IV.A. Introduction

HAROLD A. LINSTONE and MURRAY TUROFF

Skeptics from the allegedly "hard" sciences have at times considered Delphi an unscientific method of inquiry. Of course, the same attitude is often encountered in the use of subjective probability (even in the face of considerable mathematical theory developed to support the concept). The basic reason in each case is the subjective, intuitive nature of the input.

Yet Delphi is by no means unordered and unsystematic. Even in the Gordon-Helmer landmark Rand study of 1964, an analysis of certain aspects of the process itself was included.[1] The authors observed two trends: (1) For most event statements the final-round interquartile range is smaller than the initialround range. In other words, convergence of responses is more common than divergence over a number of rounds. (2) Uncertainty increases as the median forecast date of the event moves further into the future. Near-term forecasts have a smaller interquartile range than distant forecasts.

It was also observed in all early forecasting Delphis that a point of diminishing returns is reached after a few rounds. Most commonly, three rounds proved sufficient to attain stability in the responses; further rounds tended to show very little change and excessive repetition was unacceptable to participants. (Obviously this tendency should not unduly constrain the design of Policy Delphis or computerized conferencing which have objectives other than forecasting.)

We shall briefly review here some of the systematic evaluations made in recent years.

Dispersion as a Function of Remoteness of Estimate

Martino has analyzed over forty published and unpublished Delphi forecasts.[2] For every event the panel's median forecast dates (measured from the year of the exercise) and the dispersion were determined. A regression analysis was performed and the statistical significance presented in terms of the probability that the regression coefficient would be smaller than the value actually obtained if there were no trend in the data.

The results are quite clear-cut. The remoteness of the forecast date and the degree of dispersion are definitely related. The regression coefficient is in nearly all cases highly significant for a single panel addressing a related set of events. However, there is no consistent relation among different panels or within a panel when addressing unrelated events.

[1] T. J. Gordon and O. Helmer, "Report on a Long Range Forecasting Study," Rand Paper P-2982, Santa Monica, California, Rand Corporation, September 1964.

[2] J. P. Martino, "The Precision of Delphi Estimates," Technological Forecasting 1, No. 3 (1970), pp. 293-99.

Martino also finds that the dispersion is not sensitive to the procedure used: in cases where only a single best estimate year is requested the result is similar to that where 10 percent, 50 percent, and 90 percent likelihood dates are Stipulated.[3]

Distribution of Responses

Dalkey has analyzed the first-round responses by panels asked to respond to almanac-type questions, i.e., those with known numerical answers.[4] All responses to each question are standardized by subtracting the mean value and dividing by the standard deviation for that question. The resulting distribution of "standardized deviates" shows an excellent fit to a lognormal distribution. Martino has applied the same techniques to the TRW Probe II Delphi using 10 percent, 50 percent, and 90 percent likelihood dates for 1500 events.[5] Again there is a very good fit to a lognormal distribution

Optimism-Pessimism Consistency

Another interesting analysis on the TRW Probe II data was undertaken by Martino to ascertain whether a panelist tends to have a consistently optimistic or pessimistic bias to his responses.[6] With each respondent providing 10 percent, 50 percent, and 90 percent likelihood dates, three standardized deviates can be computed for each individual and a given event. Taking the means over all events of the standardized deviates for a given individual and likelihood, we find an interesting pattern. Most panelists are consistently optimistic or pessimistic with respect to the three likelihoods, i.e., there are relatively few cases where, say, the 10 percent likelihood is optimistic while the 50 percent and 90 percent likelihoods are pessimistic. Considering the totality of events the individual panelist tends to be biased optimistically or pessimistically with moderate consistency. However, the amount of the bias is not very great; an optimistic panelist is pessimistic in some of his responses and vice versa. In other words, each participant exhibits a standard deviation which is comparable to, or greater than, his mean.

[3] In the first case the interquartile range of best estimates was used, in the second case the 10 percent to 90 percent span was taken.

[4] N. C. Dalkey, "An Experimental Study of Group Opinion," Rand RM-5888-PR, Rand Corporation, Santa Monica, California, March 1969.

[5] J. P. Martino, "The Lognormality of Delphi Estimates," *Technological Forecasting* l, No. 4 (1970), pp. 355-58.

[6] J. P. Martino, "The Optimism/Pessimism Consistency of Delphi Panelists," *Technological Forecasting and Social Change* 2, No. 2 (1970), pp. 221-24.

Accuracy of Forecasts

An apparent indicator of the value of Delphi as a forecasting tool is its accuracy. Since the method was widely publicized only ten years ago, it is difficult to have sufficient hindsight perspective to evaluate its success by this measure. In any event caution is in order. The most accurate forecast is not necessarily the most useful one. Forecasts are at times most effective if they are self-fulfilling or self-defeating. The Forrester-Meadows World Dynamics model has been sponsored by the Club of Rome in the hope that it will act as an early warning system and prove to be a poor forecast. Delphi may be viewed similarly in terms of effectiveness.

We should also observe that long-range forecasts tend to be pessimistic and short-range forecasts optimistic. In the long term no solution is apparent; in the near term the solution is obvious but the difficulties of system synthesis and implementation are underestimated.[7] Thus in 1920 commercial use of nuclear energy seemed far away. By 1949 the achievement appeared reasonable and in 1964 General Electric estimated that fast breeder reactors should be available in 1970.[8] Today the estimate has moved out to the 1980s. The same pattern has been followed by the supersonic transport aircraft. Buschmann has formulated this behavior as a hypothesis and proposed an investigation in greater depth.[9] If this pattern is normal, forecasts should be adjusted accordingly, e.g., forecasts more than, say, ten years in the future brought closer in time and forecasts nearer than ten years moved out. Subsequently Robert Ament made a comparison between a 1969 Delphi study on scientific and technological developments and the 1964 Gordon-Helmer Rand study.[10] Focusing on those items forecast in both studies, he found that all items originally predicted to occur in years before 1980[11] were later shifted further into the future, i.e., the original year seemed optimistic by 1969. On the other hand, two-thirds of the items originally forecast to occur after 1980 were placed in 1969 at a date earlier than that estimated in the 1964 study. Thus we find evidence here, too, of Buschmann's suggested bias.

[7] The large cost overruns on advanced technology aerospace and electronics projects are evidence of this trend (see Chapter III, A).

[8] E. Jantsch, "Technological Forecasting in Perspective," OECD, Paris, 1967, p. 106.

[9] R. Buschmann, "Balanced Grand-Scale Forecasting," Technological Forecasting 1 (1969), p. 221.

[10] R. H. Ament, "Comparison of Delphi Forecasting Studies in 1964 and 1969," FUTURES, March 1970, p. 43.

[11] T. J. Gordon and H. R. Ament, "Forecasts of Some Technological and Scientific Developments and Their Societal Consequences," IFF Report R-6, September 1969.

Grabbe and Pyke have undertaken an analysis of Delphi forecasts of information-processing technology and applications.[12] Forecast events whose occurrence could be verified cover the time period 1968 to 1972. Although six different Delphi studies were used, eighty-two out of ninety forecasts covering this period were taken from one study-: the U.S. Navy Technological Forecast Project. The results appear to contradict the hypothesis that near-term forecasts tend to be optimistic. In this case information-processing advances forecast four to five years in the future occur sooner than expected by the panelists who were drawn largely from government laboratories. There is, of course, the possibility that these laboratories are not as close to the leading edge of technology in this field as industrial and university research and development groups. Alternatively, the meaning of "availability" of a technological application may be interpreted differently by the laboratory forecasters and by the authors of this article.

Delphi Statements

The statements which comprise the elements of a Delphi exercise inevitably reflect the cultural attitudes, subjective bias, and knowledge of those who formulate them. This was recognized by Gordon and Helmer a decade ago and led them to commence the first round with "blank" questionnaires. Every student knows that multiple-choice examinations require insight into the instructor's mode of thought as well as the substance of the questions. Misinterpretations of the given statements can arise in both superior and inferior students. Grabbe and Pyke present examples of good and poor Delphi statements.[13] Statements may be too concise, leading to excessive variations in interpretation, or too lengthy, requiring the assimilation of too many elements. Consequently, we would expect a constraint on the number of words leading to the widest agreement in interpretation. Salancik, Wenger, and Helfer have probed this question more deeply.[14] They use an information theory measure (bits) of the amount of information derivable from a distribution of responses to a Delphi statement to measure consensus and the number of words needed to describe an event as a measure of its complexity. The study uses a computer development and application Delphi study as a test case. The authors find a distinct relation between number of words used and amount of information obtained, i.e., agreement in forecast dates. Low and high numbers of words yield low consensus with medium-statement lengths producing the highest consensus. In the

[12] E. M. Grabbe and D. L. Pyke, "An Evaluation of the Forecasting of Information Processing Technology and Applications," Technological Forecasting and Social *Change 4, No.* 2 (1972), p. 143.
[13] *Ibid.*
[14] J. R. Salancik, W. Wenger, and E. Helfer, "The Construction of Delphi Event Statements," *Technological Forecasting and Social Change* 3, No. 1 (1971), pp. 65-73.

particular case considered, twenty to twenty-five words form the peak in the distribution. This study also finds that the more familiar respondents are with a specific computer application, the fewer words are needed to attain agreement. If many words are used, less information results as to the occurrence of a familiar event. On the other hand, a longer-word description raises the consensus level for unfamiliar events.

A corresponding pattern is found when expert respondents are compared to nonexperts. The latter develop increasing consensus with longer-event descriptions. The experts, however, come to very high consensus with moderate statement lengths (higher than the greatest nonexpert consensus) but fall to a very low level of agreement with long statements. Apparently the addition of words brings on an effect somewhat similar to that of disputations by Talmudic scholars about minutiae.

Basis for Respondents' Intuitive Forecast

Salancik has examined the hypothesis that the panelists in a forecasting Delphi assimilate input on *feasibility,* benefits, and potential costs of an event in an additive fashion to estimate its probable date of occurrence.[15] The subject of the test is again a panel forecast of computer applications. Separate coding of participants' reasons for their chosen dates in the three categories enables the author to make a regression analysis. The second-round median date is made a linear function of the number of positive and negative statements in each of the three categories. He finds that the multiple regression strongly supports the hypothesis. The more feasible, beneficial, or economically viable a concept is judged, the earlier it is forecast to occur. The three categories contribute about equally to the regression.

In a second study independent assessments of feasibility and benefits are rated for twenty computer applications and then combined to form the basis for a rank ordering. This ordering is then compared to the Delphi panelists' responses. Again the correlation supports the suggested model of Delphi input assimilation. This paper adds another beam of support to the idea that Delphi is a systematic and meaningful process of judgment synthesis.

[15] J. R. Salancik, "Assimilation of Aggregated Inputs into Delphi Forecasts: A Regression Analysis," Technological Forecasting and Social Change 5, No. 3 (1973), pp. 243-48.

Self-Rating of Experts

Dalkey, Brown, and Cochran tackle another aspect of Delphi: the expertise of the respondents.[16] With a given group we might consider two ways of improving its accuracy: iterating the responses and selecting a more expert subgroup. The latter process implies an ability to identify such a subgroup (e.g., by self-rating) and a potential degradation in accuracy due to the reduced group size. The authors stipulate a minimum subgroup size to counteract this degradation and they force a clear separation in self-ratings of low- and highexpertise subgroups. The experiments were carried out by the authors using 282 university students and verifiable almanac-type questions. The conclusions: (1) self-rating is a meaningful basis for identification of expertise, and (2) selection of expert subgroups improves the accuracy to a somewhat greater degree than does feedback or iteration.

One must raise the question whether an experiment based on almanac type questions serves as an adequate basis for a conclusion about the validity of self-ratings of expertise for forecasting Delphis. While the lognormality behavior exhibited a similar pattern for factual (almanac-type) and forecasting cases, this similarity might not carry over for self-ratings.

And there are other fascinating unanswered questions. Why do women rate themselves consistently lower than men? Should only the expert subgroup results be fed back to the larger group in the iteration process? How do age, education, and cultural background condition the response of individuals?

The four articles in this chapter provide us with further evaluations of the process. When we use Delphi to draw forth collective expert judgments, we are actually making two substitutions: (1) expert judgment for direct knowledge, and (2) a group for an individual. In the first article, Dalkey strives to develop some mathematically rigorous underpinnings, i.e., a start toward a theory of group estimation. It quickly becomes evident that we still have much to learn about this process. Dalkey emphasizes the concept of "realis m," or "track record," to describe the expert's estimation skill and the theory of errors for the group. But the final verdict on their applicability is by no means in.

Scheibe, Skutsch, and Schofer report on several highly instructive findings based on research in the application of Delphi to the derivation of explicit goals and objectives. Analysis of a Delphi goal-formulation experiment for urban systems planning yielded the following important results:

[16] N. Dalkey, B. Brown, and S. Cochran, "Use of Self-Ratings to Improve Group Estimates," Technological Forecasting 1, No. 3 (1970), pp. 283-91.

(1) The three-interval scaling methods used-simple ranking, a rating scale, and pair comparisons-give essentially equivalent scales. The rating scale is found to be most comfortable to use by the participants.

(2) Respondents are sensitive to feedback of the scores from the whole group and tend to move (at least temporarily) toward the perceived consensus.

(3) There is only a modest tendency for the degree of confidence of an individual with respect to a single answer to be reflected in movement toward the center of opinion, i.e., less confident members exhibit a somewhat larger movement in the second round.

(4) Stability of the distribution of the group's response along the interval scale over successive rounds is a more significant measure for developing a stopping criterion than degree of convergence. The authors propose a specific stability measure.

Next, Mulgrave and Ducanis discuss an experiment which focuses on the behavior of the dogmatic individual in successive Delphi rounds. Surprisingly, the high-dogmatism group exhibits significantly more changes than the lowdogmatism group. It is the authors' belief that the dogmatic individual looks to authority for support of his view. In the absence of a clearly defined authority, he views the median of the group response as a surrogate.

There clearly exists the possibility of an unnatural overconsensus. Conformists may "capitulate" to group pressures temporarily, on paper. It would be interesting to compare the behavior of such psychological types in a Delphi with that in a conventional committee.

Finally, Brockhoff examines a series of hypotheses on the performance of forecasting groups using the Delphi technique and face-to-face discussions in a Lockean context. He focuses on short-range forecasting and small homogeneous groups. Staff members of local banks trained in economics are queried about data concerning financial questions, banking, stock quotations, and foreign trade. Groups vary in size from eleven to four participants (the latter below the size considered minimal by Dalkey, Brown, and Cochran[17]). The Delphi process uses an interactive computer program for structuring the dialogue as well as computing intermediate and final results. The correlation of self-rating of expertise with individual or group performance, the relation between information exchange and group performance, and the relevance of almanac-type fact-finding questions for short-term forecasting analysis are among the questions examined. One may speculate whether the dogmatism aspect raised by Mulgrave and Ducanis plays a significant role in groups of the type used in Brockhoff's experiments.

For the reader the thrust of this chapter is that, to develop proper guidelines for its use, we can and should subject Delphi to systematic study and evaluation in (lie same

[17] *Ibid*.

way as has been the case with other techniques of analysis and communication. Much still needs to be learned!