ELSEVIER

International Journal of Forecasting 15 (1999) 353–375

international journal of forecasting

www.elsevier.com/locate/ijforecast

The Delphi technique as a forecasting tool: issues and analysis

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

[a]*Institute of Food Research, Norwich Research Park, Colney, Norwich NR4 7UA, UK*
[b]*Strathclyde Graduate Business School, Strathclyde University, 199 Cathedral Street, Glasgow G4 0QU, UK*

Abstract

This paper systematically reviews empirical studies looking at the effectiveness of the Delphi technique, and provides a critique of this research. Findings suggest that Delphi groups outperform statistical groups (by 12 studies to two with two 'ties') and standard interacting groups (by five studies to one with two 'ties'), although there is no consistent evidence that the technique outperforms other structured group procedures. However, important differences exist between the typical laboratory version of the technique and the original concept of Delphi, which make generalisations about 'Delphi' per se difficult. These differences derive from a lack of control of important group, task, and technique characteristics (such as the relative level of panellist expertise and the nature of feedback used). Indeed, there are theoretical and empirical reasons to believe that a Delphi conducted according to 'ideal' specifications might perform *better* than the standard laboratory interpretations. It is concluded that a different focus of research is required to answer questions on Delphi effectiveness, focusing on an analysis of the *process* of judgment change within nominal groups.

Keywords: Delphi; Interacting groups; Statistical group; Structured groups

1. Introduction

Since its design at the RAND Corporation over 40 years ago, the Delphi technique has become a widely used tool for measuring and aiding forecasting and decision making in a variety of disciplines. But what do we really understand about the technique and its workings, and indeed, how is it being employed? In this paper we adopt the perspective of Delphi as a judgment or forecasting or decision-aiding tool, and we review the studies that have attempted to evaluate it. These studies are actually sparse, and, we will argue, their attempts at evaluating the technique have been largely *inappropriate*, such that our knowledge about the potential of Delphi is still poor. In essence, this paper relates a critique of the methodology of evaluation research and suggests that we will acquire no great knowledge of the potential benefits of Delphi until we adopt a different methodological approach, which is subsequently detailed.

2. The nature of Delphi

The Delphi technique has been comprehensively reviewed elsewhere (e.g., Hill & Fowles, 1975; Linstone & Turoff, 1975; Lock, 1987; Parenté & Anderson-Parenté, 1987; Stewart, 1987; Rowe, Wright & Bolger, 1991), and so we will present a

*Corresponding author. Tel.: +44-1603-255-125.
E-mail address: gene.rowe@bbsrc.ac.uk (G. Rowe)

0169-2070/99/$ – see front matter
PII: S0169-2070(99)00018-7

brief review only. The Delphi technique was developed during the 1950s by workers at the RAND Corporation while involved on a U.S. Air Force sponsored project. The aim of the project was the application of expert opinion to the selection – from the point of view of a Soviet strategic planner – of an optimal U.S. industrial target system, with a corresponding estimation of the number of atomic bombs required to reduce munitions output by a prescribed amount. More generally, the technique is seen as a procedure to ''obtain the most reliable consensus of opinion of a group of experts . . . by a series of intensive questionnaires interspersed with controlled opinion feedback'' (Dalkey & Helmer, 1963, p. 458). In particular, the structure of the technique is intended to allow access to the positive attributes of interacting groups (knowledge from a variety of sources, creative synthesis, etc.), while pre-empting their negative aspects (attributable to social, personal and political conflicts, etc.). From a practical perspective, the method allows input from a larger number of participants than could feasibly be included in a group or committee meeting, and from members who are geographically dispersed.

Delphi is not a procedure intended to challenge statistical or model-based procedures, against which human judgment is generally shown to be inferior: it is intended for use in judgment and forecasting situations in which pure model-based statistical methods are not practical or possible because of the lack of appropriate historical/economic/technical data, and thus where some form of human judgmental input is necessary (e.g., Wright, Lawrence & Collopy, 1996). Such input needs to be used as efficiently as possible, and for this purpose the Delphi technique might serve a role.

Four key features may be regarded as necessary for defining a procedure as a 'Delphi'. These are: anonymity, iteration, controlled feedback, and the statistical aggregation of group response. Anonymity is achieved through the use of questionnaires. By allowing the individual group members the opportunity to express their opinions and judgments privately, undue social pressures – as from dominant or dogmatic individuals, or from a majority – should be avoided. Ideally, this should allow the individual group members to consider each idea on the basis of merit alone, rather than on the basis of potentially invalid criteria (such as the status of an idea's proponent). Furthermore, with the iteration of the questionnaire over a number of rounds, the individuals are given 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 questionnaire iteration, controlled feedback is provided through which the group members are informed of the opinions of their anonymous colleagues. Often feedback is presented as a simple statistical summary of the group response, usually comprising a mean or median value, such as the average 'group' estimate of the date by when an event is forecast to occur. Occasionally, additional information may also be provided, such as arguments from individuals whose judgments fall outside certain pre-specified limits. In this manner, feedback comprises the opinions and judgments of all group members and not just the most vocal. At the end of the polling of participants (i.e., after several rounds of questionnaire iteration), the group judgment is taken as the statistical average (mean/median) of the panellists' estimates on the final round. The final judgment may thus be seen as an equal weighting of the members of a staticized group.

The above four characteristics are necessary defining attributes of a Delphi procedure, although there are numerous ways in which they may be applied. The first round of the classical Delphi procedure (Martino, 1983) is unstructured, allowing the individual experts relatively free scope to identify, and elaborate on, those issues they see as important. These individual factors are then consolidated into a single set by the monitor team, who produce a structured questionnaire from which the views, opinions and judgments of the Delphi panellists may be elicited in a quantitative manner on subsequent rounds. After each of these rounds, responses are analysed and statistically summarised (usually into medians plus upper and lower quartiles), which are then presented to the panellists for further consideration. Hence, from the third round onwards, panellists are given the opportunity to alter prior estimates on the basis of the provided feedback. Furthermore, if panellists' assessments fall outside the upper or lower quartiles they may be asked to give reasons why they believe their selections are correct against the majority opinion. This procedure continues until

a certain stability in panellists' responses is achieved. The forecast or assessment for each item in the questionnaire is typically represented by the median on the final round.

An important point to note here is that variations from the above Delphi ideal do exist (Linstone, 1978; Martino, 1983). Most commonly, round one is structured in order to make the application of the procedure simpler for the monitor team and panellists; the number of rounds is variable, though seldom goes beyond one or two iterations (during which time most change in panellists' responses generally occurs); and often, panellists may be asked for just a single statistic – such as the date by when an event has a 50% likelihood of occurring – rather than for multiple figures or dates representing degrees of confidence or likelihood (e.g., the 10% and 90% likelihood dates), or for written justifications of extreme opinions or judgments. These simplifications are particularly common in laboratory studies and have important consequences for the generalisability of research findings.

3. The study of Delphi

Since the 1950s, use of Delphi has spread from its origins in the defence community in the U.S.A. to a wide variety of areas in numerous countries. Its applications have extended from the prediction of long-range trends in science and technology to applications in policy formation and decision making. An examination of recent literature, for example, reveals how widespread is the use of Delphi, with applications in areas as diverse as the health care industry (Hudak, Brooke, Finstuen & Riley, 1993), marketing (Lunsford & Fussell, 1993), education (Olshfski & Joseph, 1991), information systems (Neiderman, Brancheau & Wetherbe, 1991), and transportation and engineering (Saito & Sinha, 1991).

Linstone and Turoff (1975) characterised the growth of interest in Delphi as from non-profitmaking organisations to government, industry and, finally, to academe. This progression caused them some concern, for they went on to suggest that the 'explosive growth' in Delphi applications may have appeared '' . . . incompatible with the limited amount of controlled experimentation or academic investigation that has taken place . . . (and that) aside from some RAND studies by Dalkey, most 'evaluations' of the technique have been secondary efforts associated with some application which was the primary interest'' (p. 11). Indeed, the relative paucity of evaluation studies noted by Linstone and Turoff is still evident. Although a number of empirical examinations have subsequently been conducted, the bulk of Delphi references still concern applications rather than evaluations. That is, there appears to be a widespread assumption that Delphi is a useful instrument that may be used to measure some kind of truth and it is mainly represented in the literature in this vein. Consideration of what the evaluation studies actually tell us about Delphi is the focus of this paper.

4. Evaluative studies of Delphi

We attempted to gather together details of all published (English-language) studies involving evaluation of the Delphi technique. There were a number of types of studies that we decided not to include in our analysis. Unpublished PhD theses, technical reports (e.g., of the RAND Corporation), and conference papers were excluded because their quality is less assured than peer-reviewed journal articles and (arguably) book chapters. It may also be argued that if the studies reported in these excluded formats were significant and of sufficient quality then they would have appeared in conventional published form.

We searched through nine computer databases: ABI Inform Global, Applied Science and Technology Index, ERIC, Transport, Econlit, General Science Index, INSPEC, Sociofile, and Psychlit. As search terms we used: 'Delphi and Evaluation', 'Delphi and Experiment', and 'Delphi and Accuracy'. Because our interest is in experimental evaluations of Delphi, we excluded any 'hits' that simply reported Delphi as used in some *application* (e.g., to ascertain expert views on a particular topic), or that omitted key details of method or results because Delphi per se was not the focus of interest (e.g., Brockhoff, 1984). Several other studies were excluded because we believe they yielded no substantial findings, or

because they were of high complexity with uncertain findings. For example, a paper by Van Dijk (1990) effectively used six different Delphi-like techniques that differed in terms of the order in which each of three forms of panellist interaction were used in each of three 'Delphi' rounds, resulting in a highly complex, difficult to encode and, indeed, difficult to interpret study.

The final type of studies that we excluded from our search 'hits' were those that considered the universal *validity* or *reliability* of Delphi without reference to control conditions or other techniques. For example, the study of Ono and Wedermeyer (1994) reported that a Delphi panel produced forecasts that were over 50% accurate, and used this as 'evidence' that the technique is somehow a valid predictor of the future. However, the validity of the technique, in this sense, will depend as much on the nature of the panellists and the task as on the technique itself, and it is not sensible to suggest that the percentage accuracy achieved here is in any way generalisable or fundamental (i.e., to suggest that using Delphi will generally result in 50% correct predictions). The important questions – not asked in this study – are whether the application of the technique helped to *improve* the forecasting of the individual panellists, and how effective the technique was relative to other possible forecasting procedures. Similarly, papers by Felsenthal and Fuchs (1976), Dagenais (1978), and Kastein, Jacobs, van der Hell, Luttik and Touw-Otten (1993) – which reported data on the reliability of Delphi-like groups – are not considered further, for they simply demonstrate that a degree of reliability is possible using the technique.

The database search produced 19 papers relevant to our present concern. The references in these papers drew our attention to a further eight studies, yielding 27 studies in all. Roughly following the procedure of Armstrong and Lusk (1987), we sent out information to the authors of these papers. We identified the current addresses of the first-named authors of 25 studies by using the ISI Social Citation Index database. We produced five prototype tables summarising the methods and findings of these evaluative papers (discussed shortly), which we asked the authors to consider and comment upon, particularly with regard to our coding and interpretation of each author's own paper. We also asked the authors whether there were any evaluative studies that we had overlooked. The eight replies caused us to make two minor changes to our tables. None of the replies noted evaluative studies of which we were unaware.

Details of the Delphi evaluative studies are summarised in Tables 1 and 2. Table 1 describes the characteristics of each experimental scenario, including the nature of the task and the way in which Delphi was constituted. Table 2 reports the findings of the studies plus 'additional comments' (which also concern details in Table 1). Ideally, the two tables should be joined to form one comprehensive table, but space constraints forced the division.

We have tried to make the tables as comprehensive as possible without obscuring the main findings and methodological features by descriptions of minor or coincidental results and details, or by repetition of common information. For example, a number of Delphi studies used post-task questionnaires to assess panellists' reactions to the technique, with individual questions ranging from how 'satisfied' subjects felt about the technique, to how 'enjoyable' they found it (e.g., Van de Ven & Delbecq, 1974; Scheibe et al., 1975; Rohrbaugh, 1979; Boje & Murnighan, 1982). Clearly, to include every analysis or significant correlation involving such measures would result in an unwieldy mass of data, much of which has little theoretical importance and is unlikely to be replicated or considered in future studies. We have also been constrained by lack of space: so, while we have attempted to indicate how aspects such as the dependent variables were measured (Table 2), we have generally been forced to give only brief descriptions in place of the complex scoring formulae used. In other cases where descriptions seem overly brief, the reader is referred to the main text where matters are explained more fully.

Regarding Table 1, the studies have been classified according to their aims and objectives as 'Application', 'Process' or 'Technique-comparison' type. Application studies use Delphi in order to gauge expert opinion on a particular topic, but consider issues on the process or quality of the technique as a secondary concern. Process studies focus on aspects of the internal features of Delphi, such as the role of feedback, the impact of the size of Delphi groups, and so on. Technique-comparison studies evaluate

Table 1
Summary of the methodological features of Delphi in experimental studies[a]

Study	Study type	Delphi group size	Rounds	Nature of Delphi feedback	Nature of subjects	Task	Incentives offered?
Dalkey and Helmer (1963)	Application (process)	7	5	Individual estimates (bombing schedules)	Professionals: e.g., economists, systems analysts, electronic engineer	Hypothetical event: number of bombs to reduce munitions output by stated amount	No
Dalkey, Brown and Cochran (1970)	Process	15–20	2	Unclear	Students: no expertise on task	Almanac questions (20 per group, 160 in total)	No
Jolson and Rossow (1971)	Process	11 and 14	3	Medians (normalised probability medians)	Professionals: computing corporation staff, and naval personnel	Forecast (demand for classes: one question) Almanac (two questions)	No
Gustafson, Shukla, Delbecq and Walster (1973)	Technique comparison	4	2	Individual estimates (likelihood ratios)	Students: no expertise on task	Almanac (eight questions requiring probability estimates in the form of likelihood ratios)	No
Best (1974)	Process	14	2	Medians, range of estimates, distributions of expertise ratings, and, for one group, reasons	Professionals: faculty members of college	Almanac (two questions requiring numerical estimates of known quantity), and Hypothetical event (one probability estimation)	No
Van de Ven and Delbecq (1974)	Technique comparison	7	2	Pooled ideas of group members	Students (appropriate expertise), Professionals in education	Idea generation (one item: defining job description of student dormitory counsellors)	No
Scheibe, Skutsch and Schofer (1975)	Process	Unspecified (possibly 21)	5	Means, frequency distribution, reasons (from Ps distant from mean)	Students (uncertain expertise on task)	Goals Delphi (evaluating hypothetical transport facility alternatives via rating goals)	No
Mulgrave and Ducanis (1975)	Process	98	3	Medians, interquartile range, own responses to previous round	Students enrolled in Educational Psychology class (most of whom were school teachers)	Almanac (10 general, and 10 related to teaching) and Hypothetical (18 on what values in U.S. will be, and then should be)	No

Table 1. Continued

Study	Study type	Delphi group size	Rounds	Nature of Delphi feedback	Nature of subjects	Task	Incentives offered?
Brockhoff (1975)	Technique comparison, Process	5, 7, 9, 11	5	Medians, reasons (of those with responses outside of quartiles)	Professionals: banking	Almanac and Forecasting (8–10 items on finance, banking, stock quotations, foreign trade, for each type)	Yes
Miner (1979)	Technique comparison	4	Up to 7	Not specified	Students: from diverse undergrad populations	Problem Solving (role-playing exercise assigning workers to subassembly positions)	No
Rohrbaugh (1979)	Technique comparison	5–7	3	Medians, ranges	Students (appropriate expertise)	Forecasting (college performance re grade point average of 40 freshmen)	No
Fischer (1981)	Technique comparison	3	2	Individual estimates (subjective probability distributions)	Students (appropriate expertise)	Giving probability of grade point averages of 10 freshmen falling in four ranges	Yes
Boje and Murnighan (1982)	Technique comparison (process)	3, 7, 11	3	Individual estimates, one reason from each P	Students (no expertise on task)	Almanac (two questions Jupiter and Dollars) and Subjective Likelihood (two questions: Weight and Height)	No
Spinelli (1983)	Application (process)	20 (24 initially)	4	Modes (via histogram), distributions, reasons	Professionals: education	Forecasting (likelihood of 34 education-related events occurring in 20 years, on 1–6 scale)	No
Larreché and Moinpour (1983)	Technique comparison	5	4	Means, lowest and highest estimates, reasons	MBA students (with some expertise in business area of task)	Forecasting (market share of product based on eight communications plans; historical data provided)	No
Riggs (1983)	Technique comparison	4–5	2	Means (of point spreads)	Students	Forecasting (point spreads of two college football games taking place 4 weeks in future)	No
Parenté et al. (1984)	Process	80	2	Percentage of group who predicted event to occur, median time scale	Students	Forecasting (30 scenarios on economic, political and military events)	No
Erffmeyer and Lane (1984)	Technique comparison	4	6	Individual estimates (ranks), reasons	Students (no expertise on task)	Hypothetical event (Moon Survival Problem, involving ranking 15 items for survival value)	No

Erffmeyer, Erffmeyer and Lane (1986)	Process	4	6	Individual estimates (ranks), reasons	Students (no expertise on task)	Hypothetical event (Moon Survival Problem, involving ranking 15 items for survival value)	No
Dietz (1987)	Process	8	3	Mean, median, upper and lower quantities, plus justifications on r3	Students: state government analysts or employed in public sector firms	Forecasting (proportion of citizens voting yes on ballot in California election)	No
Sniezek (1989)	Technique comparison	5	Varied ≤5	Medians	Students (no expertise on task)	Forecasting (of dollar amount of next month's sales at campus store re four categories of items)	Yes
Sniezek (1990)	Technique comparison	5	<6	Medians	Students (no expertise on task)	Forecasting (five items concerned with economic variables: forecasts were for 3 months in the future)	Yes
Taylor, Pease and Reid (1990)	Process	59	3	Individual responses (i.e. chosen characteristics)	Elementary school teachers	Policy (identifying the qualities of effective inservice program: determining top seven characteristics)	No
Leape, Freshour, Yntema and Hsiao (1992)	Technique comparison	19, 11	2 and 3	Distribution of numerical evaluations of each of the 55 items	Surgeons	Policy (judged intraservice work required to perform each of 55 services, via magnitude estimation)	No
Gowan and McNichols (1993)	Process (application)	15 in r1	3	Response frequency or regression weights or induced (if-then) rules	Bank loan officers and small business consultants	Hypothetical event (evaluation of likelihood of a company's economic viability in future)	No
Hornsby, Smith and Gupta (1994)	Technique comparison	5	3	Statistical average of points for each factor	Students (no particular expertise but some task training)	Policy (evaluating four jobs using the nine-factor Factor Evaluation System)	No
Rowe and Wright (1996)	Process	5	2	Medians, individual numerical responses, reasons	Students (no particular expertise though items selected to be pertinent to them)	Forecasting (of 15 political and economic events for 2 weeks into future)	No

[a] *Analysis significant at the $P < 0.05$ level; NS, analysis not significant at the $P < 0.05$ level (trend only); NDA, no statistical analysis of the noted relationship was reported; >, greater/better than; =, no (statistically significant) difference between; D, Delphi; R, first round average of Delphi polls; S, staticized group approach (average of non-interacting set of individuals). When used, this indicates that separate staticized groups were used instead of (or in addition to) the average of the individual panellists from the first round Delphi polls (i.e., instead of/in addition to 'R'); I, interacting group; NGT, Nominal Group Technique; r, 'round'; P, 'panellist'; NA, not applicable.

Table 2
Results of the experimental Delphi studies[a]

Study	Independent variables (and levels)[b,c]	Dependent variables (and measures)	Results of technique comparisons	Other results, e.g. re processes	Additional comments
Dalkey and Helmer (1963)	Rounds	Convergence (difference in estimates over rounds)	NA	Increased convergence (consensus) over rounds (NDA)	A complex design where quantitative feedback from panellists was only supplied in round 4, and from only two other panellists
Dalkey et al. (1970)	Rounds	Accuracy (logarithm of median group estimate divided by true answer) Self-rated knowledge (1–5 scale)	Accuracy: D>R (NDA)	High self-rated knowledgeability was positively related to accuracy (NDA)	Little statistical analysis reported. Little detail on experimental method, e.g. no clear description of round 2 procedure, or feedback
Jolson and Rossow (1971)	Expertise (experts vs. non-experts) Rounds	Accuracy (numerical estimates on the two almanac questions only). Consensus (distribution of P responses)	Accuracy: D>R (experts) (NDA) Accuracy: D<R (non-experts) (NDA)	Increased consensus over rounds*. Claimed relationship between high r1 accuracy and low variance over rounds ('significant')	Often reported study. Used a small sample, so little statistical analysis was possible
Gustafson et al. (1973)	Techniques (NGT, I, D, S) Rounds	Accuracy (absolute mean percentage error). Consensus (variance of estimates across groups)	Accuracy: NGT>I>S>D (* for techniques) Accuracy: D<R (NS) Consensus: NGT=D=I=S	NA	Information given to all panellists to control for differences in knowledge. This approach seems at odds to rationale of using Delphi with heterogenous groups
Best (1974)	Feedback (statistic vs. reasons) Self-rated expertise (high vs. low) Rounds	Accuracy (absolute deviation of assessments from true value). Self-rated expertise (1–11 scale)	Accuracy: D>R* Accuracy: D(+reasons)>D(no reasons)* (for one of two items)	High self-rated expertise positively related to accuracy, and sub-groups of self-rated experts more accurate than non-experts*	Self-rated expertise was used as both a dependent and independent variable
Van de Ven and Delbecq (1974)	Techniques (NGT, I, D)	Idea quantity (number of unique ideas). Satisfaction (five items rated on 1–5 scales). Effectiveness (composite measure of idea quantity and satisfaction, equally weighted)	Number of ideas: D>I*; NGT>D (NS) Satisfaction: NGT>D*, I* Effectiveness: NGT*, D*>I	NA	Two open-ended questions were also asked about likes and dislikes concerning the procedures
Scheibe et al. (1975)	Rating scales (ranking, rating-scales method, and pair comparisons) Rounds	Confidence (assessed by seven item questionnaire) Change (five aspects, e.g. from r1 to r2)	NA	Little evidence of correlation between response change and low confidence, except in r2(*). False feedback initially drew responses to it (NDA)	Poor design meant nothing could be concluded about consistency of using different rating scales. Missing information on design details
Mulgrave and Ducanis (1975)	Expertise (e.g., on items about teaching Ps considered expert, on general items they non-expert) Rounds	Change (% changing answers) Dogmatism (on Rokeach Dogmatism Scale)	NA	High dogmatism related to high change over rounds*, most evident on items on which the most dogmatic were relatively inexpert	A brief study, with little discussion of the unexpected/counter-intuitive results

Brockhoff (1975)	Techniques (D, I) Group size (5, 7, 9, 11, D; 4, 6, 8, 10, I) Item type (forecasting vs. almanac) Rounds	Accuracy (absolute error) Self-rated expertise (1–5 scale)	Accuracy: D>I (almanac items) (NDA) Accuracy: I>D (forecast items) (NDA) Accuracy: D>R (NS) (Accuracy increased to r3; then decreased)	No clear results on influence of group size. Self-rated expertise not related to accuracy	Significant differences in accuracy of groups of different sizes were due to differences in initial accuracy of groups at round 1
Miner (1979)	Techniques (NGT, D, PCL)	Accuracy/Quality (index from 0.0 to 1.0) Acceptance (11 statement q'naire) Effectiveness (product of other indexes)	Accuracy: PCL>NGT>D (NS) Acceptance: D=NGT=PCL Effectiveness: PCDL>D*, NGT*	NA	Effectiveness was determined by the product of quality (accuracy) and acceptance
Rohrbaugh (1979)	Techniques (D, SJ)	Accuracy (correlation with solution) Post-group agreement (difference between group and individual post-group policy) Satisfaction (0–10 scale)	Accuracy: D=SJ Post-group agreement: SJ>D ('significant') Satisfaction: SJ>D ('significant')*	Quality of decisions of individuals increased only slightly after the group Delphi process	Delphi groups instructed to reduced existing differences over rounds. Agreement concerns degree to which panellists agreed with each other after group process had ended (post-task rating)
Fischer (1981)	Techniques (NGT, D, I, S)	Accuracy (according to truncated logarithmic scoring rule)	Accuracy: D=NGT=I=S (Procedures Main Effect: $P>0.20$)	NA	
Boje and Murnighan (1982)	Techniques (NGT, D, Iteration) Group size (3, 7, 11) Rounds	Accuracy (deviation of gp mean from correct answer) Confidence (1–7 scale) Satisfaction (responses on 10 item q'naire)	Accuracy: R>D, NGT Accuracy: Iteration procedure>R (Procedures×Trials was significant)*	No relationship between group size and accuracy or confidence. Confidence in Delphi increased over rounds*	Compared Delphi to an NGT procedure and to an iteration (no feedback) procedure
Spinelli (1983)	Rounds	Convergence (e.g., change in interquartile range)	NA	No convergence over rounds	No quoted statistical analysis. In r3 and r4, Ps encouraged to converge towards majority opinion!
Larreché and Moinpour (1983)	Techniques (D, I) Rounds	Accuracy (Mean Absolute Percentage Error) Expertise (via (a) Confidence in estimates, (b) external measure)	Accuracy: D>R* Accuracy: D>I* Accuracy: R=I	Expertise assessed by self-rating had no significant validity, but did when assessed by an external measure.* Aggregate of latter 'experts' increased accuracy over other Delphi groups*	A measure of confidence was taken as synonymous with expertise. Evidence that composing Delphi groups of the best (externally-rated) experts may improve judgement
Riggs (1983)	Techniques (D, I) Information on task (high vs. low)	Accuracy (absolute difference between predicted and actual point spreads)	Accuracy: D>I*	More accurate predictions were provided for the high information game.* Delphi was more accurate in both high and low information cases	Comparison with first round not reported. Standard information on the four teams in the two games was provided
Parenté, Anderson, Myers and O'Brien (1984)	Feedback (yes vs. no) Iteration (yes vs. no)	Accuracy (IF an event would occur; and error in days re WHEN an event would occur)	Accuracy: D>R (WHEN event occurs) Accuracy: D=R (IF event occurs)	Decomposition of Delphi seemed to show improvement in accuracy related to the iteration component, not feedback (WHEN item)	Three related studies reported: data here on ex3. Suggestion that feedback superfluous in Delphi, however, feedback used here was superficial
Erffmeyer and Lane (1984)	Techniques (NGT, D, I, Group procedure with guidelines on appropriate behavior) Rounds	Quality (comparison of rankings to 'correct ranks' assigned by NASA experts) Acceptance (degree of agreement between an individual's ranking and final grp ranking)	Quality: D>NGT*, I*, Guided gp* Quality: D>R*, also note: I>NGT* Acceptance: Guided gp procedure>D*	NA	Acceptance bears similarity to the post-task consensus ('agreement') measure of Rohrbaugh (1979)

Table 2. Continued

Study	Independent variables (and levels)[b,c]	Dependent variables (and measures)	Results of technique comparisons	Other results, e.g. re processes	Additional comments
Erffmeyer et al. (1986)	Rounds	Quality (comparison of rankings to 'correct ranks' assigned by NASA experts)	Quality: D>R*	Quality increased up to the fourth round, but not thereafter: 'stability' was achieved	Is this an analysis of data from the study of Erffmeyer and Lane (1984)?
Dietz (1987)	Weighting (using confidence vs. no) Summary measure (median, mean, Hampel, biweight, trimmed mean) Rounds	Accuracy (difference between forecast value and actual percentage) Confidence (1–5 scale)	Accuracy D>R (NS)	Weighted forecasts less accurate than non-weighted ones, though difference small. Differences in summary methods insignificant. Least confident not most likely to change opinions	Very small study with only one group of eight panellists, and hence it is not surprising that statistically significant results were not achieved
Sniezek (1989)	Techniques (D, I, Dictator, Dialectic) (repeated-measures design) Rounds	Accuracy (absolute percentage error) Confidence (1–7 scale) Influence (difference between individual's judgment and r1 group mean)	Accuracy: Dictator, D>I (NDA) Accuracy: D>R (NS)	High confidence and accuracy were correlated in Delphi*. Influence in Delphi was correlated with confidence* and accuracy*	Influence measure concerned extent to which an individual's estimates drew the group judgment. Time-series data supplied.
Sniezek (1990)	Techniques (D, I, S, Best Member (BM)) Item difficulty (mean % errors)	Accuracy (e.g., mean forecast error) Confidence (difference between upper and lower limits of Ps 50% confidence intervals)	Accuracy: D>BM (one of five items) ($P<0.10$) Accuracy: S>D (one of five items) ($P<0.10$) No. sig. differences in other comparisons	Confidence measures were not correlated to accuracy in Delphi groups	All individuals were given common information (time-series data) prior to the task. Item difficulty measured post task
Taylor et al. (1990)	None	Survivability (degree to which r1 contribution of each P is retained in later rounds) Abandonment (degree of Ps abandoning r1 contributions in favour of those of others) Demographic factors (four types)	NA	No relationship between demographic characteristics of panellists (gender, education, interest, and effectiveness and survivability/abandonment)	Abandonment and Survivability relate to influence and opinion change
Leape et al. (1992)	Techniques (D, single round mail survey, moderated D with panel discussion)	Agreement (compared mean of adjusted ratings to that gained in national survey)	Agreement: Mail survey>D>Modified D (Modified D included panel discussion)	NA	Agreement was measured by comparing panel estimates to those from a national survey. Not a true measure of accuracy
Gowan and McNichols (1993)	Feedback (statistical, regression model, if-then rule feedback)	Consensus (pairwise subject agreement after each round of decisions)	NA	If-then rule feedback gave greater consensus than statistical or regression model feedback*	Accuracy was measured, but not reported in this paper
Hornsby et al. (1994)	Techniques (NGT, D, Consensus)	Degree of opinion change (from averaged individual evaluations to final gp evaluation) Satisfaction (12 item q'naire: 1–5 scale)	Opinions changed from initial evaluations during NGT* and consensus*, but not Delphi. Satisfaction: D>consensus*, NGT*	NA	No measure of accuracy was possible. Consensus involves discussion until all group members accept final decision
Rowe and Wright (1996)	Feedback (statistical, reasons, iteration – without feedback) (repeated measures design) Rounds	Accuracy (Mean Absolute Percentage Error) Opinion change (in terms of z-score measure) Confidence (each item, 1–7 scale) Self-rated expertise (one measure, 1–7 scale) Objective expertise of Ps (via MAPES)	Accuracy: D(reas), D(start), Iterate>R (all*) Accuracy: D(reas)>D(start)>Iteration (NS) Change: Iteration>D(start), D(reas)*	Best objective experts changed least in the two feedback conditions*. High self-rated experts were most accurate on r1* but no relationship to change variable. Confidence increased over rounds in all conditions*	Although subjects changed their forecasts less in the reasons condition, when they did so the new forecasts were more accurate.* This trend did not occur in the other conditions, suggesting reasons feedback helped discriminality of change

[a] See Table 1 for explanation of abbreviations.

[b] For the number of rounds (i.e. levels of independent variable 'Round'), see Table 1.

[c] In many of the studies, a number of different questions are presented to the subjects so that 'items' may be considered to be an independent variable. Usually, however, there is no specific rationale behind analysing responses to the different items, so we do not consider 'items' an I.V. here.

Delphi in comparison to external benchmarks, that is, with respect to other techniques or means of aggregating judgment (for example, interacting groups).

In Table 2, the main findings have been coded according to how they were presented by their authors. In some cases, claims have been made that an effect or relationship was found even though there was no backing statistical analysis, or though a statistical test was conducted but standard significance levels were not reached (because of, for example, small sample sizes, e.g. Jolson & Rossow, 1971). It has been argued that there are problems in using and interpreting '*P*' figures in experiments and that effect sizes may be more useful statistics, particularly when comparing the results of independent studies (e.g., Rosenthal, 1978; Cohen, 1994). In many Delphi studies, however, effect sizes are not reported in detail. Thus, Table 2 has been annotated to indicate whether claimed results were significant at the $P = 0.05$ level (*); whether statistical tests were conducted but proved non-significant (NS); or whether there was no direct statistical analysis of the result (NDA). Readers are invited to read whatever interpretation they wish into claims of the statistical significance of findings.

5. Findings

In this section, we consider the results obtained by the evaluative studies as summarised in Table 2, to which the reader is referred. We will return to the details in Table 1 in our subsequent critique.

5.1. Consensus

One of the aims of using Delphi is to achieve greater *consensus* amongst panellists. Empirically, consensus has been determined by measuring the variance in responses of Delphi panellists over rounds, with a reduction in variance being taken to indicate that greater consensus has been achieved. Results from empirical studies seem to suggest that variance reduction is typical, although claims tend to be simply reported unanalysed (e.g., Dalkey & Helmer, 1963), rather than supported by analysis (though see Jolson & Rossow, 1971). Indeed, the trend of reduced variance is so typical that the phenomenon of increased 'consensus', per se, no longer appears to be an issue of experimental interest.

Where some controversy does exist, however, is in whether a reduction in variance over rounds reflects true *consensus* (reasoned acceptance of a position). Delphi has, after all, been advocated as a method of reducing group pressures to *conform* (e.g., Martino, 1983), and both increased consensus and increased conformity will be manifest as a convergence of panellists' estimates over rounds (i.e., these factors are confounded). It would seem in the literature that reduced variance has been interpreted according to the position on Delphi held by the particular author/s, with proponents of Delphi arguing that results demonstrate consensus, while critics have argued that the 'consensus' is often only 'apparent', and that the convergence of responses is mainly attributable to other social-psychological factors leading to conformity (e.g., Sackman, 1975; Bardecki, 1984; Stewart, 1987). Clearly, if panellists are being drawn towards a central value for reasons other than a genuine acceptance of the rationale behind that position, then inefficient process-loss factors are still present in the technique.

Alternative measures of consensus have been taken, such as 'post-group consensus'. This concerns the extent to which individuals – after the Delphi process has been completed – individually agree with the final group aggregate, their own final round estimates, or the estimates of other panellists. Rohrbaugh (1979), for example, compared individuals' post-group responses to their aggregate group responses, and seemed to show that reduction in 'disagreement' in Delphi groups was significantly less than the reduction achieved with an alternative technique (Social Judgment Analysis). Furthermore, he found that there was little increase in agreement in the Delphi groups. This latter finding seems to suggest that panellists were simply altering their estimates in order to conform to the group without actually changing their opinions (i.e., implying conformity rather than genuine consensus).

Erffmeyer and Lane (1984) correlated post-group individual responses to group scores and found there to be significantly more 'acceptance' (i.e., significantly higher correlations between these measures) in

an alternative structured group technique than in a Delphi procedure, although there were no differences in acceptance between the Delphi groups and a variety of other group techniques (see Table 2). Unfortunately, no analysis was reported on the difference between the correlations of the post-group individual responses and the first and final round group aggregate responses.

An alternative slant on this issue has been provided by Bardecki (1984), who reported that – in a study not fully described – respondents with more extreme views were more likely to drop out of a Delphi procedure than those with more moderate views (i.e., nearer to the group average). This suggests that consensus may be due – at least in part – to attrition. Further empirical work is needed to determine the extent to which the convergence of those who do not (or cannot) drop out of a Delphi procedure are due to either true consensus or to conformity pressures.

5.2. Increased accuracy

Of main concern to the majority of researchers is the ability of Delphi to lead to judgments that are more accurate than (a) initial, pre-procedure aggregates (equivalent to equal-weighted staticized groups), and (b) judgments derived from alternative group procedures. In order to ensure that accuracy can be rapidly assessed, problems used in Delphi studies have tended to involve either short-range forecasting tasks, or tasks requiring the estimation of almanac items whose quantitative values are already known to the experimenters and about which subjects are presumed capable of making educated guesses. In evaluative studies, the long-range forecasting and policy formation items (etc.) that are typical in Delphi applications are rarely used.

Table 2 reports the results of the evaluation of Delphi with regards to the various benchmarks, while Tables 3–5 collate and summarise the comparisons according to specific benchmarks.

5.2.1. Delphi versus staticized groups

The average estimate of Delphi panellists on the first round – prior to iteration or feedback – is equivalent to that from a staticized group. Comparing a final round Delphi aggregate to that of the first round is thus, effectively, a *within-subjects* comparison of techniques (Delphi versus staticized group); when comparison occurs between Delphi and a separate staticized group then it is a *between-subjects* one. Clearly, the former comparison is preferable, given that it controls for the highly variable influence of *subjects*. Although the comparison of round averages should be possible in every study considering Delphi accuracy/quality, a number of evaluative studies have omitted to report round differences (e.g., Fischer, 1981; Riggs, 1983). Comparisons of relative accuracy of Delphi panels with first round aggregates and staticized groups are reported in Table 3.

Evidence for Delphi effectiveness is equivocal, but results generally support its advantage over first round/staticized group aggregates by a tally of 12 studies to two. Five studies have reported *significant* increases in accuracy over Delphi rounds (Best, 1974; Larreché & Moinpour, 1983; Erffmeyer & Lane, 1984; Erffmeyer et al., 1986; Rowe & Wright, 1996), although the two papers of Erffmeyer et al. may be reports of separate analyses on the same data (this is not clear). Seven more studies have produced qualified support for Delphi: in five cases, Delphi is found to be better than statistical or first round aggregates more often than not, or to a degree that does not reach statistical significance (e.g., Dalkey et al., 1970; Brockhoff, 1975; Rohrbaugh, 1979; Dietz, 1987; Sniezek, 1989), and in two others it is shown to be better under certain conditions and not others (i.e., in Parenté et al., 1984, Delphi accuracy increases over rounds for predicting 'when' an event might occur, but not 'if' it will occur; in Jolson & Rossow, 1971, it increases for panels comprising 'experts', but not for 'non-experts').

In contrast, two studies found no substantial difference in accuracy between Delphi and staticized groups (Fischer, 1981; Sniezek, 1990), while two others suggested 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 than independent staticized 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.

Table 3
Comparisons of Delphi to staticized (round 1) groups[a]

Study	Criteria	Trend of relationship (moderating variable)	Additional comments
Dalkey et al. (1970)	Accuracy	D>R (NDA)	81 groups became more accurate, 61 became less
Jolson and Rossow (1971)	Accuracy Accuracy	D>R (experts) (NDA) D<R ('non experts') (NDA)	For both items in each case
Gustafson et al. (1973)	Accuracy	D<S (NDA) D<R (NDA)	D worse than S on six of eight items, and than R on seven of eight items
Best (1974)	Accuracy	D>R*	For each of two items
Brockhoff (1975)	Accuracy	D>R (almanac items) D>R (forecast items)	Accuracy generally increased to r3, then started to decrease
Rohrbaugh (1979)	Accuracy	D>R (NDA)	No direct comparison made, but trend implied in results
Fischer (1981)	Accuracy	D=S	Mean group scores were virtually indistinguishable
Boje and Murnighan (1982)	Accuracy	R>D*	Accuracy in D decreased over rounds for three out of four items
Larreché and Moinpour (1983)	Accuracy	D>R*	
Parenté et al. (1984)	Accuracy Accuracy	D>R (WHEN items) (NDA) D=R (IF items) (NDA)	The findings noted here are the trends as noted in the paper
Erffmeyer and Lane (1984)	Quality	D>R*	Of four procedures, D seemed to improve most over rounds
Erffmeyer et al. (1986)	Quality	D>R*	Most improvement in early rounds
Dietz (1987)	Accuracy	D>R	Error of D panels decreased from r1 to r2 to r3
Sniezek (1989)	Accuracy	D>R	Small improvement over rounds but not significant
Sniezek (1990)	Accuracy	S>D (one of five items)	No reported comparison of accuracy change over D rounds
Rowe and Wright (1996)	Accuracy Accuracy	D (reasons f'back)>R* D (stats f'back)>R*	

[a] See Table 1 for explanation of abbreviations.

5.2.2. Delphi versus interacting groups

Another important comparative benchmark for Delphi performance is provided by interacting groups. Indeed, aspects of the design of Delphi are intended to pre-empt the kinds of social/psychological/political difficulties that have been found to hinder effective communication and behaviour in groups. The results of studies in which Delphi and interacting groups have been compared are presented in Table 4.

Research supports the relative efficacy of Delphi over interacting groups by a score of five studies to one with two ties, and with one study showing task-specific support for both techniques. Support for Delphi comes from Van de Ven and Delbecq (1974), Riggs (1983), Larreché and Moinpour (1983),

Table 4
Comparisons of Delphi to interacting groups[a]

Study	Criteria	Trend of relationship (moderating variable)	Additional comments
Gustafson et al. (1973)	Accuracy	I>D (NDA)	D worse than I on five of eight items
Van de Ven and Delbecq (1974)	Number of ideas	D>I*	
Brockhoff (1975)	Accuracy	D>I (almanac items) (NDA) I>D (forecast items) (NDA)	Comparisons were between I and r3 of D
Fischer (1981)	Accuracy	D=I	Mean group scores were indistinguishable
Larreché and Moinpour (1983)	Accuracy	D>I*	
Riggs (1983)	Accuracy	D>I*	Result was significant for both high and low information tasks
Erffmeyer and Lane (1984)	Quality	D>I*	
Sniezek (1989)	Accuracy	D>I (NDA)	Small sample sizes
Sniezek (1990)	Accuracy	D=I	No difference at $P = 0.10$ level on comparisons on the five items

[a] See Table 1 for explanation of abbreviations.

Erffmeyer and Lane (1984), and Sniezek (1989). Fischer (1981) and Sniezek (1990) found no distinguishable differences in accuracy between the two approaches, 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 than interacting groups for *almanac* items, but the reverse being the case for *forecasting* items.

5.2.3. Delphi versus other procedures

Although evidence suggests that Delphi does generally lead to improved judgments over staticized groups and unstructured interacting groups, it is clearly of interest to see how Delphi performs in comparison to groups using other structured procedures. Table 5 presents the findings of studies which have attempted such comparisons.

A number of studies have compared Delphi to the Nominal Group Technique or NGT (also known as the 'estimate-talk-estimate' procedure). NGT uses the basic Delphi structure, but uses it in face-to-face meetings that allow discussion between rounds. Clearly, there are practical situations where one technique may be viable and not the other, for example NGT would seem appropriate when a job needs to be done quickly, while Delphi would be apt when experts cannot meet physically; but which technique is preferable when both are options? Results of Delphi–NGT comparisons do not fully answer this question. Although there is some evidence that NGT groups make more accurate judgments than Delphi groups (Gustafson et al., 1973; Van de Ven & Delbecq, 1974), other studies have found no notable differences in accuracy/quality between them (Miner, 1979; Fischer, 1981; Boje & Murnighan, 1982), while one study has shown Delphi superiority (Erffmeyer & Lane, 1984).

Other studies have compared Delphi to: groups in which members were required to argue both for and against their individual 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 interaction (Erffmeyer & Lane, 1984); groups whose information exchange was structured according to Social Judgment Analysis (Rohrbaugh, 1979); and groups following a Problem Centred Leadership (PCL) approach (Miner, 1979). The only studies that found any substantial differences between Delphi and the comparison procedure/s are those of

Table 5
Comparisons of Delphi to other structured group procedures[a]

Study	Criteria	Trend of relationship (moderating variable)	Additional comments
Gustafson et al. (1973)	Accuracy	NGT>D*	NGT-like technique was more accurate than D on each of eight items
Van de Ven and Delbecq (1974)	# of ideas	NGT>D	NGT gps generated, on average, 12% more unique ideas than D gps
Miner (1979)	Quality	PCL>NGT>D	PCL was more effective than D ($P < 0.01$) (see Table 2)
Rohrbaugh (1979)	Accuracy	D=SJ	No P statistic reported due to nature of study/analysis
Fischer (1981)	Accuracy	D=NGT	Mean group scores were similar, but NGT very slightly better than D
Boje and Murnighan (1982)	Accuracy	D=NGT	On the four items, D was slightly more accurate than NGT gps on final r
Erffmeyer and Lane (1984)	Quality	D>NGT* D>Guided Group*	Guided gps followed guidelines on resolving conflict
Sniezek (1989)	Accuracy	Dictator=D=Dialectic (NDA)	Dialectic gps argued for and against positions. Dictator gps chose one member to make gp decision
Sniezek (1990)	Accuracy	D>BM (one of five items)	BM=Dictator (above): gp chooses one member to make decision

[a] See Table 1 for explanation of abbreviations.

Erffmeyer and Lane (1984), which found Delphi to be better than groups that were given instructions on resolving conflict, and Miner (1979), which found that the PCL approach (which involves instructing group leaders in appropriate group-directing skills) to be significantly more 'effective' than Delphi.

6. A critique of technique-comparison studies

Much of the criticism of early Delphi studies centred on their 'sloppy execution' (e.g., Stewart, 1987). Among specific criticisms were claims that Delphi questionnaires were poorly worded and ambiguous (e.g., Hill & Fowles, 1975) and that the analysis of responses was often superficial (Linstone, 1975). Reasons given for the poor conduct of early studies ranged from the technique's 'apparent simplicity' encouraging people without the requisite skills to use it (Linstone & Turoff, 1975), to suggestions that the early Delphi researchers had poor backgrounds in the social sciences and lacked acquaintance with appropriate research methodologies (e.g., Sackman, 1975).

Over the past few decades, empirical evaluations of the Delphi technique have taken on a more systematic and scientific nature, typified by the controlled comparison of Delphi-like procedures with other group and individual methods for obtaining judgments (i.e., technique comparison). Although the methodology of these studies has provoked a degree of support (e.g., Stewart, 1987), we have queried the relevance of the majority of this research to the general question of Delphi efficacy (e.g., Rowe et al., 1991), pointing out that most of such studies have used versions of Delphi somewhat removed from the 'classical' archetype of the technique (consider Table 1).

To begin with, the majority of studies have used structured first rounds in which event statements – devised by experimenters – are simply presented to panellists for assessment, with no opportunity for

them to indicate the issues they believe to be of greatest importance on the topic (via an unstructured round), and thus militating against the construction of coherent task scenarios. The concern of researchers to produce simple experimental designs is commendable, but sometimes the consequence is oversimplification and designs/tasks/measures of uncertain relevance. Indeed, because of a desire to verify rapidly the accuracy of judgments, a number of Delphi studies have used almanac questions involving, for example, estimating the diameter of the planet Jupiter (e.g., as in Gustafson et al., 1973; Mulgrave & Ducanis, 1975; Boje & Murnighan, 1982). Unfortunately, it is difficult to imagine how Delphi might conceivably benefit a group of unknowledgeable subjects – who are largely guessing the order of magnitude of some quantitatively large and unfamiliar statistic, such as the tonnage of a certain material shipped from New York in a certain year – in achieving a reasoned consideration of others' knowledge, and thus of improving the accuracy of their estimates. However, even when the items are sensible (e.g., of short-term forecasting type), their pertinence must also depend on the relevance of the expertise of the panellists: 'sensible' questions are only sensible if they relate to the domain of knowledge of the specific panellists. That is, student subjects are unlikely to respond to, for example, a task involving the forecast of economic variables, in the same way as economics experts – not only because of lesser knowledge on the topic, but because of differences in aspects such as their motivation (see Bolger & Wright, 1994, for discussion of such issues).

Delphi, however, was ostensibly designed for use with experts, particularly in cases where the variety of relevant factors (economic, technical, etc.) ensures that individual panellists have only limited knowledge and might benefit from communicating with others possessing different information (e.g., Helmer, 1975). This use of diverse experts is, however, rarely found in laboratory situations: for the sake of simplification, most studies employ homogenous samples of undergraduate or graduate students (see Table 1), who are required to make judgments about scenarios in which they can by no means be considered expert, and about which they are liable to possess similar knowledge (indeed, certain studies have even tried to standardise knowledge prior to the Delphi manipulation; e.g., Riggs, 1983). Indeed, the few laboratory studies that have used non-students have tended to use professionals from a single domain (e.g., Brockhoff, 1975; Leape et al., 1992; Ono & Wedermeyer, 1994). This point is important: if varied information is not available to be shared, then what could possibly be the benefit of any group-like aggregation over and above a statistical aggregation procedure?

There is some evidence that panels composed of relative experts tend to benefit from a Delphi procedure to a greater extent than comparative aggregates of novices (e.g., Jolson & Rossow, 1971; Larreché & Moinpour, 1983), suggesting that typical laboratory studies may *underestimate* the value of Delphi. Indeed, the relevance of the expertise of members of a *staticized* group, and the extent to which their judgments are shared/correlated, have been shown to be important factors related to accuracy in the theoretical model of Hogarth (1978), which has been empirically supported by a number of studies (e.g., Einhorn, Hogarth & Klempner, 1977; Ashton, 1986).

The absolute and relative expertise of a set of individuals is also liable to interact with the effectiveness of different group techniques – not only statistical groups and Delphi, but also interacting groups and techniques such as the NGT. Because the nature of such relationships are unknown, it is likely that, for example, though a Delphi-like procedure might prove more effective than an NGT approach at enhancing judgment with a particular set of individuals, a panel possessing different 'expertise' attributes might show the reverse trend. The uncomfortable conclusion from this line of argument is that it is difficult to draw grand conclusions on the relative efficacy of Delphi from technique-comparison studies that have made no clear attempt at specifying or describing the characteristics of their panels and the relevant features of the task. Furthermore, any particular demonstration of Delphi efficacy cannot be taken as an indicator of the more general validity of the technique, since the demonstrated effect will be partly dependent on the specific characteristics of the panel and the task.

A second major difference between laboratory Delphis and the Delphi ideal lies in the nature of the feedback presented to panellists (see Table 1). The feedback recommended in the 'classical' Delphi comprises medians or distributions plus arguments from panellists whose estimates fall outside the upper and lower quartiles (e.g., Martino, 1983). In the majority of experimental studies, however, feedback usually comprises only medians or means (e.g., Jolson & Rossow, 1971; Riggs, 1983; Sniezek, 1989, 1990; Hornsby et al., 1994), or a listing of the individual panellists' estimates or response distributions (e.g., Dalkey & Helmer, 1963; Gustafson et al., 1973; Fischer, 1981; Taylor, Pease & Reid, 1990; Leape et al., 1992; Ono & Wedermeyer, 1994), or some combination of these (e.g., Rohrbaugh, 1979). Although feedback of reasons or rationales behind individuals' estimates has sometimes taken place (e.g., Boje & Murnighan, 1982; Larreché & Moinpour, 1983; Erffmeyer & Lane, 1984; Rowe & Wright, 1996) this form of feedback is rare. This variability of feedback, however, must be considered a crucial issue, since it is through the medium of feedback that information is conveyed from one panellist to the next, and by limiting feedback one must also limit the scope for improving panellist aggregate accuracy. The issue of feedback will be considered more fully in the next section: suffice it to say here that, once more, it is apparent that simplified laboratory versions of Delphi tend to differ significantly from the ideal in such a way that they may limit the effectiveness of the technique.

From the discussion so far it has been suggested that the majority of the recent studies that have attempted to address Delphi effectiveness have discovered little of general consequence about Delphi. That research has yielded so little of substance would suggest that there are definitional problems with Delphi that militate against standard administrations of the procedure (e.g., Hill & Fowles, 1975). Furthermore, the lack of concern of experimenters in following even the fairly loose definitional guidelines that do exist suggests that there is a general lack of appreciation of the importance of factors like the nature of feedback in contingently influencing the success of procedures like Delphi. The requirement of empirical social science research to use simplification and reductionism in order to study highly complex phenomena seems to be at the root of the problem, such that experimental Delphis have tended to use artificial tasks (that may be easily validated), student subjects, and simple feedback in the place of meaningful and coherent tasks, experts/professionals, and complex feedback. But simplification can go too far, and in the case of much of the Delphi research it would appear to have done so.

In order that research is conducted more fruitfully in the future, two major recommendations would appear warranted from the above discussion. The first is that a more precise definition of Delphi is needed in order to prevent the technique from being misrepresented in the laboratory. Issues that particularly need clarification include: the type of feedback that constitutes Delphi feedback; the criteria that should be met for convergence of opinion to signal the cessation of polling; the selection procedures that should be used to determine number and type of panellists, and so on. Although suggestions on these issues do exist in the literature, the fact that they have largely been ignored by researchers suggests that they have not been expressed as explicitly or as strongly as they might have been, and that the importance of definitional precision has been underestimated by all.

The second recommendation is that a much greater understanding is required of the factors that might influence Delphi effectiveness, in order to prevent the technique's potential utility from being underestimated through accurate representations used in unsympathetic experimental scenarios. It would seem likely that such research would also help in clarifying precisely what those aspects of Delphi administration are that need more specific definition. For example, if 'number of panellists' was empirically demonstrated to have no influence on the effectiveness of variations of Delphi, then it would be clear that no particular comment or qualifier would be required on this topic in the technique definition.

7. Findings of process studies

Technique-comparison studies ask the question: 'does Delphi work?', and yet they use technique

forms that differ from one study to the next and that often differ from the ideal of Delphi. Process studies ask: 'what is it about Delphi that makes it work?', and this involves both asking and answering the question: 'what is Delphi?'

We believe that the emphasis of research should be shifted from technique-comparison studies to process studies. The latter should focus on the way in which an initial statistical group is transformed by the Delphi process into a final round statistical group. To be more precise, the transformation of statistical group characteristics must come about through changes in the nature of the panellists, which are reflected by changes over rounds in their judgments (and participation), which are manifest as changes in the individual accuracy of the 'group' members, their judgmental intercorrelations, and indeed – when panellists drop out over rounds – even their group size (as per Hogarth, 1978).

One conceptualisation of the problem is to view judgment change over rounds as coming about through the interaction of internal factors related to the individual (such as personality styles and traits), and external factors related to technique design and the general environment (such as the influence of feedback type and the nature of the judgment task). 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 judgment, and which are related to change in the desired direction (i.e., towards more accurate judgment). Only by understanding such relationships will it be possible to explain the contingent differences in judgment-enhancing utility of the various available procedures which include Delphi. Importantly from this perspective – and unlike in technique-comparison studies – analysis should be conducted at the level of the individual panellist, rather than at the level of the nominal group, since the non-interacting nature of Delphi ensures that individual responses and changes are central and measurable and (unlike interacting groups) the ultimate group response is effectively 'related to the sum of the parts' and is not holistically 'greater than' these (e.g., Rowe et al., 1991). The results of those few studies that we consider to be of the process type are collated in subsequent sections.

7.1. The role of feedback

Feedback is the means by which information is passed between panellists so that individual judgment may be improved and debiasing may occur. But there are many questions about feedback that need to be answered. For example, which individuals change their judgments in response to Delphi feedback – the least confident, the most dogmatic, the least expert? What is it about feedback that encourages change? Do different types of feedback cause different types of people to change? And so on. Although studies such as those of Dalkey and Helmer (1963) and Dalkey et al. (1970) would appear to have examined the role of feedback, this is not the case, since feedback is confounded by the unknown role of 'iteration', and since only single feedback formats were used in each study. What studies like those of Scheibe et al. (1975) do reveal, however, is the powerful influence of feedback: in their case, by demonstrating the pull that even *false* feedback can have on panellists. Evidence from most Delphi studies shows that convergence towards the 'group' average is typical (see section on Consensus).

However, more systematic efforts at assessing the role of feedback have been attempted by other authors. Parenté et al. (1984), in a series of experiments, used student subjects to forecast 'if' and 'when' a number of newsworthy events would occur. In their third study they attempted to separate out the iteration and feedback components by decomposing Delphi. They appeared to demonstrate that, although neither iterated polling nor consensus feedback had any apparent effect upon group 'if' accuracy, iteration *alone* resulted in increased accuracy for 'when' a predicted event would occur – while feedback alone did not. Similarly, Boje and Murnighan (1982) found decreased accuracy over rounds for a 'standard' Delphi procedure, while a control procedure that simply entailed the iteration of the questionnaire (with no feedback) resulted in increased accuracy (perhaps because of the opportunity for individuals to further reflect on the problem). However, as noted earlier, Delphi studies have tended to involve only superficial types of feedback. In Parenté et al. (1984), the feedback comprised modes and medians,

while in the study of Boje and Murnighan (1982), individuals' justifications for their estimates were used as feedback along with the estimates themselves (no median or other average statistic was used).

If different types of feedback differentially influence individuals to change their judgments over rounds, then comparisons across the various Delphi versions become difficult. Unfortunately, there are a number of reasons – empirical, intuitive and theoretical – to suggest that such differential effects do exist. In Rowe and Wright (1996) we compared 'iteration', 'statistical' feedback (means and medians) and 'reasons' feedback (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 when receiving 'reasons' feedback than the other types, when they *did* change their forecasts they tended to become more accurate – which was not the case in the 'iteration' and 'statistical' conditions.

That feedback of a more profound nature can improve the performance of Delphi groups has also been shown by Best (1974). He found that, for one of two task items, a Delphi group that was given feedback of reasons in addition to a median and range of estimates was significantly more accurate than a Delphi group that was simply provided with the latter information. Gowan and McNichols (1993) considered the relative influences of three types of Delphi feedback: statistical, regression model, and if-then rules. They could not measure the accuracy of judgments (the task involved considering the economic viability of companies on the basis of financial ratios), but they did find that (if-then) rule feedback induced a significantly greater degree of 'consensus' than the other types.

That different feedback types are differentially related to factors such as accuracy, change, and consensus should be of no surprise. Indeed, in the field of social psychology there has been much research on judgment, conformity, and opinion change in interacting groups, and how these relate to the type of information exchanged (e.g., see Deutsch & Gerard, 1955; Myers, 1978; Isenberg, 1986), but such concerns seem not to have infiltrated the Delphi domain. Although the use of 'reasons' in feedback, as prescribed in the 'classical' definition, might lead to improved Delphi performance, this is not to say that 'proper' Delphi feedback is liable to be the best type – merely that it is liable to lead to trends in judgment change that will differ from those in many evaluative studies of 'Delphi'. Indeed, given the limited nature of recommended feedback in the classic procedure, the question remains as to how effective even that type will be, given that the majority of individual panellists are allowed so little input (e.g., those within the upper and lower quartiles are not required to justify their estimates, even though they might be made for different – and even mutually incompatible – reasons).

7.2. The nature of panellists

A number of studies have considered the role of Delphi panellists and how their attributes relate to criteria such as the effectiveness (e.g., accuracy) of the procedure. One of the main attributes of panellists is their 'expertise' or 'knowledgeability'. As discussed earlier, Delphi is intended for use by disparate experts, and yet most empirical studies have used inexpert (often student) panels. Intuitively, the use of experts makes sense, but what does research show? A number of 'process' studies have considered self-rated expertise, essentially to determine whether self-ratings might be useful for selecting panels. Perhaps unsurprisingly – given the number of ways in which self-ratings may be taken – results have been equivocal, with a number of studies suggesting that self-assessment is a valid procedure (e.g., Dalkey et al., 1970; Best, 1974; Rowe & Wright, 1996), and others suggesting that it is not (Brockhoff, 1975; Larreché & Moinpour, 1983). But these results say more about the utility of a particular expertise measure than the role of experts in Delphi.

Jolson and Rossow (1971), however, found that accuracy increased over rounds for expert groups but not for inexpert ones (though they used only two groups making judgments on only two problems). Similarly, Riggs (1983) found that panels were more accurate in predicting the result of a college football game on which they had more information (i.e., were more 'expert') than one on which they had less

(although a more interesting comparison might have been of the relative improvement in the two cases). Rowe and Wright (1996) found that the most accurate Delphi panellists on first rounds changed their estimates less over rounds than those who were initially less accurate (and hence who were, arguably, less 'expert'). Importantly, this result appears to support the Theory of Errors (see, for example, Parenté & Anderson-Parenté, 1987), in which accuracy is improved over rounds as a consequence of the panel experts 'holding-out', while the less-expert panellists 'swing' towards the group average. The utility of expertise has been reviewed elsewhere, with evidence suggesting that there is an interaction between expertise and the nature of the task, so that expertise is only helpful up to a certain level for *forecasting* tasks, but of greater importance for estimation tasks (e.g., Welty, 1974; Armstrong, 1985). More controlled experiments are required to examine how expertise interacts with aspects of the Delphi technique, and how it relates to accuracy improvement over rounds.

Panellist confidence has been studied from a number of perspectives. One conceptualisation of confidence is as an outcome measure. For example, Sniezek (1992) has pointed out that panel confidence may be the only available measure of the quality of a decision (e.g., since one cannot determine the accuracy of forecasts a priori), and from this sense it is important that 'confidence' in some way correlates to other quality measures. Although Rowe and Wright (1996) found that both confidence and accuracy increased over rounds, they found no clear relationship between the accuracy and confidence of the *individual* panellists. Conversely, Boje and Murnighan (1982) found that while confidence increased over rounds, accuracy decreased. Sniezek has also compared panellist confidence to accuracy with contradictory results (evidence for a positive relationship in Sniezek, 1989; but no evidence for such a relationship in Sniezek, 1990). Dietz (1987) attempted to weight panellist estimates according to how confident they were, but found that such a weighting process gave less accurate results than a standard equal-weighting one.

From these inconsistent results in Delphi studies, we conclude that the use of confidence as a measure of quality is generally inappropriate (Armstrong, 1985, reviews studies on 'confidence' more broadly). A more interesting conceptualisation of confidence is as a potential predictor of panellists' propensity to change their estimates in the face of feedback. Scheibe et al. (1975) found a positive relationship between these factors (confidence and change), but Rowe and Wright (1996) found no evidence for this. We therefore have no consistent evidence that initial confidence explains judgment change over Delphi rounds.

Two studies have considered the impacts of a number of personality factors on the Delphi process. Taylor et al. (1990) found no relationship between four demographic factors (e.g., gender and education) and the 'effectiveness' of Delphi, or whether panellists dropped out. Mulgrave and Ducanis (1975) considered panellist dogmatism, finding that the most dogmatic panellists changed judgments the most over rounds – although the authors had no explanation for this rather counter-intuitive result and reported no statistical analysis.

The impact of the *number* of panellists has been considered by Brockhoff (1975) (who used groups of five, seven, nine, and 11) and Boje and Murnighan (1982) (using groups of three, seven, and 11). Neither of these studies found a consistent relationship between panel size and effectiveness criteria.

8. Conclusion

This paper reviews research conducted on the Delphi technique. In general, accuracy tends to increase over Delphi rounds, and hence tends to be greater than in comparative staticized groups, while Delphi panels also tend to be more accurate than unstructured interacting groups. The technique has shown no clear advantages over other structured procedures.

Various difficulties exist in research of this technique-comparison type, however. Our main concern is with the sheer variety of technique formats that have been used as representative of Delphi, varying from the technique 'ideal' (and from each other) on aspects such as the type of feedback used and the nature of the panellists. If one uses a technique

format that varies from the 'ideal' on a factor that is shown to influence the performance of the technique, then one is essentially studying a *different* technique. One tentative conclusion from this is that the recommended manner of comprising Delphi groups (as commonly used in real-world applications, e.g. Martino, 1983) may well lead to *greater* enhancement of accuracy/quality than might be expected to arise from the typical laboratory panel. In this case, it is possible that potential exists for Delphi (conducted properly) to produce results far superior to those that have been demonstrated by research.

Indeed, the critique of technique-comparison studies is arguably pertinent to wider research concerned with determining which is the best of the various potential judgment-enhancing techniques. We believe that the focus of research needs to shift from comparisons of vague techniques, to studies of the processes related to judgment change and accuracy *within* groups. We suggest there should be more research on the role of feedback in Delphi, and how aspects of the task, the measures, and the panellists interact to determine how first round Delphi groups are transformed to final round groups. We need to understand the underlying processes of techniques before we can hope to determine their contingent utilities.

References

Armstrong, J. S. (1985). Long range forecasting: from crystal ball to computer, 2nd ed., Wiley, New York.

Armstrong, J. S., & Lusk, E. J. (1987). Return postage in mail surveys: a meta-analysis. *Public Opinion Quarterly 51*(2), 233–248.

Ashton, R. H. (1986). Combining the judgments of experts: how many and which ones? *Organizational Behaviour and Human Decision Processes 38*, 405–414.

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

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

Bolger, F., & Wright, G. (1994). Assessing the quality of expert judgment: issues and analysis. *Decision Support Systems 11*, 1–24.

Boje, D. M., & Murnighan, J. K. (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: Linstone, H., & Turoff, M. (Eds.), The Delphi method: techniques and applications, Addison-Wesley, London.

Brockhoff, K. (1984). Forecasting quality and information. *Journal of Forecasting 3*, 417–428.

Cohen, J. (1994). The Earth is round ($p < .05$). *American Psychologist December*, 997–1003.

Dagenais, F. (1978). The reliability and convergence of the Delphi technique. *The Journal of General Psychology 98*, 307–308.

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

Dalkey, N. C., & Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. *Management Science 9*, 458–467.

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *Journal of Abnormal and Social Psychology 51*, 629–636.

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

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

Erffmeyer, R. C., Erffmeyer, E. S., & Lane, I. M. (1986). The Delphi technique: an empirical evaluation of the optimal number of rounds. *Group and Organization Studies 11*(1), 120–128.

Erffmeyer, R. C., & Lane, I. M. (1984). Quality and acceptance of an evaluative task: the effects of four group decision-making formats. *Group and Organization Studies 9*(4), 509–529.

Felsenthal, D. S., & Fuchs, E. (1976). Experimental evaluation of five designs of redundant organizational systems. *Administrative Science Quarterly 21*, 474–488.

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.

Gowan, J. A., & McNichols, C. W. (1993). The efforts of alternative forms of knowledge representation on decision making consensus. *International Journal of Man–Machine Studies 38*, 489–507.

Gustafson, D. H., Shukla, R. K., Delbecq, A., & Walster, G. W. (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.

Helmer, O. (1975). Foreward. In: Linstone, H., & Turoff, M. (Eds.), The Delphi method: techniques and applications, Addison-Wesley, London.

Hill, K. Q., & Fowles, J. (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 Behaviour and Human Performance 21*, 40–46.

Hornsby, J. S., Smith, B. N., & Gupta, J. N. D. (1994). The impact of decision-making methodology on job evaluation outcomes. *Group and Organization Management 19*, 112–128.

Hudak, R. P., Brooke, P. P., Finstuen, K., & Riley, P. (1993). Health care administration in the year 2000: practitioners' views of future issues and job requirements. *Hospital and Health Services Administration 38*(2), 181–195.

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

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

Kastein, M. R., Jacobs, M., Van der Hell, R. H., Luttik, K., & Touw-Otten, F. W. M. M. (1993). Delphi, the issue of reliability: a qualitative Delphi study in primary health care in the Netherlands. *Technological Forecasting and Social Change 44*, 315–323.

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

Leape, L. L., Freshour, M. A., Yntema, D., & Hsiao, W. (1992). Small group judgment methods for determining resource based relative values. *Medical Care 30*, NS28–NS39.

Linstone, H. A. (1975). Eight basic pitfalls: a checklist. In: Linstone, H., & Turoff, M. (Eds.), The Delphi method: techniques and applications, Addison-Wesley, London.

Linstone, H. A. (1978). The Delphi technique. In: Fowles, J. (Ed.), Handbook of futures research, Greenwood Press, Westport, CT.

Linstone, H. A., & Turoff, M. (1975). The Delphi method: techniques and applications, Addison-Wesley, London.

Lock, A. (1987). Integrating group judgments in subjective forecasts. In: Wright, G., & Ayton, P. (Eds.), Judgmental forecasting, Wiley, Chichester.

Lunsford, D. A., & Fussell, B. C. (1993). Marketing business services in central Europe. *Journal of Services Marketing 7*(1), 13–21.

Martino, J. (1983). Technological forecasting for decision making, 2nd ed., American Elsevier, New York.

Miner, F. C. (1979). A comparative analysis of three diverse group decision making approaches. *Academy of Management Journal 22*(1), 81–93.

Mulgrave, N. W., & Ducanis, A. J. (1975). Propensity to change responses in a Delphi round as a function of dogmatism. In: Linstone, H., & Turoff, M. (Eds.), The Delphi method: techniques and applications, Addison-Wesley, London.

Myers, D. G. (1978). Polarizing effects of social comparisons. *Journal of Experimental Social Psychology 14*, 554–563.

Neiderman, F., Brancheau, J. C., & Wetherbe, J. C. (1991). Information systems management issues for the 1990s. *MIS Quarterly 15*(4), 474–500.

Olshfski, D., & Joseph, A. (1991). Assessing training needs of executives using the Delphi technique. *Public Productivity and Management Review 14*(3), 297–301.

Ono, R., & Wedermeyer, D. J. (1994). Assessing the validity of the Delphi technique. *Futures 26*, 289–304.

Parenté, F. J., Anderson, J. K., Myers, P., & O'Brien, T. (1984). An examination of factors contributing to Delphi accuracy. *Journal of Forecasting 3*(2), 173–182.

Parenté, F. J., & Anderson-Parenté, J. K. (1987). Delphi inquiry systems. In: Wright, G., & Ayton, P. (Eds.), Judgmental forecasting, Wiley, Chichester.

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.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin 85*(1), 185–193.

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

Rowe, G., Wright, G., & Bolger, F. (1991). The Delphi technique: a re-evaluation of research and theory. *Technological Forecasting and Social Change 39*(3), 235–251.

Sackman, H. (1975). Delphi critique, Lexington Books, Lexington, MA.

Saito, M., & Sinha, K. (1991). Delphi study on bridge condition rating and effects of improvements. *Journal of Transport Engineering 117*, 320–334.

Scheibe, M., Skutsch, M., & Schofer, J. (1975). Experiments in Delphi methodology. In: Linstone, H., & Turoff, M. (Eds.), The Delphi method: techniques and applications, Addison-Wesley, London.

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*(1), 5–19.

Sniezek, J. A. (1992). Groups under uncertainty: an examination of confidence in group decision making. *Organizational Behavior and Human Decision Processes 52*(1), 124–155.

Spinelli, T. (1983). The Delphi decision-making process. *Journal of Psychology 113*, 73–80.

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

Taylor, R. G., Pease, J., & Reid, W. M. (1990). A study of survivability and abandonment of contributions in a chain of Delphi rounds. *Psychology: A Journal of Human Behavior 27*, 1–6.

Van de Ven, A. H., & Delbecq, A. L. (1974). The effectiveness of nominal, Delphi, and interacting group decision making processes. *Academy of Management Journal 17*(4), 605–621.

Van Dijk, J. A. G. M. (1990). Delphi questionnaires versus individual and group interviews: a comparison case. *Technological Forecasting and Social Change 37*, 293–304.

Welty, G. (1974). The necessity, sufficiency and desirability of experts as value forecasters. In: Leinfellner, W., & Kohler, E. (Eds.), Developments in the methodology of social science, Reidel, Boston.

Wright, G., Lawrence, M. J., & Collopy, F. (1996). The role and validity of judgment in forecasting. *International Journal of Forecasting 12*, 1–8.

Biographies: Gene ROWE is an experimental psychologist who gained his PhD from the Bristol Business School at the University of the West of England (UWE). After some years at UWE, then at the University of Surrey, he is now at The Institute of Food Research (IFR), Norwich. His research interests have ranged from expert systems and group decision support, to judgment and decision making more generally. Lately, he has been involved in research on risk perception and public participation mechanisms in risk assessment and management.

George WRIGHT is a psychologist with an interest in the judgmental aspects of forecasting and decision making. He is Editor of the *Journal of Behavioral Decision Making* and an Associate Editor of the *International Journal of Forecasting* and the *Journal of Forecasting*. He has published in such journals as *Management Science*, *Current Anthropology*, *Journal of Direct Marketing*, *Memory and Cognition*, and the *International Journal of Information Management*.