larreché and Moinpour (1983) demonstrated that one could achieve better accuracy in an estimation task by aggregating only the estimates of those identified as most expert according to an external measure of expertise (but not when expertise was assessed according to panelists' confidence estimates). Best (1974) found that subgroups of experts—determined by self-rating—were more accurate than subgroups of non-experts. In these studies, the researchers effectively gave experts a weighting of one and non-experts a weighting of zero, although weighting does not have to be all or nothing.

The central problem in variable weighting of the judgments of experts is determining how to weight them. In forecasting tasks, objective measures of expertise are unlikely to be available, unless the task is repetitive with detailed records of past performance, such as for weather forecasts. Generally, there will not be enough appropriate data to adequately

rate all panelists, perhaps because their experiences are non-comparable, or because the current problem is subtly different from past problems, or because no objective measurements of past performance exist. Indeed, even if these criteria were satisfied, learning might have taken place since the most recent assessment (Lock 1987), or good past performance may have been due to chance. In any case, situations prone to objective measurement are likely to be situations in which the objective data can be used in econometric or extrapolative models. Those approaches might be preferable because they do not rely on any subjective components (Armstrong 1985). Weighting schemes based on something other than objective data, such as panelist ratings of their own confidence or expertise, have not generally been shown to be valid indicators of expertise in judgment and forecasting tasks. For example, although Best (1974) and Rowe and Wright (1996) seemed to find that self-ratings can have some validity, other studies have found no relationship between self-ratings and objective expertise (e.g., in Delphi research, Brockhoff, 1975; Larreché and Moinpour 1983; Dietz 1987, Sniezek 1990). Identifying expertise is a bottleneck in applying differential weighting in mathematical aggregation. (This principle is similar to Dawes', 1982, findings on the weighting of *information;* the equal weighting of variables in linear models is a strategy that is difficult to better for a variety of reasons.)

- **In phrasing questions, use clear and succinct definitions and avoid emotive terms.**

How a question is worded can lead to significant response biases. By changing words or emphasis, one can induce respondents to give dramatically different answers to a question. For example, Hauser (1975) describes a 1940 survey in which 96 percent of people answered yes to the question "do you believe in freedom of speech?" and yet only 22 percent answered yes to the question "do you believe in freedom of speech to the extent of allowing radicals to hold meetings and express their views to the community?" The second question is consistent with the first; it simply entails a fuller definition of the concept of freedom of speech. One might therefore ask which of these answers more clearly reflects the views of the sample. Arguably, the more apt representation comes from the question that includes a clearer *definition* of the concept of interest, because this should ensure that the respondents are all *answering the same question.* Researchers on Delphi per se have shown little empirical interest in question wording. Salancik, Wenger and Heifer (1971) provide the only example of which we are aware; they studied the effect of question length on initial panelist consensus and found that one could apparently obtain greater consensus by using questions that were neither "too short" nor "too long." This is a generally accepted principle for wording items on surveys: they should be long enough to define the question adequately so that respondents do not interpret it differently, yet they should not be so long and complicated that they result in information overload, or so precisely define a problem that they demand a particular answer. Also, questions should not contain emotive words or phrases: the use of the term "radicals" in the second version of the freedom-of-speech question, with its potentially negative connotations, might lead to emotional rather than reasoned responses.

- **Frame questions in a balanced manner.**

Tversky and Kahneman (1974, 1981) provide a second example of the way in which question framing may bias responses. They posed a hypothetical situation to subjects in which human lives would be lost: if subjects were to choose one option, a certain number

of people would *definitely* die, but if they chose a second option, then there was a *probability* that more would die, but also a chance that less would die. Tversky and Kahneman found that the proportion of subjects choosing each of the two options changed when they phrased the options in terms of people surviving instead of in terms of dying (i.e., subjects responded differently to an option worded "60 percent will survive" than to one worded "40 percent will die," even though these are logically identical statements). The best way to phrase such questions might be to clearly state both death and survival rates (balanced), rather than leave half of the consequences implicit. Phrasing a question in terms of a single perspective, or numerical figure, may provide an anchor point as the focus of attention, so biasing responses.

- **Avoid incorporating irrelevant information into questions.**

In another study, Tversky and Kahneman (1974) presented subjects with a description or personality sketch of a hypothetical student, "Tom W." They asked the subjects to choose from a number of academic fields that field in which Tom was most likely to be a student. They found that subjects tended to ignore information about base rates (i.e., the relative numbers of students in the various fields) and instead focused on the personality information. Essentially, because Tom W. was "intelligent, although lacking in true creativity" and had a need for "order and clarity" he was seen as more likely to be, for example, an engineering student than a social science student, even though the statistical likelihood might be for the opposite option. We will not explain all possible reasons for this effect here. One possibility, however, is that subjects may see irrelevant information in a question or statement as relevant because it is included, and such information should therefore be avoided. Armstrong (1985) suggests that no information is better than worthless information. Payne (1951), Noelle-Neuman (1970), and Sudman and Bradburn (1983) also give practical advice on wording questions.

- **When possible, give estimates of uncertainty as frequencies rather than probabilities or odds.**

Many applications of Delphi require panelists to make either numerical estimates of the probability of an event happening in a specified time period, or to assess their confidence in the accuracy of their predictions. Researchers on behavioral decision making have examined the adequacy of such numerical judgments. Results from these findings, summarized by Goodwin and Wright (1998), show that sometimes judgments from *direct* assessments (what is the probability that…?) are inconsistent with those from *indirect* methods. In one example of an indirect method, subjects might be asked to imagine an urn filled with 1,000 colored balls (say, 400 red and 600 blue). They would then be asked to choose between betting on the event in question happening, or betting on a red ball being drawn from the urn (both bets offering the same reward). The ratio of red to blue balls would then be varied until a subject was *indifferent* between the two bets, at which point the required probability could be inferred. Indirect methods of eliciting subjective probabilities have the advantage that subjects do not have to verbalize numerical probabilities. Direct estimates of *odds* (such as 25 to 1, or 1,000 to 1), perhaps because they have no upper or lower limit, tend to be more extreme than direct estimates of *probabilities* (which must lie between zero and one). If probability estimates derived by different methods for the same event are inconsistent, which method should one take as the true index of degree of belief? One way

to answer this question is to use a single method of assessment that provides the most consistent results in repeated trials. In other words, the subjective probabilities provided at different times by a single assessor for the same event should show a high degree of agreement, given that the assessor's knowledge of the event is unchanged. Unfortunately, little research has been done on this important problem. Beach and Phillips (1967) evaluated the results of several studies using direct estimation methods. Test-retest correlations were all above 0.88, except for one study using students assessing odds, where the reliability was 0.66.

Gigerenzer (1994) provided empirical evidence that the untrained mind is not equipped to reason about uncertainty using subjective probabilities but is able to reason successfully about uncertainty using frequencies. Consider a gambler betting on the spin of a roulette wheel. If the wheel has stopped on red for the last 10 spins, the gambler may feel subjectively that it has a greater probability of stopping on black on the next spin than on red. However, ask the same gambler the relative frequency of red to black on spins of the wheel and he or she may well answer 50-50. Since the roulette ball has no memory, it follows that for each spin of the wheel, the gambler should use the latter, relative frequency assessment (50-50) in betting. Kahneman and Lovallo (1993) have argued that forecasters tend to see forecasting problems as unique when they should think of them as instances of a broader class of events. They claim that people's natural tendency in thinking about a particular issue, such as the likely success of a new business venture, is to take an "inside" rather than an "outside" view. Forecasters tend to pay particular attention to the distinguishing features of the particular event to be forecast (e.g., the personal characteristics of the entrepreneur) and reject analogies to other instances of the same general type as superficial. Kahneman and Lovallo cite a study by Cooper, Woo, and Dunkelberger (1988), which showed that 80 percent of entrepreneurs who were interviewed about their chances of business success described this as 70 percent or better, while the overall survival rate for new business is as low as 33 percent. Gigerenzer's advice, in this context, would be to ask the individual entrepreneurs to estimate the proportion of new businesses that survive (as they might make accurate estimates of this relative frequency) and use this as an estimate of their own businesses surviving. Research has shown that such interventions to change the required response mode from subjective probability to relative frequency improve the predictive accuracy of elicited judgments. For example, Sniezek and Buckley (1991) gave students a series of general knowledge questions with two alternative answers for each, one of which was correct. They asked students to select the answer they thought was correct and then estimate the probability that it was correct. Their results showed the same general overconfidence that Arkes (2001) discusses. However, when Sniezek and Buckley asked respondents to state how many of the questions they had answered correctly of the total number of questions, their frequency estimates were accurate. This was despite the fact that the same individuals were generally overconfident in their subjective probability assessments for individual questions. Goodwin and Wright (1998) discuss the usefulness of distinguishing between single-event probabilities and frequencies. If a reference class of historic frequencies is not obvious, perhaps because the event to be forecast is truly unique, then the only way to assess the likelihood of the event is to use a subjective probability produced by judgmental heuristics. Such heuristics can lead to judgmental overconfidence, as Arkes (2001) documents.

- **Use coherence checks when eliciting estimates of probabilities.**

Assessed probabilities are sometimes incoherent. One useful *coherence* check is to elicit from the forecaster not only the probability (or confidence) that an event will occur, but also the probability that it will not occur. The two probabilities should sum to one. A variant of this technique is to *decompose* the probability of the event not occurring into the occurrence of other possible events. If the events are mutually exclusive and exhaustive, then the addition rule can be applied, since the sum of the assessed probabilities should be one. Wright and Whalley (1983) found that most untrained probability assessors followed the additivity axiom in simple two-outcome assessments involving the probabilities of an event happening and not happening. However, as the number of mutually exclusive and exhaustive events in a set increased, more forecasters became supra-additive, and to a greater extent, in that their assessed probabilities added up to more than one. Other coherence checks can be used when events are interdependent (Goodwin and Wright 1998; Wright, et al. 1994).

There is a debate in the literature as to whether decomposing analytically complex assessments into analytically more simple marginal and conditional assessments of probability is worthwhile as a means of simplifying the assessment task. This debate is currently unresolved (Wright, Saunders and Ayton 1988; Wright et al. 1994). Our view is that the best solution to problems of inconsistency and incoherence in probability assessment is for the pollster to show forecasters the results of such checks and then allow interactive resolution between them of departures from consistency and coherence. MacGregor (2001) concludes his review of decomposition approaches with similar advice.

When assessing probability distributions (e.g., for the forecast range within which an uncertainty quality will lie), individuals tend to be *overconfident* in that they forecast too narrow a range. Some response modes fail to counteract this tendency. For example, if one asks a forecaster initially for the median value of the distribution (the value the forecaster perceives as having a 50 percent chance of being exceeded), this can act as an anchor. Tversky and Kahneman (1974) were the first to show that people are unlikely to make sufficient adjustments from this anchor when assessing other values in the distribution. To counter this bias, Goodwin and Wright (1998) describe the "probability method" for eliciting probability distributions, an assessment method that de-emphasizes the use of the median as a response anchor. McClelland and Bolger (1994) discuss overconfidence in the assessment of probability distributions and point probabilities. Wright and Ayton (1994) provide a general overview of psychological research on subjective probability. Arkes (2001) lists a number of principles to help forecasters to counteract overconfidence.

CONDITIONS FOR THE USE OF DELPHI

Delphi can be used to elicit and combine expert opinions under the following conditions:

- **When expert judgment is necessary because the use of statistical methods is inappropriate.**

Research shows that human judgment compares poorly to the output of statistical and computational models that are based on the same data. For example, linear models that

ascribe weights to predictor variables and then sum these to arrive at a value for a criterion variable (the event being judged or forecast) have been shown to be more accurate than people estimating the criterion according to their own judgment (Meehl 1954). In essence, people are inconsistent in their judgments and unable to deal with large amounts of data and to combine information (Stewart 2001). Evidence suggests using statistical techniques whenever this is feasible.

In many forecasting situations, however, the use of statistical models is either impractical or impossible. This may be because obtaining historical or economic or technical data is costly or impossible. Even when such data exist, one must be sure that future events will not make the historical data unusable. With little information, one must rely on opinion, and Delphi is a useful method for eliciting and aggregating expert opinion.

- **When a number of experts are available.**

When a forecasting situation requires the use of human judgment and several experts are available, one must then decide which experts to use *(who* and *how many)* and how to use them. Delphi requires a number of experts; if research showed that individuals generally forecast as well as (or better than) several experts combined, we would not recommend using Delphi or any other approach requiring multiple experts. Research suggests, however, that traditional and statistical groups tend to outperform individuals in a variety of judgmental tasks (Hill 1982). Groups possess at least as much knowledge as any one of their members, while traditional interacting groups provide the opportunity for the debiasing of faulty opinions and the synthesis of views. Therefore, when a number of experts are available, research suggests that we *should* use several experts, and Delphi might be appropriate for eliciting and combining their opinions.

- **When the alternative is simply to average the forecasts of several individuals.**

When a forecasting task must rely on judgment and numerous experts are available, the individuals and their forecasts may be combined in several ways. In the most straightforward, individuals give their forecasts without interacting, and these forecasts are weighted equally and statistically aggregated. Researchers have compared the accuracy of such statistical groups to Delphi groups in two ways: through a straightforward comparison of the two approaches, and through a comparison of the quality of averaged estimates on the first round and on the final round in a Delphi procedure. The first, pre-interaction round is equivalent to a statistical group in every way except for the instructions given to individuals: Delphi panelists are led to expect further polling and feedback from others, which may lead panelists to consider the problem more deeply and possibly to make better "statistical group" judgments on that first round than individuals who do not expect to have their estimates used as feedback for others. A first-round Delphi may, however, provide a better benchmark for comparison than a separate statistical group, because the panelists in the two "conditions" are the same, reducing a potential source of great variance.

We (Rowe and Wright 1999) have reviewed the evidence for the relative values of statistical groups and Delphi groups. Although it should be possible to compare averages over rounds in every study of Delphi accuracy or quality, researchers in a number of evaluative studies do not report the differences between rounds (e.g., Fischer 1981, Riggs 1983). Nevertheless, we found that results generally support the advantage of Delphi groups over first-round or statistical groups by a tally of 12 studies to two. In five studies, the research-

ers reported significant increases in accuracy over Delphi rounds (Best 1974; Larreché and Moinpour 1983; Erffmeyer and Lane 1984; Erffmeyer, Erfrmeyer and Lane 1986; Rowe and Wright 1996), although in their two papers, Erffmeyer and colleagues may have been reporting separate analyses on the same data (this is not clear). Seven more studies produced qualified support for Delphi: in five cases, researchers found Delphi to be better than statistical or first-round groups more often than not, or to a degree that did not reach statistical significance (Dalkey, Brown and Cochran 1970; Brockhoff 1975; Rohrbaugh 1979; Dietz 1987; Sniezek 1989), and in two others, researchers found Delphi to be better under certain conditions and not others: Parenté et al (1984) found that Delphi accuracy increased over rounds for predicting "when" an event might occur, but not "if" it would occur; Jolson and Rossow (1971) found that accuracy increased for panels comprising "experts," but not for "non-experts."

In contrast, researchers in only two studies found no substantial difference in accuracy between Delphi and statistical groups (Fischer 1981 and Sniezek 1990—although Sniezek's panelists had common information; hence there could be no basis for Delphi improvements), and researchers in two studies found that Delphi accuracy was worse. Gustafson et al. (1973) found that Delphi groups were less accurate than both their first-round aggregates (for seven out of eight items) and independent statistical groups (for six out of eight items), while Boje and Murnighan (1982) found that Delphi panels became less accurate over rounds for three out of four items. The weight of this evidence, however, suggests that Delphi groups should be used instead of statistical groups when feasible, because evidence generally shows that they lead to more accurate judgments. Intuitively, this is what we would expect, given the additional interaction that takes place during Delphi *following* the averaging of first-round estimates.

- **When the alternative is a traditional group.**

A more common manner of using multiple experts is in a traditional group meeting. Unfortunately, a variety of social, psychological, and political difficulties may arise during group meetings that can hinder effective communication and behavior. Indeed, Delphi was designed to improve upon the traditional group by adding structure to the process. Results generally suggest that Delphi groups are more accurate than traditional groups. In a review of the literature, we found that Delphi groups outperformed traditional groups by a score of five studies to one, with two ties, and with one study showing task-specific support for both techniques (Rowe and Wright 1999). Support for Delphi comes from Van de Ven and Delbecq (1974), Riggs (1983), Larreché and Moinpour (1983), Erffmeyer and Lane (1984), and Sniezek (1989). Fischer (1981) and Sniezek (1990) found no distinguishable differences in accuracy between the two approaches (although Sniezek's subjects had common information), while Gustafson et al. (1973) found a small advantage for interacting groups. Brockhoff (1975) seemed to show that the nature of the task is important, with Delphi being more accurate with almanac items, but less accurate with forecasting items (although the difference might reflect task difficulty as much as content).

These studies, seem to show that collections of individuals make more accurate judgments and forecasts in Delphi groups than in unstructured groups, and that Delphi should be used in preference. One point of caution, however, is that the groups used in Delphi studies are usually highly simplified versions of real-world groups; the latter comprise individuals with a high degree of expertise on the problem topic who genuinely care about the result of their meeting and have some knowledge of the strengths and weaknesses of

their colleagues (or think they do) on which basis they might be able to selectively accept or reject their opinions. It may be that in a richer environment, the extra information and motivation brought to a task by those in a traditional group may make it of greater value than the limiting Delphi procedure. But this is conjecture and does not cause us to reverse our recommendation based on evidence.

Delphi has also been compared to other procedures that add some structure to the group process. Some of these can be considered formal procedures, while others are experimental variants that might form the basis of distinct techniques in the future. Delphi has been compared to groups whose members were required to argue both for and against their judgments (the 'dialectic' procedure [Sniezek 1989]); groups whose judgments were derived from a single, group-selected individual (the 'dictator' or 'best member' strategy (Sniezek 1989, 1990)); groups that received rules on how to interact appropriately (Erffmeyer and Lane 1984); groups whose information exchange was structured according to social judgment analysis (Rohrbaugh 1979); and groups following a problem-centered leadership (PCL) approach (Miner 1979). The only studies that revealed any substantial differences between Delphi and the comparison procedures are those of Erffmeyer and Lane (1984), which showed Delphi to be more effective than groups given instructions on resolving conflict, and Miner (1979), which showed that the PCL approach (which involves instructing group leaders in appropriate group-directing skills) to be significantly more effective than Delphi ("effectiveness" here being a measure comprising the product of measures of "quality" and "acceptance"). Given the equivocal nature of the results of these studies, we will not belabor their details here. On the basis of this limited evidence, however, there appears to be no clear rationale for adopting any of these techniques in preference to Delphi.

IMPLICATIONS FOR PRACTITIONERS

In the Principles section, we discussed how best to conduct a Delphi procedure, and in the Conditions section we discussed those situations in which Delphi might be useful. The practitioner should consider other factors, however, before deciding to implement a Delphi group. We do not describe these factors as principles or conditions because they generally relate to opinions and are not supported by evidence.

The possible utility of Delphi is increased in a number of situations. When experts are geographically dispersed and unable to meet in a group, Delphi would seem an appropriate procedure. It would enable members of different organizations to address industry-wide problems or forecasts, or experts from different facilities within a single organization to consider a problem without traveling to a single location. Indeed, experts with diverse backgrounds who have no history of shared communication are liable to have different perspectives, terminologies, and frames of reference, which might easily hinder effective communication in a traditional group. Such difficulties could be ironed out by the facilitator or monitor team before the structured rounds of a Delphi.

Delphi might also be appropriate when disagreements between individuals are likely to be severe or politically unpalatable. Under such circumstances, the quality of judgments and decisions is likely to suffer from motive conflicts, personality clashes, and power games. Refereeing the group process and ensuring anonymity should prove beneficial.

Finally, the practitioner should be aware of the expense of conducting a Delphi exercise compared to the alternatives. Expenses to be considered include the cost of employing a facilitator or monitor team (or the time required if the Delphi is done in-house), the price of materials and postage, and the delay in obtaining a forecast (because of the time taken in polling and collating results). It should be possible to automate Delphi to some extent, perhaps conducting it electronically through the use of e-mail, the internet, or electronic conference sites, and this would require different costs, skills, and resources. These considerations are not negligible: although research generally shows that Delphi groups outperform statistical and traditional groups, differences in the quality of estimates and forecasts are not always high, and the gain in response quality from a Delphi panel may be outweighed by the time and expense needed to conduct the procedure. For important forecasts where even small improvements in accuracy are valuable, one has greater incentive to use Delphi.

IMPLICATIONS FOR RESEARCHERS

The literature contains hundreds of papers on Delphi procedures, but most concern *applications* in which Delphi is used as a tool for aggregating expert judgments and which focus on the final judgment or forecast. Accounts of experimental evaluations of the technique are scarce, and even these have been criticized. Much of the criticism of the early evaluative studies (for example, those carried out at the RAND Corporation) centered on their "sloppy execution" (e.g., Stewart 1987). Among specific criticisms are claims that Delphi questionnaires tended to be poorly worded and ambiguous (Hill and Fowles 1975) and that the analysis of responses was often superficial (Linstone 1975). Explanations for the poor conduct of early studies have ranged from the technique's apparent simplicity encouraging people without the requisite skills to use it (Linstone and Turoff 1975) to suggestions that the early Delphi researchers had poor backgrounds in the social sciences and hence lacked acquaintance with appropriate research methodologies (Sackman 1975). Although more recent research has generally been conducted by social scientists using standard experimental procedures, little evidence has accumulated regarding how best to conduct Delphi and when to use it. We have relied on the findings of these recent studies to formulate tentative principles and conditions, but the topic requires more concerted and disciplined study.

We believe that recent research has been somewhat misdirected, with too much emphasis on "Technique-Comparison" studies at the expense of "Process" studies (Rowe et al. 1991, Rowe and Wright 1999). Studies of the former type tend to compare Delphi to other procedures to answer the question "is Delphi (relatively) good or bad?", while studies of the latter type ask "why is Delphi good or bad?" Because the answer to the first question is generally "it depends…", and because researchers asking this question tend to show little concern for the factors on which effectiveness depends, we are left little the wiser. This lack of control of mediating factors has generally been associated with the use of simplified versions of Delphi that vary from the technique ideal in ways that might be expected to decrease effectiveness. For example, researchers performing evaluative studies generally use naive subjects (students) instead of experts, use artificial tasks (e.g., estimating almanac questions) instead of meaningful ones, and provide only limited feedback (means or

medians) instead of rationales. Indeed, one might argue that the kinds of techniques researchers use in some of these studies are barely Delphis at all. Using simplified versions of the technique is not always wrong; indeed, it is appropriate when conducting controlled experiments aimed at understanding basic processes within Delphi. But using simplified versions in studies aimed at comparing Delphi to other procedures is akin to holding a race to see whether dogs are faster than cats and then using a Pekinese to represent the dogs instead of a greyhound. To truly understand Delphi, we need to focus on what it is about Delphi that makes it work, and consequently, how we should ideally specify Delphi (so that we can identify the greyhound!). We need controlled studies on the influences of feedback, panel compositions and sizes, and tasks.

With regard to understanding structured group processes, we particularly need to discover which panelists *change* their estimates over rounds (for this determines whether panels become more or less accurate), and what it is about the technique and task circumstances that encourage them to do so. This will enable us to determine what facets of Delphi help panelists improve their judgments and what do not, with implications for the principles of conducting Delphi.

Few studies have focused on understanding how panelists' judgments change. One theory is that the improvement in accuracy over Delphi rounds comes about because the more-expert panelists (the hold outs) maintain their judgments over rounds, while the less-expert panelists (the swingers) alter their judgments towards the group average (Parenté and Anderson-Parenté 1987). If this occurs, it can be shown that the group average will move towards the average of the expert subset over rounds and hence towards the true answer. We have produced some evidence supporting this theory, finding that the more-accurate Delphi panelists on the first round (the more expert) changed their estimates *less* over subsequent rounds than did the less-accurate (less expert) panelists, so that the average group value shifted towards that of the more accurate panelists with a corresponding increase in group accuracy (Rowe and Wright 1996).

Other theories can be constructed to explain opinion change during the Delphi process, however, and these might describe the empirical data better than the above theory. For example, a confidence theory might predict that it is the least-confident individuals who change their estimates the most over rounds, rather than the least expert. This would suggest that when confidence is appropriate (when it correlates with objective expertise), Delphi would lead to more accurate judgment, and when it is not, judgment quality would decline. (Regarding this hypothesis, Scheibe, Skutsch and Schofer [1975] found a positive relationship between high confidence and low change, but Rowe and Wright [1996] found no evidence for this.) If this theory had any validity, it would have implications for the selection of Delphi panelists.

Future research should focus on formulating competing theories and determining empirically which fits observations best. Researchers should also recognize the complexity of Delphi-task interactions, and pay more attention to possible mediating variables related to the nature of the panelists, the precise nature of the task, and the characteristics of the technique.

SUMMARY

When human judgment is required in forecasting situations, the key issue is how best to elicit and use expert opinion. Judgments derived from multiple experts—that is, from groups—are generally more accurate than those of individual experts. However, group processes often lead to suboptimal judgments, and one solution to this is to structure the interaction of experts using such approaches as Delphi. We have distilled the following principles for using expert opinion, which have implications for defining best practice in the design and application of structured groups:

- Use experts with appropriate domain knowledge.
- Use heterogenous experts.
- Use between five and 20 experts.
- For Delphi feedback, provide the mean or median estimate of the panel plus the rationales from all panelists for their estimates.
- Continue Delphi polling until the responses show stability. Generally, three structured rounds is enough.
- Obtain the final forecast by weighting all the experts' estimates equally and aggregating them.
- In phrasing questions, use clear and succinct definitions and avoid emotive terms.
- Frame questions in a balanced manner.
- Avoid incorporating irrelevant information into questions.
- When possible, give estimates of uncertainty as frequencies rather than probabilities or odds.
- Use coherence checks when eliciting estimates of probabilities.

In spite of the inconsistent application of these principles in empirical examples of Delphi, research has shown that Delphi-like groups perform judgmental and forecasting tasks more effectively than other judgmental approaches. Studies support the advantage of Delphi over traditional groups (in terms of increased accuracy) by five to one with one tie, and its advantage over statistical groups by 12 to two with two ties. More consistent application of the above principles may lead to better performance of structured groups in the future.

REFERENCES

Arkes, H. (2001), "Overconfidence in judgmental forecasting," in J. .S. Armstrong (ed.), *Principles of Forecasting.* Norwell, MA.: Kluwer Academic Publishers.

Armstrong, J. S. (1985), *Long Range Forecasting: From Crystal Ball to Computer,* 2nd ed., New York: Wiley. (Full text at http://hops.wharton.upenn.edu/forecast.)

Bardecki, M.J. (1984), "Participants' response to the Delphi method: An attitudinal perspective," *Technological Forecasting and Social Change,* 25, 281-292.

Beach, L. R. & L. D. Phillips (1967), "Subjective probabilities inferred from estimates and *bets," Journal of Experimental Psychology,* 75, 354–259.

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

Boje, D. M. & 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 & M. Turoff (eds.), *The Delphi Method: Techniques and Applications.* London: Addison-Wesley.

Cooper, A., C. Woo & W. Dunkelberger (1988), "Entrepreneurs perceived chances of success," *Journal of Business Venturing,* 3, 97–108.

Dalkey, N.C., B. Brown & S. W. Cochran (1970), "The Delphi Method III: Use of self-ratings to improve group estimates," *Technological Forecasting,* 1, 283–291.

Dawes, R. M. (1982), "The robust beauty of improper linear models in decision making," in D. Kahneman, P. Slovic & A. Tversky (eds.), *Judgement Under Uncertainty: Heuristics and Biases.* Cambridge: Cambridge University Press.

Dietz, T. (1987), "Methods for analyzing data from Delphi panels: Some evidence from a forecasting study," *Technological Forecasting and Social Change,* 31, 79-85.

Erffmeyer, R. C., E. S. Erffmeyer & I. M. Lane (1986), "The Delphi technique: An empirical evaluation of the optimal number of rounds," *Group and Organization Studies,* 11, 120-128.

Erffmeyer, R. C. & I .M. Lane (1984), "Quality and acceptance of an evaluative task: The effects of four group decision-making formats," *Group and Organization Studies,* 9, 509-529.

Fischer, G. W. (1981), "When oracles fail—a comparison of four procedures for aggregating subjective probability forecasts," *Organizational Behavior and Human Performance,* 28, 96–110.

Gigerenzer, G. (1994), "Why the distinction between single event probabilities and frequencies is important for psychology (and vice-versa)," in G. Wright and P. Ayton (eds.), *Subjective Probability.* Chichester, U.K.: Wiley.

Goodwin, P. & G. Wright (1998), *Decision Analysis for Management Judgment,* 2nd ed. Chichester, U.K.: Wiley.

Gustafson, D. H., R. K. Shukla, A. Delbecq & G. W. Walster (1973), "A comparison study of differences in subjective likelihood estimates made by individuals, interacting groups, Delphi groups and nominal groups," *Organizational Behavior and Human Performance,* 9, 280–291.

Mauser, P.M. (1975), *Social Statistics in Use.* New York: Russell Sage.

Hill, G. W. (1982), "Group versus individual performance: Are N+1 heads better than one?" *Psychological Bulletin,* 91, 517–539.

Hill, K. Q. & J. Fowles (1975), "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.

Jolson, M. A. & G. Rossow (1971), "The Delphi process in marketing decision making," *Journal of Marketing Research,* 8, 443–448.

Kahneman, D. & D. Lovallo (1993), "Timid choices and bold forecasts: A cognitive perspective on risk taking," *Management Science,* 39, 17–31.

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

Linstone, H. A. (1975), "Eight basic pitfalls: A checklist," in H. Linstone and M. Turoff (eds.), *The Delphi Method: Techniques and Applications.* London: Addision-Wesley.

Linstone, H. A. & M. Turoff (1975), *The Delphi Method: Techniques and Applications.* London: Addision-Wesley.

Lock, A. (1987), "Integrating group judgments in subjective forecasts," in G. Wright and P. Ayton (eds.), *Judgmental Forecasting.* Chichester, U.K.: Wiley.

MacGregor, D. G. (2001), "Decomposition for judgmental forecasting and estimation," in J. S. Armstrong (ed.), *Principles of Forecasting.* Norwell, MA.: Kluwer Academic Publishers.

Martino, J. (1983), *Technological Forecasting for Decision Making,* (2nd ed.). New York: American Elsevier.

McClelland, A. G. R. & F. Bolger (1994), "The calibration of subjective probabilities: Theories and models 1980–1994," in G. Wright and P. Ayton (eds.), *Subjective Probability.* Chichester, U.K.: Wiley.

Meehl, P. E. (1954), *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence.* Minneapolis: University of Minnesota Press.

Miner, F. C. (1979), "A comparative analysis of three diverse group decision making approaches," *Academy of Management Journal,* 22, 81–93.

Noelle-Neuman, E. (1970), "Wanted: Rules for wording structured questionnaires," *Public Opinion Quarterly,* 34, 90–201.

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

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

Payne, S.L. (1951), *The Art of Asking Questions.* Princeton, NJ: Princeton University Press.

Riggs, W. E. (1983), "The Delphi method: An experimental evaluation," *Technological Forecasting and Social Change,* 23, 89–94.

Rohrbaugh, J. (1979), "Improving the quality of group judgment: Social judgment analysis and the Delphi technique," *Organizational Behavior and Human Performance,* 24, 73–92.

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

Rowe, G. & G. Wright (1999), "The Delphi technique as a forecasting tool: Issues and analysis," *International Journal of Forecasting,* 15, 353-375. (Commentary follows on pp. 377-381.)

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

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

Salancik, J. R., W. Wenger & E. Helfer (1971), "The construction of Delphi event statements," *Technological Forecasting and Social Change,* 3, 65–73.

Scheibe, M., M. Skutsch & J. Schofer (1975), "Experiments in Delphi methodology," in H. Linstone and M. Turoff (eds.), *The Delphi Method: Techniques and Applications.* London: Addison-Wesley.

Sniezek, J. A. (1989), "An examination of group process in judgmental forecasting," *International Journal of Forecasting,* 5, 171–178.

Sniezek, J. A. (1990), "A comparison of techniques for judgmental forecasting by groups with common information," *Group and Organization Studies,* 15, 5–19.

Sniezek, J. A. & T. Buckley (1991), "Confidence depends on level of aggregation," *Journal of Behavioral Decision Making,* 4, 263–272.

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

Stewart, T.R. (2001), "Improving reliability in judgmental forecasts," in J. S. Armstrong (ed.), *Principles of Forecasting.* Norwell, MA.: Kluwer Academic Publishers.

Sudman, S. & N. Bradburn (1982), *Asking Questions.* San Francisco: Josey-Bass.

Tversky, A. & D. Kahneman (1974), "Judgment under uncertainty: Heuristics and biases," *Science,* 185, 1124–1131.

Tversky, A. & D. Kahneman (1981), "The framing of decisions and the psychology of choice," *Science,* 211, 453–458.

Van de Ven, A. H. & A. L. Delbecq (1971), "Nominal versus interacting group processes for committee decision making effectiveness," *Academic Management Journal,* 14, 203–213.

Van de Ven, A. H. & A. L. Delbecq (1974), "The effectiveness of nominal, Delphi, and interacting group decision making processes," *Academy of Management Journal,* 17, 605–621.

Welty, G. (1974), "The necessity, sufficiency and desirability of experts as value forecasters," in W. Leinfellner and E. Kohler (eds.), *Developments in the Methodology of Social Science.* Boston: Reidel.

Wright, G. & P. Ayton (1994), *Subjective Probability.* Chichester, U.K.: Wiley.

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

Wright, G., C. Saunders & P. Ayton (1988), "The consistency, coherence and calibration of holistic, decomposed and recomposed judgmental probability forecasts," *Journal of Forecasting,* 7, 185–199.

Wright, G. & P. Whalley (1983), "The supra-additivity of subjective probability," in B. Stigum & F. Wenstop (eds.), *Foundations of Risk and Utility Theory with Applications.* Dordrecht: Reidel.