Delphi Critique

Expert Opinion, Forecasting, and Group Process

Harold Sackman
The Rand Corporation

Lexington Books
D.C. Heath and Company
Lexington, Massachusetts
Toronto London

Library of Congress Cataloging in Publication Data

Sackman, Harold.
Delphi critique.

Bibliography: p.
1. Decision-making. 2. Consensus (Social Sciences) 3. Questionnaires. 4. Forecasting. 5. Public opinion polls. I. Title.
HM73.S23 001.4'33 74-14858
ISBN 0-669-96156-0

Published simultaneously in Canada.

Printed in the United States of America.

International Standard Book Number: 0-669-96156-0

Library of Congress Catalog Card Number: 74-14858

Contents

List of Figures

List of Tables

1 Scope and Direction of the Inquiry

Organizing a meaningful critique of Delphi presented many problems. After considering various alternatives, a four-step schema was adopted. First, raise the various types of definitive issues pertinent to a Delphi critique. Second, evaluate conventional Delphi against established professional standards for opinion questionnaires, and scientific standards for empirical experimentation with human subjects. Third, evaluate Delphi in terms of its unique assumptions, principles, and methodology. Finally, summarize basic conclusions and make recommendations concerning the future use of Delphi.

Scope of the Inquiry

Before proceeding with the critique, three caveats on the scope and limitations of this study should be made. Delphi has been used for a vast array of applications in business, science, education, medicine, and other areas, both broad and specialized. The total literature has been estimated to include several hundred titles; a substantial number of these are proprietary or otherwise inaccessible. The author has been able to examine approximately 150 Delphi studies conducted at Rand and elsewhere. (See the semiannotated listing of Delphi and related publications in the Appendix.) The author makes no claim to having examined all the literature, particularly all the applications literature.

The focus of this study is on Delphi principles and methodology. The literature that has been reviewed contains the basic writings of the originators and key practitioners of Delphi, both within and outside Rand, with critical coverage of Delphi principles, assumptions, and procedures. Evaluative inferences from methodology to application are admittedly based on illustrative examples rather than on direct examination of all relevant studies. The validity of such inferences should be judged on the coherence of arguments put forth and the representativeness of examples used.

Another constraint is the elusiveness of a fixed, universally agreed upon working definition of Delphi. Many variants have emerged, some departing widely from the Delphi procedure associated with its Rand origins. An attempt is made in the next chapter to present a definition and characterization of "conventional Delphi." The term "Delphi" in this report refers primarily to "conventional Delphi," which may or may not apply to Delphi variants, depending upon the issues and the context.

A third caveat is that this study does not compare Delphi systematically with competing techniques. A comparison of Delphi with such techniques as simulation, trend extrapolation, gaming, morphological models, scenarios, relevance trees, input-output tables, contextual mapping, brainstorming, dialectical planning, critical path methodology, etc. would require an independent review and evaluation of each of these techniques and the systematic comparison of each with the others for key objectives and application areas. Undoubtedly, such a comprehensive critical appraisal of the methodology of the entire field of forecasting and planning techniques is long overdue, for much the same reasons that in-depth Delphi critiques are overdue. (For an instructive initial comparison and rating of these and related techniques, and for an appreciation of the magnitude of the task, see Rosove 1967, and Sackman and Citrenbaum 1972). As desirable as such an undertaking might be, this evaluation is necessarily limited to a comparison of conventional Delphi with scientific questionnaire development and experimental methodology with human subjects, and to questioning many of the basic assumptions and methods of the technique as it is currently being applied.

Outline of the Evaluation

The next chapter sketches key methodological issues associated with the complete cycle of conventional or characteristic Delphi studies. The discussion proceeds in the subsequent chapter to an evaluation of Delphi against professional standards for social experimentation and for opinion questionnaires established by the American Psychological Association and other national professional organizations. Analysis of conventional Delphi indicates that it does not satisfactorily meet the numerous experimental and methodological standards cited for test design, item analysis, subject sampling, reliability, validity, administration, interpretation of findings, and warranted social use.

The main body of the critique reviews methodological principles and key assumptions associated with Delphi. This analysis reveals: considerable evidence that results based on opinions of laymen and "experts" are indistinguishable in most cases; aggregate raw opinion presented as systematic prediction; technical shortcomings, such as untested and uncontrolled halo effects in the application of Delphi questionnaires; unsystematic and nonreplicable definition, sampling, and use of "experts"; manipulated group suggestion rather than real consensus; ambiguity in results stemming from vague questions; acceptance of snap judgments on complex issues; and the virtual absence of a vigorous critical methodological literature, even though hundreds of Delphi studies have been published. The accuracy of the technique, in generating forecasts and other "expert" estimates is necessarily suspect as long as Delphi questions are not empirically linked to objective and independently verifiable external validation criteria. These liabilities are counterbalanced primarily by a popular demand for systematic expert opinion, and by the convenience, low cost, and simplicity of the method. It is argued that such advantages are inconsequential if the Delphi concept, method, and results are inherently untrustworthy.

The analysis concludes that conventional Delphi is basically an unreliable and scientifically unvalidated technique in principle and probably in practice. In the absence of a comprehensive survey of the extensive applications literature, it is suggested, but not proven, that the results of most Delphi experiments are probably unreliable and invalid. Even variations of conventional Delphi should not be encouraged unless they explicitly attempt to meet the challenge of generally accepted standards of rigorous empirical experimentation in the social sciences. Except for its possible value as an informal exercise for heuristic purposes, Delphi should be replaced by demonstrably superior, scientifically rigorous questionnaire techniques and associated experimental procedures with human subjects.

As the preferred alternative to conventional Delphi, professionals, funding agencies, and users are urged to work with social scientists with psychometric training who can apply rigorous questionnaire techniques and scientific human experimentation procedures tailored to their specific needs. The final recommendation is that conventional Delphi be dropped from institutional, corporate, and

government use until its principles, methods, and fundamental applications can be established experimentally as scientifically tenable.

2 Delphi Issues

This chapter identifies certain methodological issues and characterizes "conventional Delphi" for the purposes of this critique. The chronological framework for a Delphi study follows a problem-solving sequence: establishment of objectives, formulation of the problem, solution testing, and the write-up and dissemination of results. In the Delphi context, objectives include needs, goals, basic value assumptions, and expected payoffs. Formulation of the problem is accomplished through the design of the questionnaire and its experimental implementation. Solution testing includes iterative field administration and scoring of responses to the questionnaire. The last stage involves the interpretation of results by the Delphi director in communicating findings to others. Each stage is briefly examined to provide a chronological chain of methodological issues as a framework for this evaluation.

Delphi Objectives

Early Delphi studies at Rand were primarily concerned with scientific and technological forecasting. They were viewed as experiments with what was thought to be an interesting, and possibly useful, new technique. From these humble beginnings, Delphi has spread rapidly, with hundreds of studies appearing in the United States, accompanied by growing use in other countries, including extensive use in the United Kingdom (Currill 1972) and recent use in the Soviet Union (Martino 1973) and in Japan. Delphi applications have grown in all directions to include forecasting of many social phenomena, including human attitudes and values (Reisman et al. 1969), and even the "quality of life" (Dalkey, Rourke, Lewis and Snyder 1972). A large and growing roster of major firms have used Delphi for diverse purposes (see Appendix). Applications have expanded until, broadly considered, they are virtually indistinguishable from the questionnaire technique. Advocates, such as Turoff

(1971), have expanded the scope of Delphi as a general-purpose vehicle for distributed human communication and consensus, and for group problem solving. Delphi has been propelled at an increasingly accelerated rate into the general field of questionnaire design and development not only for "experts," but for nonexperts as well. The core question arises: How does Delphi rate in comparison with competing approaches in the well-established fields of questionnaire design and application in the social sciences?

The payoff of a Delphi study is typically a presentation of observed expert concurrence in a given application area where none existed previously. This assumes that participating panelists are experts in the subject area, and that the reported consensus was obtained through reliable and valid procedures. Proponents of Delphi (Dalkey 1969) stress three quintessential attributes that contribute to authentic consensus and valid results: anonymity of panelists, statistical response, and iterative polling with feedback. Is the trust placed in these central assumptions warranted?

In any decision to use Delphi, there are various cost-effectiveness considerations. How much does a Delphi study cost in time and effort for the director and panelists, and how are such investments related to the usefulness of the final results? An associated issue is the attractiveness of Delphi as a quick and easy way to solicit rational expert opinion in an unknown area. Do such positive payoffs exist?

Formulation of the Problem

The next step in a Delphi study is the formulation of the problem, the design of the questionnaire, and its application. How effectively is the area of inquiry defined and delimited by the Delphi investigator? Is there an effort to make questionnaires bias free? Are his assumptions spelled out? Are there explicit hypotheses, and are they operationally defined? Has the relevant literature been reviewed and systematically evaluated? Have baseline statistics and qualitative characteristics of the area of inquiry been documented and spelled out so that respondents derive their forecasts and opinions from a common specification of the current state of the art?

In developing the questionnaire, many technical considerations arise. Is the questionnaire an informal, ad hoc collection of items?

Or is it systematically designed as a standardized instrument to be administered under rigorously controlled conditions? How are the items constructed? How large was the original pool of items, how were they derived, and what pilot procedures were used for item analysis to prune them down to the final set used for the study? What psychometric scaling approach was selected (e.g., Thurstone, Likert, or Guttman psychometric scales, or econometric scales; see Pill 1971) and what factors determine the selection?

Then there are problems concerning the panelist sample to which the questionnaire is applied. What is an "expert" in the target application field, and how are such experts operationally defined? How many panelists are used? What are the expected levels of statistical precision of the results relative to planned sample size for the dispersion of responses anticipated? Can the selected panelist sample be systematically related to an objectively defined population with measurable sampling parameters? Is the choice of experts random or is it selective? Are sampling procedures rigorously defined (see Cochran 1963) relative to hypothesis testing for opinion polling?

Test and Interpretation

In administering the questionnaire, many problematic issues arise. How are dropouts handled in the results? Which items should be dropped, modified, or retained in their original form in successive Delphi rounds? What kind of feedback, how much feedback, and in what form should it be presented to panelists? When is the point of diminishing returns reached in successive iterations? How long should the intervals be between successive rounds, and how can participants be encouraged to respond promptly to expedite turnaround time? What is the tradeoff between more items and a longer form versus fewer items with less data in relation to study objectives? Does the director reinforce and encourage conformist or dissenting behavior in successive rounds? In working with distributed Delphi by mailed questionnaires and iterative polling, what opportunities exist for misusing the technique?

In the final stage of writeup and dissemination of results, the main problems center around the analysis and interpretation of findings. Should only descriptive results be presented, or should all

statistics be accompanied by standard errors of estimate, clearly indicating the empirical level of precision? Is it misleading to present only interquartile ranges in graphic portrayal of Delphi results, or should the full range and true dispersion of results also be presented? Should first-round results be presented showing the full dispersions of expert opinion? How strongly should the expert halo effect be exploited, or should it be controlled in evaluating results? Should the procedure and the interpretation give weight to adversary or consensus positions?

How strongly should procedural, administrative, statistical, and experimental limitations be stressed in the final publication? Are results put forth as scientific prediction or as conglomerate opinion? Has provision been made for replication testing or validity generalization in follow-on studies?

Definition of Conventional Delphi

The above review of the Delphi cycle provides a backdrop for the characterization of "conventional Delphi" as it is used in this critique. These are briefly described below under the categories of objectives, subjects, and techniques.

The application objective of conventional Delphi may be the forecasting of specified events, long-term or short-term; it may be the generation of quantitative estimates (costs, market demand, number of users, etc.) from a set of participants; or it may be aimed at qualitative evaluations (qualitative scales of agreement, disagreement, preferences among alternatives). The range of application objectives thus includes any type of quantitative or qualitative rating scale, and as such is coextensive with questionnaires broadly considered.

Other key objectives for conventional Delphi may be singled out, including consensus of participants and heuristic goals. The consensus intent of Delphi is typically oriented toward controlled and rational exchange of iterated opinion leading toward optimal convergence of opinion achievable within the framework of the technique. The heuristic objective views Delphi as an educational technique to help participants, the director, and users to explore a problem area more throughly, leading to greater insight on the target problem.

Conventional Delphi is primarily concerned with experts, but may also use other subject groups who may be informed to a greater or lesser extent in the target area of inquiry, but who do not qualify as experts. Although this report focuses on the Delphi concept of expert, it is also directed at the growing use of nonexperts. More broadly, this critique is concerned with the operational sampling procedures used in selecting Delphi subjects, expert or otherwise.

The technique category is the most detailed. Conventional Delphi, as used in this report, exhibits the following characteristics:

1. The format is typically, but not always, a paper and pencil questionnaire; it may be administered by mail, in a personal interview, or at an interactive, online computer console. The basic data-presentation and data-collection technique is the structured, formal questionnaire in each case.

2. The questionnaire consists of a series of items using similar or different scales, quantitative or qualitative, concerned with study objectives.

3. The questionnaire items may be generated by the director, participants, or both.

4. The questionnaire is accompanied by some set of instructions, guidelines, and ground rules.

5. The questionnaire is administered to the participants for two or more rounds; participants respond to scaled objective items; they may or may not respond to open-end verbal requests.

6. Each iteration is accompanied by some form of statistical feedback, which usually involves a measure of central tendency, some measure of dispersion, or perhaps the entire frequency distribution of responses for each item.

7. Each iteration may or may not be accompanied by selected verbal feedback from some participants, with the types and amounts of feedback determined by the director.

8. Individual responses to items are kept anonymous for all iterations. However, the director may list participants by name and affiliation as part of the study.

9. Outliers (upper and lower quartile responses) may be asked by the director to provide written justification for their responses.

10. Iteration with the above types of feedback is continued until convergence of opinion, or "consensus," reaches some point of diminishing returns, as determined by the director.

11. Participants do not meet or discuss issues face to face, and they may be geographically remote from one another.

It should be apparent that a one-sentence or even one-paragraph definition of "conventional Delphi" is not possible without leaving out many significant details and qualifications that receive substantial attention in this report. Generally speaking, the working definition of Delphi for this study embodies the "quintessential" model originating at Rand, with many related variations that more or less follow the iterative questionnaire format with anonymous statistical feedback.

This completes the review of issues raised by the conventional Delphi cycle and permits an evaluative comparison of Delphi with professional standards for opinion questionnaires and experimentation with human subjects.

3 Delphi Versus Social Science Standards

This section presents key standards in professional questionnaire design and use, and shows how Delphi measures up to them. The evaluative criteria are quoted from "Standards for Educational and Psychological Tests and Manuals," published by the American Psychological Association (1966). This publication was jointly prepared by a committee representing three national organizations: the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education. This committee worked over a period of five years in conjunction with numerous measurement specialists and test publishers.

The manual is currently undergoing revision under the auspices of the APA Office of Scientific Affairs. Public hearings on the proposed draft have been held in Washington, D.C. The provisional table of contents of the proposed version is shown in Table 3-1.

It can not be too strongly emphasized that these guidelines represent responsible efforts to establish exemplary scientific standards in a controversial area with a history of continuing abuse on the part of some test developers, and with a history of continuing misunderstanding and undereducation on the part of the public. Whether or not the reader identifies himself as a social scientist, he should be aware that there is a vast and highly germane literature reflecting an organized professional effort to serve the public interest.

Buros (1965), after dedicating a distinguished lifetime to professional quality control in the public domain for the testing field, concluded that only partial success is possible with the inevitable collusion between test promoters and a gullible public that expects far more from tests than they can possibly deliver. The carryover to Delphi, as this report shows, is more than mere coincidence. In the absence of any tradition with such guidelines, Delphi practitioners, participants, and users neglect such standards only at their own peril.

Some may still argue that Delphi is not a conventional test, though it usually assumes the form of an iterative paper-and-pencil questionnaire. As such, they argue, Delphi is exempt, and the APA guidelines should not apply. A cursory review of the selected items, however, reveals that the guidelines deal with bedrock questions concerning sampling, controls, reliability of measures, and criterion validity which are universal to all scientific experimentation with human subjects. If Delphi is to be treated seriously as a professional technique, it must be judged by basic, minimum standards applicable to all empirical social science.

The historical precursors of Delphi in the opinion-polling and social-psychological literature were most explicit in applying rigorous questionnaire design and sampling techniques against the methods and findings of their studies. Cantril (1938) and McGregor (1938), in independent studies on predictions of social events, emphasized the severe limitations of questionnaire format and procedures and in the representativeness of subject sampling for any generalizations of their results. Kaplan, Skogstad, and Girshick, in a landmark study on "The Prediction of Social and Technological Events" (1950) presented a detailed listing of sampling, reliability, and validity problems encountered in this field in relation to rigorous questionnaire and polling standards. As a direct historical offshoot of these pioneering efforts, the Delphi technique does not possess or warrant any special dispensation exempting it from such scientific standards.

Delphi proponents may protest that concern with experimental method in the application of Delphi questionnaires is "misguided" because Delphi is a tool, and a tool, once developed, does not have to be experimentally administered each time it is used. This may be fine for weighing scales, rulers, compasses, spectrometers, voltmeters, and other measurement instruments frequently used in the physical sciences. However, it does not apply to questionnaires, nor to paper-and-pencil testing broadly considered, nor to Delphi in particular. A questionnaire is reliable and valid only to the extent that it is administered under conditions that replicate the basic experimental controls under which it was originally designed, tested, and validated. This means that each administration of the questionnaire is viewed as an experimental replication of operationally designed conditions for individual response, collection of data, scoring, and interpretation. The layman's failure to realize that

Table 3-1
Standards for Development and Use of Educational and Psychological Tests

Table of Contents

questionnaire tests are replicated experiments leads to abuses of the technique, noncomparability of results, and a general increase in measurement-error variance.

Delphi iteration of questionnaires with feedback is a definitive empirical experimental procedure with human subjects in its own right. Neglect of standard experimental guidelines may lead to uncontrolled variations in results and inability to define, replicate, and validate method and findings. This neglect may be acceptable for an informal exploratory technique, but it is unacceptable for a rigorous

social science experiment. The compounding of methodological problems generated by an unscientific approach to the conduct of Delphi studies is described and illustrated in this section.

Scope of Standards

While the standards are quoted verbatim from the manual,[a] the author is fully responsible for the evaluative Delphi commentary. The manual covers paper-and-pencil testing broadly considered, and obviously, many of the standards do not pertain directly to Delphi. In what follows a representative subset of key standards relevant to Delphi is cited, accompanied by evaluative commentary. The citations cover introductory, interpretive, validity, reliability, and administrative/scoring standards, taken from applicable sections in the APA manual.

In the direct quotes that follow, material is reproduced verbatim except for one term. The word *manual* is replaced by *test documentation*. This is done because it was found that individuals unfamiliar with psychometrics found it difficult to understand the scope and intent of test "manuals." "Test documentation," which refers to test materials, instructions, controls, and reports of empirical results, norms, interpretations and recommendations for use, is less likely to cause unintentional confusion for the layman in relating the guidelines to Delphi.

In the introduction, the manual states:

> These recommended standards cover not only tests as narrowly defined, but also most published devices for diagnosis, prognosis, and evaluation. . . .
>
> The present standards apply to devices which are distributed for use as a basis for practical judgments rather than solely for research. Most tests which are made available for use in schools, clinics, and industry are of this practical nature (p. 3).

Conventional Delphi studies, as applied prognostications or as predictions of technological and social developments for a variety of end-users, fall under the general purview of the manual.

[a]Copyright 1966 by the American Psychological Association. Reprinted by permission. (Applies to all quotes in this chapter.)

Interpretive Standards

From the section on interpretations of findings, two items are selected. Ratings accompany each standard listed in the manual. Ratings are ESSENTIAL, VERY DESIRABLE, or DESIRABLE.

> B4.2. When the statistical significance of a relationship is reported, the statistical report should be in a form that makes clear the sensitivity or power of the significance test. ESSENTIAL (p. 11).

Statistical significance is rarely reported in Delphi studies, either for precision of estimates or for tests of the significance of mean or median differences between two or more forecasts. Consensus and precision are implied from suggestive graphs, not from standard errors of estimates. With small samples and large dispersions, many forecasts do not differ significantly from one another, but are shown to do so by implication if not by explicit statement.

> B4.4. The test documentation should state clearly what interpretations are intended for each subscore as well as for the total test. ESSENTIAL (p. 12).

This standard is especially pertinent to Delphi studies where forecasts are made on a broad and diverse target area. Each forecast should be individually and separately tested for dispersion of consensus, systematic correlations with other items, and for significance of forecasted differences against other items as is done with quantitative scores in conventional questionnaire item analyses (Anastasi 1968).

The author has never seen the full three-dimensional matrix of items versus panelists versus rounds analyzed by a common statistical vehicle, such as analysis of variance, to test for main and interaction effects. Nor are items compared for homogeneity of variance, linearity, and type of empirical frequency distributions for applying such tests. With small samples, interquartile Delphi graphs are no substitute for rigorous statistical testing of individual items and item subsets. This is not a pedantic frill; differential statistical reliability requires differential interpretation of findings.

Except for a study by Derian and Morize (1973), the author has not seen a factor analysis of Delphi items, also part of the standard repertoire in test-item analysis. Factor analysis is valuable for prun-

ing out redundant items that are highly intercorrelated or are "saying the same thing" by eliciting the same response from subjects. This type of item "padding" is thus hidden from the end-user who interprets results at face value.

If these interpretative standards were respected, quantitative Delphi findings would not be presented in simplistic, descriptive form to potential users. They would then not be taken at face value by users who are unaware of statistical and sampling limitations.

Empirical Validity

The next items are drawn from the "Validity" section of the APA manual. The keynote standard for this section:

C1. Test documentation should report the validity of the test for each type of inference for which it is recommended. If its validity for some suggested interpretation has not been investigated, that fact should be made clear. ESSENTIAL (p. 15).

This standard provides obvious protection for potential users of test results by requiring the test publisher to indicate whether his test rests on his (vested) opinion (face validity), indirect validity (e.g., correlations with related areas), or more direct forms of validity testing (e.g., empirical experimentation or real-world performance measurement). With Delphi, panel opinion is reported with little or no subsequent effort to test results against actual or related events (except for a small number of studies discussed later in this report). The results are usually simply aggregations of iterative opinions. For example, Gordon and Helmer (1964) went no further than to show medians and quartiles and some descriptive scatter-plots for their classic forecasting study, and Nanus, Wooten and Borko (1973) simply show frequency distributions and list some percentages for quantitative results in their study of the social impact of multinational computers. Measures of central tendency are put forth, however, as systematic and concurred forecasts of specified events by experts.

The Delphi method typically measures very small sample attitudes toward future events at a given time. It does not measure the events themselves, nor does it incorporate systematic hypotheses and empirical feedback from such events. The leap from raw opinion to future events under these conditions is strictly an act of faith.

The next selected standard is found under "Content Validity." It refers to item definition and item sampling.

C3. If a test performance is to be interpreted as a sample of performance or a definition of performance in some universe of situations, the test documentation should indicate clearly what universe is represented and how adequate is the sampling. ESSENTIAL (p. 15).

When an area of inquiry has been selected for a Delphi study, as a first step in determining content validity, has the area been adequately formulated and defined? We rarely find systematic reviews of application literature in Delphi studies leading to a careful, state-of-the-art definition of the target domain. Such reviews should extract the best of precursor studies and define basic assumptions and bounds of the inquiry. We often encounter an amorphous sociotechnological area (scientific advances, quality of life, etc.) where the universe of situations may be virtually indistinguishable from future society broadly considered.

The second step in determining content validity is demonstrating that the selected items comprising the questionnaire represent a systematic sampling of key elements of the target area of inquiry. If a particular problem area has been chosen for a Delphi forecast, has a taxonomy been developed for subproblems, embedding situations, resources, and classes of problem-solvers? If so, has it been used as the basis for a representative and comprehensive selection of items?

For example, in using Delphi to forecast computer developments, as was done in Parsons and Williams's widely cited study (1968), content validity preparation would call for a systematic taxonomy of hardware, software, peripheral equipment, communications and applications, perhaps along the lines of the classification scheme used by the Computing Reviews of the Association for Computing Machinery. If the entire computer field is to be covered, or some specified subset, the correspondence between final selected items and the specified area should be spelled out. Such taxonomies, and such accountability in matching items against the target universe, are rarely seen in the Delphi literature.

Standards for Use of Experts

The next two standards are the only references in the APA manual to the use of experts in test design and analysis. It should come as no

surprise that the social sciences have abandoned the use of experts as an integral part of scientific methodology. In test construction and analysis, the role of experts in generating and contributing questionnaire items to the initial item pool is well recognized and is consistent with current practice. However, the use of experts as the principal and exclusive method for validating tests has been discredited. For example, in World War II, the unreliable "expert" opinions of experienced, professional interviewers were dropped in favor of more effective standardized objective testing procedures. (See Thorndike's account (1949) of the Aviation Psychology Program of the Army Air Force in World War II.)

Another example of the use of experts in the field of economics is revealing. Zarnowitz (1965) studied eight independent forecasts of the gross national product from 1953 to 1963 derived from "expert" opinion. The average observed absolute error for experts was $10 billion, or about 2 percent of the GNP during this period. Zarnowitz found that simple arithmetic extrapolation of the increase occurring in the previous year yielded an average absolute error of $12 billion, effectively the same as the average expert prediction. Zarnowitz conducted studies of other economic indices and obtained similar results.

When we leave the area of short-term forecasting in economics, where extensive baseline statistical indicators are available, and enter the more nebulous areas of psychological and psychiatric diagnosis and prognosis, the record of expert clinical opinion is and has been in a state of disarray. In "The Discontent Explosion in Mental Health," Hersch (1969) explicated the bankruptcy in theory and practice of the unscientific use of clinical experts in empirical research on psychotherapy.

After reviewing some forty large-scale programs involving man-machine system experimentation in his comprehensive book covering the work in this area since World War II, Parsons (1972) concluded that the reliance of system designers on the opinions and preferences of "so-called expert system operators" is "foolhardy." He pointed out that such experts "may provide suggestive leads, but are not reliable guides, as demonstrated by their repeated disagreement with objective data" (p. 553). These examples illustrate the repeated failures and frustrations encountered in the use of experts in diverse social science areas.

C3.1. When experts have been asked to judge whether items are an appropriate sample of a universe or are correctly scored, the test documentation should describe the relevant professional experience and qualifications of the experts and the directions under which they made their judgments. VERY DESIRABLE (p. 15).

Delphi exercises guarantee anonymity of individual responses to encourage free expression of opinion. Some studies list the names of panelists and, in fewer cases, list their professional affiliations. The author was not able to find any studies listing professional training and scaled experience levels qualifying each individual as possessing the skills required to meet an objective criterion as an "expert." This "very desirable" standard is effectively neglected in Delphi practice.

C3.11. When the items are selected by experts, the extent of agreement among independent judgments should be reported. DESIRABLE (p. 16).

This standard makes an explicit distinction between independent and dependent expert judgment, which goes to the heart of Delphi iteration "with feedback." The first round is basically designed to secure independent expert judgment. The second and successive rounds produce strictly correlated, or biased, judgments. The use of standardized statistical techniques for hypothesis testing based on random sampling assumptions, which may offer no major problems for independent first-round judgments, becomes difficult and problematic in successive rounds, a methodological shortcoming apparently unnoticed by Delphi practitioners. All rationalizations about reconsidering, incorporating new information, and converging toward consensus can not hide the fact that independent judgment is destroyed once the participant knows how others have responded to each item. If Delphi can make no claims concerning independent expert opinion, does Delphi feedback develop insight into the issues for improved collective judgment in successive rounds?

Theoretical Standards

The next standard refers to long-term predictions and overlaps substantively with the notion of forecasting.

C4.41 . If a test is recommended for long-term predictions, but comparisons with concurrent criteria only are presented, the test documentation should emphasize that the validity of predictions is undetermined. ESSENTIAL (pp. 17-18).

Delphi practice essentially neglects long-term longitudinal validation and typically dissociates itself from any systematic comparisons with even second-string concurrent criteria (e.g., short-term interpretations of long-term trends). Panelists often disagree over what exists "today," and with rare exceptions, Delphi practitioners make no effort to present panelists with a precise report on "where we are" to establish a baseline for projections into the future. On both counts, for this "essential" standard, Delphi forecasting results should be explicitly presented to potential users as conjectures of undetermined validity.

Delphi practitioners object to this conclusion, pointing out that Delphi has been proven "valid" and "accurate" in a few relatively recent studies involving almanac-type items (Dalkey 1969) and for relatively short-term predictions (Martino 1972). Established almanac items (e.g., population of a city or gross national income at a particular point in time) are not in any substantive way generalizable to long-range forecasts. What they share in common is the trivial property that we all can exercise opinions on each item, hardly a sound basis for generalizing from simple descriptive facts anchored in the past to complex events in the future.

Martino (1972) reports forthcoming work comparing earlier Delphi predictions with outcomes. The original estimates, as in the Gordon, Helmer study (1964), were derived from pooled respondent opinion, and the outcomes were also determined by pooled opinion. The abuses of such a post hoc subjective approach should be obvious, leaving the central issue of Delphi validity and accuracy unresolved.

The next standard applies to identification of the characteristics of participating panelists.

C5.2. The validity sample should be described in test documentation in terms of those variables known to be related to the quality tested, such as age, sex, socioeconomic status, and level of education. Any selective factor determining the composition of the sample should be indicated. ESSENTIAL (p. 19).

Delphi studies, having promised anonymity to participants, typically do not report key population characteristics of panelists such as those cited in this standard. Such specification of "expert" samples would permit more effective evaluation of the adequacy of the expert sample. For example, a long-range forecasting study might benefit from inputs from relatively youthful panelists who are more likely to be living in, and directly shaping, the world they are forecasting; lower-class or minority members, if the socioeconomic items cut across their future; more women panelists, if they are underrepresented; (Dalkey [1969], Borko [1970], and Bedford [1972] have shown systematic quantitative and qualitative differences by sex in Delphi responses); wider geographical distribution of panelists, if they are concentrated in one or two locales. The author has not encountered any studies where panelists have been asked to provide detailed personal data for sampling profiles. Anonymity can still be honored if panelist characteristics are presented as statistical aggregates.

The next standard applies particularly to the pitfalls inherent in the voluntary participation of Delphi panelists.

C5.3. If the validity sample is made up of records accumulated haphazardly or voluntarily submitted by test users, this fact should be stated in the test documentation, and the test users should be warned that the group is not a systematic or random sample of any specifiable population. Probable selective factors and their presumed influence on test variables should be stated. ESSENTIAL (p. 19).

Panelist dropout is one of the well-known hazards of Delphi. Delphi dropout rates are probably quite high. Although he cited no empirical data, Martino (1972) asserted that response rates to first-round questionnaires "typically ran 50 percent or less." In the only study the author has been able to find on Delphi dropouts, Bedford (1972) noted that dropouts in a study on home communication services were less motivated to participate in the study (responded to fewer questionnaire items), and more significantly, dropouts were considerably more critical of the overall study, the utility of questionnaire items, and relative stress placed on various factors such as "lack of concern for sociological and psychological considerations."

There is no question but that some selective factors operate to

determine the hard-core group that sticks with the study through all iterations. The reasons may be positive, such as strong motivation and interest in the target area, or negative, such as a high proportion of personal acquaintances of the director, or of those in his professional circle. Perhaps those who disagree strongly with the design and content of the questionnaire, and those who question initial results (as in Bedford's study), drop out more often than those who have confidence in the study and the procedure, or who play along with minimum effort. To the extent that any systematic panelist sampling effects are known, they should be stated explicitly and taken into account in the evaluation of results. If the original expert sampling is unknown, and if the dropout rate is also unknown, the sample on which the final results are based is doubly suspect. This double indemnity is probably the rule, not the exception for Delphi studies.

A recent memo sent to me by Brownlee Haydon illustrates the possibilities of serious social abuse of conventional Delphi in picking a stacked panel of experts in a controversial area with major vested interests.

If you are a regular reader of *The New Yorker*, you may already have seen the series entitled "Annals of Industry—Casualties of the Workplace" currently appearing in that magazine. The November 12, 1973 installment describes a classic case of the misuse or perversion of the Delphi process.

As I read it, Arthur D. Little, Inc. has undertaken for the Department of Health, Education and Welfare (Occupational Safety and Health Administration) to use the Delphi method to arrive at a consensus on the proper level of exposure to asbestos fibres (2, 5, 12, 30 fibres of greater than 5 microns length per cubic centimeter of air) to be established as a government safety standard. What is almost unbelievable is the choice of "experts"—apparently members of the asbestos manufacturing community and their "medical experts" along with a few (too few) independent medical researchers in the field of asbestos-induced cancer!

Dr. Selikoff was the only member of the expert health panel in the Delphi study who had not been a paid consultant of, or whose investigations into asbestos-related disease had not been supported by, some segment of the asbestos industry. In this *New Yorker* article by Brodeur (1973), Dr. Selikoff indicated that there was no sense in guessing about the biological effects of asbestos when mortality studies of asbestos workers had already shown the effects.

The next standard appears under the section concerned with

"construct validity," which refers to the interpretation of theoretical constructs on which tests are based. This standard raises the key issue of accountability for the interpretation of Delphi results.

C7.1. The test documentation should indicate the extent to which the proposed interpretation has been substantiated and should summarize investigations of the hypotheses derived from the theory. ESSENTIAL (p. 23).

This requirement is largely ignored in Delphi practice, where a descriptive approach characterizes the presentation of results. The reasons, theories, and hypothetical constructs of expert panelists are covert, rather than overt. Panelists are asked for opinions, and the occasional rationale from panelists is typically very brief, uneven, and often absent in final reports. This haphazard manner of collecting and reporting data underscores the casual opinionative essence of Delphi. There are many levels of opinions ranging from snap judgments to carefully organized and well-defended documentation of positions systematically linked to interpretive concepts of construct validity. Although Delphi practitioners may point out occasional exceptions, snap judgments are apparently the rule for most Delphi questionnaire items, as shown below.

Bedford (1972) appears to be the only investigator who has solicited, classified, and analyzed all panelist comments, in his Delphi study on home communications services (for a sample of 1253 responses). His analysis of open-end verbal responses has led him to defect from "traditional Delphi with its heavy emphasis on statistical feedback" toward a structured adversary procedure "stressing the importance of assumptions, qualifications, interpretation of general trends, and criticism of co-panelist's remarks" (p. 43).

Questionnaire Reliability

The next section in the APA manual concerns test reliability. The first selected standard indicates minimal statistical requirements for reliability reporting.

D3. Reports of reliability studies should ordinarily be expressed in the test documentation in terms of variances for error components (or their square roots) or standard errors of measurement, or product-moment reliability coefficients. ESSENTIAL (p. 29).

Delphi studies invariably tend to ignore such "essential" considerations of test and item reliability. For example, Sahr (1970) presents some fifty pages filled with descriptive quantitative data comparing three Delphi studies conducted at the Institute for the Future. At no point does he report a single statistic indicating "variances, standard errors of measurement or product-moment reliability coefficients" required by this standard. Dalkey (1969) has made an initial attempt in this direction by indicating increasing reliability of medians with increasing sample size of panelists—a surprise-free result. (The standard error of measures of central tendency generally vary inversely with the square root of sample size.) He does not present standard errors of medians for individual item results as minimally required by this standard. Dalkey does present split-half (odd-even) reliabilities for some results, with coefficients usually varying between .4 and .6. This reported level of reliability is marginal for useful questionnaires. Furthermore, these are for end results with nonindependent or feedback-affected opinions, as discussed earlier. Reliability of first-round results would provide more meaningful coefficients for rigorous statistical testing. Dalkey's attempt to measure reliability is the exception rather than the rule for the descriptive statistics characteristic of the Delphi literature.

For example, Martino (1972) attempts to demonstrate the reliability of Delphi by listing several analogous items in presumably independent studies that resulted in "similar" predictions. No correlation coefficients or other statistical indices are reported; no account is presented of deleted items or discordant items; and no attempt is made to describe comparability of test conditions for final results. A study by McLoughlin (1969) is cited in which two groups of experts provided independent forecasts for fifty-five identical questionnaire items. The obtained standard deviation of the differences of the medians between the two groups was 3.54 years for events expected to occur before 1990. Martino concludes that this result shows a "high degree of consistency." On the contrary, assuming a 5 percent level of significance, this finding means that the "true" median forecast falls somewhere between ± 7 years of the obtained forecast (± two standard deviations), which is hardly the basis for a "high degree of consistency." A 95 percent confidence belt of 14 years is not very good for forecasts of events expected to occur within 20 years.

The next standard cited also applies to test reliability, in particular the stability of results.

D6. Test documentation should indicate to what extent test scores are stable, that is, how nearly constant the scores are likely to be if a test is repeated after time has lapsed. Test documentation should also describe the effect of any such variation on the usefulness of the test. The time interval to be considered depends on the nature of the test and on what interpretation of the test scores is recommended. ESSENTIAL (pp. 30-31).

This "essential" standard says, as applied to Delphi, that the questionnaire should be replicated at a later time on an independent sample of panelists, following original procedures, so that earlier results can be compared with later results to determine test reliability over time. No such replications are reported in the Delphi literature. This type of reliability is especially important for Delphi, because the method presumably measures attitudes toward the future, which change to a greater or lesser extent with changing conditions and independent panels. The absence of such studies, and the lack of interpretations of the underlying dynamics of attitude changes toward the future, is a major methodological and theoretical shortcoming in Delphi.

Some Delphi proponents object to a study of the underlying dynamics of attitudes toward the future, as distinct from and peripheral to the domain of Delphi opinion technology. The argument is that opinions are quite different from attitudes, particularly if they are concerned with technical subjects. Such a position reflects the isolation of Delphi from the mainstream of social science. The author concurs with Anastasi (1968), who says: "Opinion is sometimes differentiated from attitude, but the proposed distinctions are neither consistent nor logically defensible. More often the two terms are used interchangeably." (p. 480). In this report the two terms are used more or less synonymously.

The validity of any testing instrument can not be greater than its reliability; that is, a test can not correlate more highly with any external validation criterion than its correlation with itself (reliability). If Delphi results prove unstable in a given area over the short run, as with attitude fluctuations over time, its value as a prognostic instrument is likely to be worthless over the long run. Longitudinal

reliability studies of this type are essential for any defensible use of Delphi or its derivatives.

Experimental Sampling Standards

The final section of the APA manual covers sampling scales and norms. The next standard overlaps to some extent with prior discussion, but is worth emphasizing.

F6.11. Norms reported in test documentation should be based on a well-planned sample rather than on data collected primarily on the basis of availability. ESSENTIAL (p. 35).

Selection of panelists for Delphi studies tends to reflect expediency rather than a "well-planned sample," particularly when investigators are not accountable for sample specification under the anonymity clause. Heavy Delphi dropout rates can only compound and aggravate this shortcoming.

The next listed standard specifically warns against a standard Delphi practice of developing norms (generalizations) from small samples of panelists.

F6.31. If the sample on which norms are based is small or otherwise undependable, the user should be cautioned explicitly in the test documentation regarding the possible magnitude of errors arising in interpretation of scores. ESSENTIAL (p. 36).

If Delphi investigators made it common practice to report standard errors of estimates for small samples, it would be apparent to all that higher levels of precision, larger samples, and well-defined samples would be required. This is particularly true where medians are reported rather than means, since the standard error of medians is usually larger than mean errors. It is also the case for forecasts far into the future, where observed dispersions are typically very large, precision poor, and more extensive sampling necessary. Martino (1972), for example, has demonstrated an increasing dispersion of forecasts in many Delphi studies the farther away the expected year of occurrence.

The next standard describes a practice consistently neglected in the Delphi literature.

F6.4. Test documentation should report whether scores vary for groups differing on age, sex, amount of training, and other equally important variables. ESSENTIAL (p. 36).

The tacit Delphi assumption is that the pooled opinion of experts is better than that of any subgroup of experts. This may or may not be the case for any given area of Delphi inquiry. The fact remains, however, that there may be systematic effects related to the kinds of sampling characteristics mentioned in this standard. It behooves the Delphi investigator to test for such effects and to report them rather than to assume uncritically that the whole is axiomatically better than any of its parts. Dalkey (1969) has demonstrated sex differences for almanac items; Borko (1970) lists substantial sex and professional differences for library and information science research items; and Derian and Morize (1973) show systematic differences between types of medical specialists (researchers versus clinicians) in medical forecasting.

Conclusion

This concludes the tour through portions of the APA manual of standards relevant to Delphi. It should be abundantly clear that conventional Delphi neglects virtually every major area of professional standards for questionnaire design, administration, application, and validation. In no sense is Delphi found to be a serious contender in scientific questionnaire development and in the experimentally controlled and replicable application of questionnaires.

But this is not the whole story by any means. Many key areas remain to round out the picture. Only the methodology common to any questionnaire instrument has been covered. The special characteristics of Delphi remain to be reviewed and evaluated.

4 Delphi Evaluation: Backdrop

In this chapter, the historical precursors to Delphi are cited from the social psychological literature, and then the critical Delphi literature is reviewed. The chapter concludes with the Delphi evaluation scheme presented in the form of ten questions. Subsequent chapters analyze responses to these ten questions for a final assessment of Delphi.

Evaluative Delphi Literature

The pre-Delphi literature, mentioned earlier, anticipated many of the evaluative problems encountered in the use of opinion to forecast social and technological events. McGregor (1938) and Cantril (1938), from social psychological approaches, found the forecasting process using questionnaires provided a medium for projecting personal values and attitudes of the respondents. They made no claims for the validity of the technique in forecasting social events, nor for the ability of experts to predict complex social events any better than nonexperts. McGregor's conclusion summarizes his findings.

> The amount of information possessed by the predictor, and his sophistication or expertness are shown to have little significance in the determination of predictions concerning complex social phenomena. The quality of information as determined by ambiguity and importance is much more decisive (p. 203).

Cantril obtained similar results and concluded: "Whenever the prediction of a social event is based wholly or in part upon an internal frame of reference, objectivity is rare, if not impossible, because of ego-involvement" (p. 388). Both studies illustrate further the difficulties encountered in the use of opinion, expert or otherwise in predicting events.

Kaplan, Skogstad, and Girshick (1950) summarized the difficulties they encountered in trying to generalize from their results in

social and technological forecasting by questionnaire as fundamentally a problem of sampling. They concluded:

> The most serious question raised by a study of prediction is whether the analysis is made on a statistically stable population. The difficulties are threefold: those concerning the group of predictors, those concerning the questions asked, and those concerning procedure (p. 108).

These authors were skeptical of their findings because of uncontrolled and unknown individual differences between subjects, obvious differences between questionnaire items precluding extrapolations to related areas, and the limitations of the procedure, such as subjective factors in experimenters' judgment, time constraints in selecting items, multiple choice and probabilistic format of items, and discrepancies between use of judgment of subjects under experimental conditions as compared to use of experts under more realistic conditions. This was a pivotal study, one that provided key leads for initial Delphi developments. Unfortunately, scientific admonitions concerning statistical representativeness and experimental rigor, as we have seen in the previous section, were disregarded by Delphi originators.

The critical literature on Delphi is uneven and sparse. Quinn (1971) has described limitations of forecasting in general that apply to Delphi, including such factors as surprise events, inadequate or biased data, and unpredictable interactions. Pill (1971) explores various limitations of Delphi and, in connection with its reliance on human intuition, suggests that "perhaps the Delphi technique should be less allied with science than with metaphysics" (p. 61). Milkovich, Annoni, and Mahoney (1972) emphasize the loss of valuable data because Delphi participants are not allowed to interact directly. Weaver (1969, 1970) suggests that Delphi pays inadequate attention to psychological values and attitudes toward the future. (See Fishbein 1967, for a comprehensive introduction to the methodological literature on attitude testing.) Morris (1971) has criticized Delphi for not capitalizing on the extensive mathematical literature on the theory of subjective probabilities (e.g., Bayesian analysis); in the previous section, we have seen that this criticism applies not only to advanced probabilistic analyses, but also to elementary statistical treatment of raw Delphi data.

Derian and Morize (1973) criticize conventional Delphi for taking

the central tendency of pooled opinion at face value as a best estimate of expert opinion. Through the use of factor analysis of Delphi participants in their study, they found subgroups of experts clustering together with consistent opinions. They recommend analyses of subgroups such as research specialists, clinicians, and surgeons, rather than composite consensus.

As mentioned earlier, Bedford (1972) found so many shortcomings in conventional Delphi that he developed an independent technique called SPRITE: Sequential Polling and Review of Interacting Teams of Experts. In a comparative Delphi study on future home communication services, he found no consistent statistical differences in forecasting results between housewives and experts, and he found the qualitative responses more useful than the quantitative results. This led Bedford to drop the traditional Delphi emphasis on consensus, to move toward "controlled conflict" between contrasting groups, and to drop statistical feedback in favor of qualitative arguments. SPRITE is an example of nonconventional Delphi.

Weaver (1972) has probably contributed the most extensive critical review of Delphi uncovered in our survey of the literature. He asserts that the "vast majority" of Delphi studies "tend to be uncritical" and "promotional." He believes, "Delphi panels cater to the power structure" (p. 21). Delphi studies reviewed "suffer from technical limitations" subject to experimenter bias in collating and summarizing responses, subjectivity, lack of alternatives, and no checks on wording or order of items. Weaver asserts, "There is serious sterility in the process of summarizing mass information into narrowly terse statements. There is a serious absence of any effort to probe beneath the surface for explanations" (p. 21).

In discussing needed changes in Delphi, Weaver makes several recommendations. He suggests a shift away from mere description of events toward explaining events. He would drop anonymity, statistical feedback of dates and probabilities, and "consensus forcing procedures." He questions the notion "that convergence improves the accuracy of a forecast." Weaver would add face-to-face interaction and direct confrontation to ensure exchange of assumptions, arguments, and conclusions, and cites an example of such an exercise conducted at the International Adult Education Seminar, at Syracuse University. Weaver believes that the elimination of anonymity and statistical feedback and the introduction of face-to-

face confrontation still represents a recognizable variant of Delphi. It seems to this reviewer that with the rejection of the three "quintessential" elements of conventional Delphi (anonymity, iteration, and statistical feedback), any resemblance between Weaver's recommended interactive group process and conventional Delphi is strictly coincidental. Weaver's recommended approach closely resembles Heller's method of "group feedback analysis" (1969), which was developed independently of Delphi.

In his summary, Weaver asserts:

> At present Delphi forecasts come up short because there is little emphasis on the grounds or arguments which might convince policy-makers of the forecasts' reasonableness. There are insufficient procedures to distinguish hope from likelihood. Delphi at present can render no rigorous distinction between reasonable judgment and mere guessing; nor does it clearly distinguish priority and value statements from rational arguments, nor feelings of confidence and desirability from statements of probability.

Weaver concludes by urging his recommended changes in conventional Delphi and by stressing its value as an educational and heuristic tool as distinguished from a forecasting instrument.

Delphi Evaluation Scheme

The author generated a list of advantages and disadvantages of Delphi in his review of the literature, as a preparatory exercise to develop a data base for this critique. The disadvantages soon vastly outstripped the advantages. Approximately two hundred negative criticisms were compiled. These were arrayed as ten key questions, which are presented below.

The advantages of conventional Delphi, at least in this reviewer's estimation, are primarily low cost, versatile application to virtually any area where "experts" can be found, ease of administration, minimal time and effort on the part of the director and panelists, and the simplicity, popularity, and directness of the method. However, these and related advantages are characteristically obtained by unwarranted assumptions in method and approach and by seriously compromising the reliability, validity, and integrity of final results. Thus, such advantages are inconsequential if the

conventional Delphi concept, method, and results are inherently untrustworthy.

The ten key questions for conventional Delphi are:

1. Is the Delphi concept of the expert and its claim to represent valid expert opinion scientifically tenable, or is it overstated?
2. Are Delphi claims of the superiority of group over individual opinion, and of the superiority of remote and private opinion over face-to-face encounter, meaningful and valid generalizations?
3. Is Delphi consensus authentic or specious consensus?
4. Does Delphi anonymity reinforce scientific accountability or unaccountability in method and findings?
5. Does Delphi systematically encourage or discourage the adversary process and exploratory thinking?
6. Are Delphi questions, particularly forecasting questions, precise and meaningful?
7. Are Delphi responses precise and unambiguous?
8. Are Delphi results meaningful and unambiguous?
9. Is Delphi primarily concerned with collections of snap-judgment opinions of polled individuals from unknown samples, or is it concerned with coherent predictions, analyses, or forecasts of operationally defined and systematically studied behaviors or events?
10. Does Delphi represent a critical tradition, or is it uncritically isolated from the mainstream of scientific questionnaire development and behavioral experimentation? And does Delphi set a desirable or an undesirable precedent for interdisciplinary science in the professional planning and policy studies community?

Each of these questions is discussed in subsequent chapters.

5 Delphi Evaluation: Expert Opinion

As emphasized earlier, it is almost impossible to find current psychometric or social science literature on "experts." For example, the author was not able to find any continuing, systematic studies on experts in the recent *Psychological Abstracts* (except for highly specialized applications in legal testimony and clinical diagnosis), nor in Berelson and Steiner's (1964) inventory of findings on human behavior, in the *United Nation's Encyclopedia of Social Science* (Gould and Kolb 1964), in the *International Encyclopedia of Social Science* (Sills 1968), nor in many other social science texts that he has examined. Sole reliance on the use of expert opinion for scientific validation has long been discredited. There is a very extensive literature on psychometric scales for judgments, attitudes and opinions for a variety of tests (Anastasi 1968), for specified subject populations, but not for "experts."

Problems and Pitfalls With Experts

In assembling a relatively small group of experts, for the typical Delphi procedure, the director is tempted to select panelists he knows or colleagues recommended by his acquaintances, because it is easier and faster, with fewer rejections. Perhaps the fastest way to discourage a Delphi study is for the director to fight uphill against a high dropout rate from panelists. The resulting sample of "experts" is likely to include people with similar backgrounds and interests, who think along similar lines. Such groups may also tend to comprise an elite with a vested interest in promoting the area under Delphi investigation. Expert panels are often selected from accessible experts, and this accessibility is largely covert. Delphi reports characteristically offer little or no information about panelist selection and provide no safeguards against such abuses.

Top names in the field under investigation lend prestige to the Delphi study. The inclusion of prestigious individuals acts as a

magnet to attract others less prestigious. However, the prestige personalities may be counterproductive: the younger and more obscure panelists may be more highly motivated to work harder at the questionnaire and provide more carefully considered responses. There is always the choice between the older, established professional versus the young Turk. Representation of the entire spectrum is probably better than taking sides, at least to help assure more diversified opinion. Turoff (1971) and Martino (1972), alarmed by uncontrolled panelist dropout rates, and concerned with the need for higher levels of panelist motivation and more carefully reasoned responses, recommend budgetary provision for honoraria for panelist time and effort.

The use of experts leads to a serious technical limitation of the Delphi questionnaire: the fallacy of the halo effect, in this case the expert halo effect. This is the tendency of respondents to be unduly influenced by any favorable or unfavorable characteristic of the questionnaire which colors and contaminates their judgment. For example, a highly desired technological event may systematically receive more optimistic forecasts than a neutral event.

Delphi is enmeshed in a pervasive expert halo effect. The director, the panelists, and the users of Delphi results tend to place excessive credence on the output of "experts." Panelists bask under the warm glow of a kind of mutual admiration society. The director has the prestige of pooled authority behind his study, and the uncritical user is more likely to feel snug and secure under the protective wing of an impressive phalanx of experts.

The result of the expert halo effect for Delphi is to make no one accountable. The director merely reports expert opinion objectively according to prescribed procedure; he is not responsible or liable for outcomes. The panelist obligingly follows the ritual, protected at all points by faceless anonymity. The user can always claim that he was simply following the best advice available, and that he is not responsible for what the experts say. Everyone has an out, no one needs to take any serious risks, and no one is ultimately accountable. With so much to gain, so little to invest at such low risk, no wonder the method is so popular. The Delphi belief structure is psychologically held together by the cementing influence of the expert halo effect.

A tacit, largely unchallenged assumption of Delphi is that authentic experts do in fact exist for predicting the extermely complex socioeconomic-technological events so common in Delphi questionnaires. Closer scrutiny reveals this to be wishful thinking. Many of these events are initial forays into unknown areas requiring unknown skills, hence, unknown "experts." Even if such events are understood to some extent, they typically presuppose a fantastic array of real, not shallow skills in diverse and far-ranging fields, such as economics, public policy, esoteric technologies, individual and group psychology, law, medicine, etc., which is simply beyond the ken of any living mortal. When we match predictions of complex sets of social events against "experts," we get something like the fabled blind men examining the Indian elephant. If we think of experts as *idiots savants*, we suddenly avoid the trap of the expert halo effect.

Experts Versus Nonexperts

Another central postulate in the Delphi epistemology of experts is that they will in fact provide significantly better and substantially different responses than nonexperts. Practically every Delphi practitioner asserts that Delphi outputs are only as good as the expert inputs, admonishing us with the GIGO principle (garbage in/garbage out).

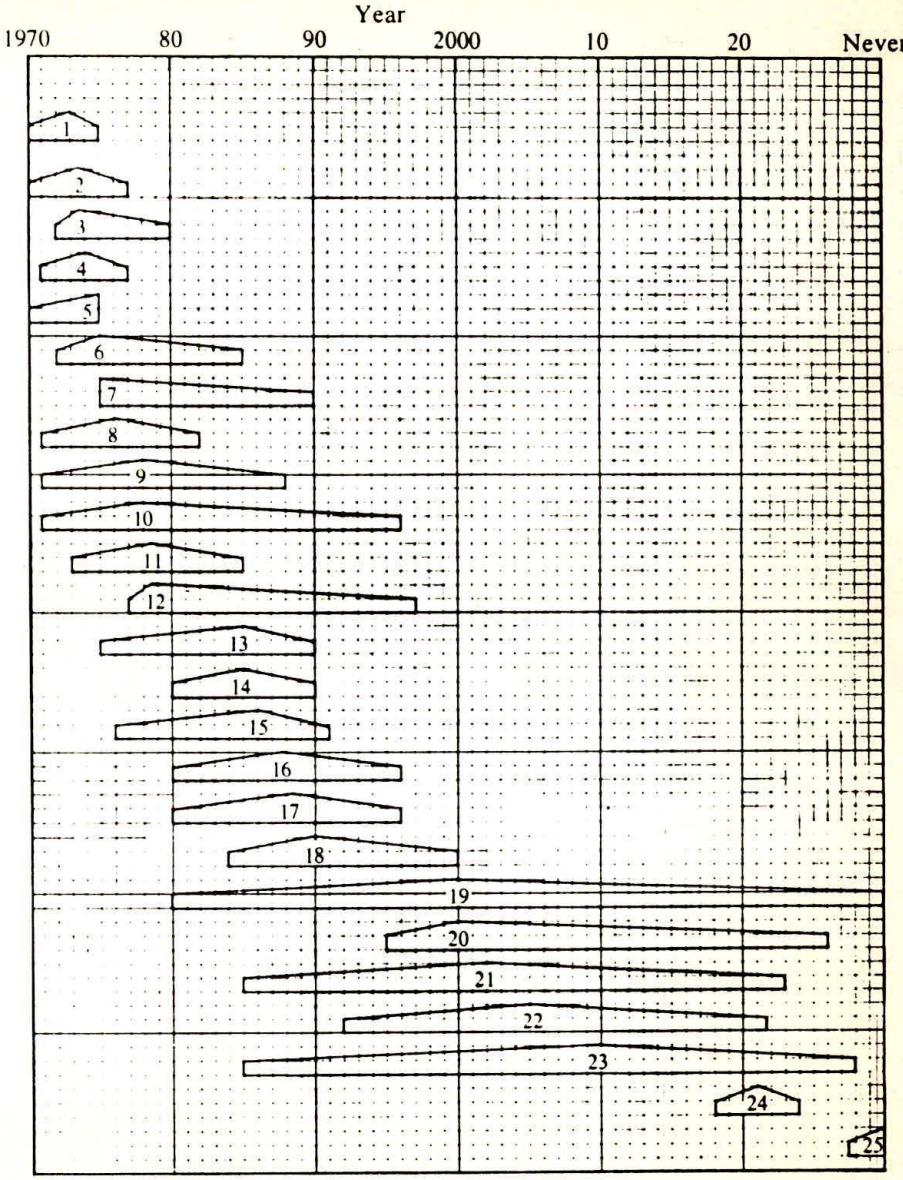
Suppose, however, that it can be proven that any informed group of individuals in the object area of inquiry can provide individual and group Delphi opinions essentially indistinguishable from those of the experts. It would follow, then, that Delphi results merely represent informed opinion rather than expert opinion.

Personal experience with graduate student predictors brought this potential expert fallacy to the author's attention. In connection with a graduate-level course on computers and society, the author asked his students to give their independent estimates of expected order of occurrence of each of the events in automation (computer technology) and general scientific advances originally investigated by Gordon and Helmer in their 1964 Delphi study. (See figure 5-1 for results with automation items.) After the students ranked the listed events, they were told the "true" ranks listed by the experts in the original study, and calculated a Spearman rank coefficient (product-moment correlation of ranks). This provided each student with a correlation coefficient comparing his first-round estimates with the medians of the "experts." Over the years, we have consis-

1. Increase by a factor of 10 in capital investment in computers for automated process control
2. Air traffic control—positive and predictive track on all aircraft
3. Direct link from stores to banks to check credit and to record transactions
4. Widespread use of simple teaching machines
5. Automation of office work and services, leading to displacement of 25 percent of current work force
6. Education becoming a respectable leisure pastime
7. Widespread use of sophisticated teaching machines
8. Automatic libraries looking up and reproducing copy
9. Automated looking up of legal information
10. Automatic language translater—correct grammar
11. Automated rapid transit
12. Widespread use of automatic decision making at management level for planning
13. Electronic prosthesis (radar for the blind, servomechanical limbs)
14. Automated interpretation of medical symptoms
15. Construction on a production line of computers with motivation by 'education'
16. Widespread use of robot services
17. Widespread use of computers in tax collection
18. Availability of a machine which 'comprehends' standard IQ tests and scores above 150
19. Evolution of a universal language from automated communication
20. Automated voting, in the sense of legislating through automated plebiscite
21. Automated highways adaptive automobile autopilots
22. Remote facsimile newspapers and magazines printed at home
23. Direct electromechanical interaction between man and computer
24. International agreements which guarantee certain economic minima to the world's population as a result of high production from automation
25. Centralized (possibly random) wire tapping

Source: Adapted from Gordon and Helmer 1964.

Figure 5-1. Delphi Results for Progress in Automation



TECHNOLOGICAL PROGRESS in automation as predicted by a panel of experts has been obtained by investigators at the RAND Corporation using the Delphi technique. The length of each bar represents various estimates put forward by the 'middle half' of the panel. In each case one quarter—the 'lower quartile'—proposed dates earlier than that at which the bar begins and another quarter—the 'upper quartile'—give dates beyond that marking the end of the bar. Each bar has a peak value which represents the median date estimated.

tently found median rank-correlations for classes of about a dozen students at about .70 for both areas for first-round estimates. These results are roughly equivalent to the upper levels of reliability for Delphi judgments described earlier from Dalkey (1969).

In a nutshell, "informed" graduate students provided essentially the same forecasts as "experts." The students did have the advantage of making their predictions some six years later than those in the original study, items were not presented randomly, there was no iteration with feedback, and standardized instructions were not rigorously observed in these informal classroom exercises. If this equivalence holds under controlled experimental conditions, anyone with some professional training in broad target fields could play the Delphi game, and it wouldn't make any difference in the results.

The tests using graduate students were not conducted as rigorous experiments, and the results have not been written up or reported in the literature. No claims are made for the validity of the findings; they are presented here to point up a central hypothesis. Some critical experimental studies comparing experts with less informed individuals and with nonexperts have been performed in the Delphi literature and in precursor studies. This is a central empirical question that can be very easily tested.

At the beginning of this section, the studies of Cantril (1938) and McGregor (1938) were cited. In these studies, the expertness of the forecaster was shown to have little or no significance in the determination of predictions of complex social events. More precisely, no statistically significant differences in such predictions were found between students and teachers, laymen and professionals, in tests which involved a combined respondent sample of over six hundred subjects. Predictions were demonstrably linked to values and attitudes toward the subject matter.

Kaplan, Skogstad, and Girshick (1950) applied a forecasting questionnaire on 152 social and technological events to twenty-six subjects representing the entire spectrum from senior professional to layman. Part of the study involved administration of a general knowledge paper-and-pencil test on "Current Social Problems" and "Science." The better-informed subjects (upper half) performed only slightly better than the less-informed subjects (lower half); average accuracy scores for short-term predictions were 56 percent and 50 percent respectively. This result is in the expected direction, but is not statistically significant with respect to a test for the mean

difference between proportions for this sample. Further, the amount of the difference, as indicated by the authors, is not substantial. Thus, these pre-Delphi studies indicate that expertise either makes no difference, or only a trivial difference, in forecasting a variety of social and technological events.

Much the same results occur with Delphi studies. In Campbell's doctoral dissertation on forecasting short-term economic indicators (1966), level of expertise was tested in terms of self-confidence ratings. He correlated these ratings for each item against forecasting accuracy and found the results did not differ significantly from a median correlation of zero. Campbell concluded that selecting the most self-confident members of a group was not an effective means of identifying the most accurate forecasters.

Campbell had additional information for a further test of the relation of expertise to accuracy in forecasting. Of the two seminar groups tested, one group was older and more experienced in professional economic forecasting than the other group. The more experienced group did obtain more accurate median forecasts more often than the less experienced group in a paired-comparison test, but the results were not statistically significant for Delphi and non-Delphi groups matched against each other for sixteen economic indicators. (Since Campbell did not report statistical comparisons, the author applied the nonparametric sign test used by Campbell in similar comparisons and obtained confirmation of the null hypothesis for Delphi and non-Delphi groups). The pooled results showed twenty more accurate forecasts for the more expert group, ten for the less expert group, and two ties, which meets a 10 percent level of significance.

Dalkey (1969), also using self-confidence ratings of expertness for each item, was able to compare those "more expert" against those "less expert" for almanac-type questions. "The basic hypothesis being tested was that a subgroup of more knowledgeable individuals could be selected in terms of their self-rating, and that this group in general would be more accurate than the total group. In every case this hypothesis was not confirmed" (p. 68).

In a subsequent almanac item study, Dalkey, Brown and Cohran (1969) did find that "significant improvements in accuracy of group estimates can be obtained with proper use of self-ratings" (p. v). Close examination of "proper use" reveals rather arbitrary ex post facto statistical verifications that have dubious generality for

other studies (e.g., at least seven subjects in high and low subgroups, with no overlap in self-ratings between subgroups, which eliminated many of the subgroups). Such arbitrary ad hoc statistical procedures capitalize on chance fluctuations in the experimental sample. A more appropriate statistic would include all data, such as a correlation coefficient showing both the statistical significance and strength of the association between self-ratings and accuracy.

Bedford's study (1972) is probably the most relevant to the issue at point: are there demonstrable forecasting differences between experts and nonexperts? Bedford matched a group of twenty-five housewives against a group of twenty-six experts in "communications, consumer behavior, sociology, and futurism generally" in a two-round Delphi study on "The Future of Communications Services in the Home." Bedford found, using a long and extensive questionnaire, "remarkably few differences between the experts and the housewives on the panel" (p. 1). His results support the contention that level of expertise makes little difference in exploratory socioeconomic forecasts.

Similar results were obtained by Reisman, Mantel, Dean, and Eisenberg (1969) in a comparative Delphi study. Evaluative ratings of laymen correlated highly with ratings of experts for 250 social service packages handled by the agencies of the Jewish Community Federation of Cleveland. These results also tend to support the hypothesis that opinions for evaluative social areas of inquiry tend to be independent of level of expertise.

What is the box score for the null hypothesis that there are no demonstrable differences between predictions of experts and nonexperts for socioeconomic-technological events? The McGregor (1938) and Cantril (1938) studies unequivocally indicate that such differences do not exist for complex social events impacting on personal values. The Bell Canada study by Bedford (1972) indicates that no demonstrable differences were shown between housewives and experts for sociotechnological developments. Campbell's analysis of self-confidence ratings also supports the null hypothesis, in that no correlation was obtained with accuracy of short-term economic forecasts. Dalkey's 1969 study showed no differences in almanac-item estimates with respect to ratings of self-confidence. Reisman et al. (1969) showed similar responses from laymen and experts in evaluations of social services. These studies collectively indicate that it doesn't make any difference how

expert the respondent is, or how confident he feels about his opinion, when forecasting or estimating a wide variety of social, economic, and technological phenomena.

Studies that show some differences in responses between different levels of expertise are marginal at best. The Kaplan, Skogstad and Girshick study showed a statistically nonsignificant trend in the correct direction, with more "knowledgeable" subjects contributing more accurate short-term forecasts. Campbell's data (1966) also showed a statistically nonsignificant trend in the expected direction, with his more experienced group tending to give more accurate forecasts than the less experienced group. The Dalkey, Brown, and Cochran study (1969) showed statistically marginal results in the expected direction for self-confidence ratings.

If Delphi investigators can not demonstrate statistically significant and substantial differences between experts and nonexperts, then it must be concluded that the Delphi emphasis on the use of experts is misplaced. Available experimental data indicate that this conclusion is probably the most accurate generalization for most Delphi applications. If statistically significant, but low-order correlations are found, the expert concept is only marginal and virtually worthless from a practical point of view. The above experimental data indicate that this might be the case in a small proportion of well-defined and highly specialized applications. If significant and substantial differences are found, a stronger case may be made for Delphi expert opinion for the target area of inquiry. The above experimental data offer no evidence of substantial difference between experts and nonexperts.

Conclusions

Looking back at the central issue of the Delphi concept and use of experts as discussed in this chapter, we find the following shortcomings:

1. The concept of expert is virtually meaningless in experiments dealing with complex social phenomena.
2. Sole or primary reliance on expert opinion in the social sciences has long been discredited and now has no serious advocates.
3. Anonymous panels chosen in unspecified ways increase the likelihood of contaminated, elitist "expert" samples.

4. There exists an uncontrolled and unknown expert halo effect in Delphi contributing to expert oversell.

5. Collective expert opinion directly reinforces unaccountability for Delphi results for all concerned: the director, panelists, and users.

6. Experts and nonexperts consistently give indistinguishable responses in forecasting or evaluating social phenomena impacting on common values and attitudes.

7. There is no explicit matching of skills required by Delphi questions against objectively measureable skills of the panelists.

The difficulties associated with the Delphi concept of "expert" does not and should not imply that all and any use of experts is necessarily bankrupt. The originators of Delphi should be credited with clearly sensing and trying to respond to strong social demand for exploiting expert opinion more effectively. For example, in a survey of sixty-five corporations, Hayden (1970) found that 69 percent used diverse expert panel consensus techniques, and of these, 26 percent used Delphi. This example is probably indicative of the widespread informal and formal use of "experts" throughout society. The proper use of expert talent remains a major problem of our time. We know precious little about the dynamics, the use, and the abuse of experts in our society. Substantive treatment of this problem, however, is beyond the scope of this analysis.

6 Delphi Evaluation: Group Process

This chapter addresses the next four Delphi evaluation questions, which cover various facets of group process with conventional Delphi. These questions cover group opinion, group consensus, anonymity, and adversary process.

Direct Confrontation Versus Private Opinion

Much of the popularity and acceptance of Delphi rests on the claim of the superiority of group over individual opinions, and the preferability of private opinion over face-to-face confrontation. Martino (1972), for example, flatly asserts, "It should be remembered that Delphi represents a distinct improvement over either individual experts or face-to-face panels" (p. 27).

Democratic process rests on the secret ballot, where voting is performed in private. Group opinion is a time-honored corrective against individual excesses. And how many of us have either been bullied in heated group exchanges or have bullied others when we had the opportunity? Besides, who wants to take the time and effort to travel to a meeting and listen to every panelist defend his expertise to the rest of the group? A quick and incisive statement of the issues on paper and an equally quick indication of individual opinion, also on paper and in the familiar privacy of your own office, as advocated by Delphi, has almost irresistible practical appeal as a sensible and cost-effective solution to the problem of sampling expert opinion.

On the other hand, each of us can probably recount numerous examples where individuals were more effective than groups in arriving at informed opinion, where confrontation clarified the issues and made honest communication possible, where introversion and isolation led to unfortunate aberrations of opinion and outlook.

The experimental data comparing individual and group performance offer no convincing conclusions on either side of these broad issues, although the literature extends over many decades. After

reviewing the early literature in this area (1920-57) Lorge, Fox, Davitz, and Brenner (1958) indicate that superiority of the group or the individual is relative to stipulated experimental tasks and conditions, varies enormously with individual differences, and is shot through with methodological difficulties in generalizing from experimental to real-world situations. In a more recent review of the experimental literature, Maier (1967) concludes that the comparative effectiveness of individuals versus groups varies widely and depends upon the tradeoff of the assets and liabilities of both approaches in the unique applied setting. He emphasizes the crucial role played by experienced group leaders acting as neutral facilitators in achieving successful group outcomes.

If we look for Delphi studies comparing groups and individuals, we find a near vacuum. Dalkey (1969) compared face-to-face with anonymous Delphi interaction for the almanac-type items mentioned earlier. He found a statistically nonsignificant tendency toward more accurate opinion in the anonymous setting. Dalkey's procedure involved picking group "leaders" randomly, which flies in the face of effective group procedure and effectively stacks the odds against successful group interaction. Farquhar (1970) compared group-versus-anonymous Delphi interaction for a complex software estimation task and consistently obtained substantially better results in the face-to-face group.

Campbell's dissertation (1966) is frequently cited by Delphi proponents as definitive evidence of the superiority of Delphi group opinion compared with face-to-face confrontation in traditional expert panels. Campbell worked out a careful experimental design as far as subject sampling is concerned, randomly assigning graduate student participants to experimental Delphi panels and control confrontation groups (which he calls uncontrolled-interaction groups). His criterion measure consisted of accuracy in forecasting sixteen short-term statistical economic indicators; a flaw in this part of his study is that these sixteen measures are only partially independent, which vitiates the integrity of statistical tests based on assumptioms of independence. Campbell used nonparametric statistics in comparing median forecasting performance of his experimental (Delphi) and control groups (confrontation) and apparently demonstrated statistically significantly better forecasting in his two matched Delphi groups.

His conclusion, however, is based on a straw man type of comparison, similar in certain respects to the token conventional group structure used by Dalkey (1969). Campbell's control groups were leaderless and remained leaderless, which undoubtedly led to considerable floundering and nonmission posturing and competition. The simple institution of an elected chairman to organize each group and identify with the problem, as occurs in conventional committees, might have altered results substantially. The confrontation groups were force-fitted into a Delphi-type format to make quantitative forecasts more directly comparable. For example, meetings were kept within fixed periods of time, whether or not the group wanted such a procedure, with one meeting corresponding to each round of the Delphi panels; discussion of each economic indicator was also pegged to a fixed period of time regardless of success or failure in achieving closure or consensus; and each meeting required open individual polling for statistical comparability of estimates between experimental and control groups, whether or not the group wanted to follow such a procedure. These Procrustean constraints break most of the rules for professional or enlightened group problem solving. The oppressiveness of these artificial confrontation groups may have undermined group motivation and morale to the point where the meetings became counterproductive, and the comparison spurious. Accordingly, Campbell's study can not be viewed as a serious comparison of the effectiveness of Delphi and conventional panels for his criterion measures.

The results raise additional methodological problems. Campbell did not compare the forecasting results of both types of groups against trend extrapolations of his selected economic time series, even though these series were available on a quarterly basis. It may be that simple arithmetic extrapolation (as mentioned earlier in connection with Zarnowitz's critical review of expert economic forecasting) or perhaps more sophisticated multiple regression analyses might provide results as good or better than those obtained with expert groups. Finally, quarterly forecasting is hardly a criterion vehicle for an expert panel when reliable and extensive baseline statistical data are available for fine-grain, short-term trend forecasting. Any generalization from such results would have to be limited to very short-range forecasting.

The alleged superiority of anonymous Delphi opinion over face-to-face opinion, and its converse, are unprovable general propositions. They can not be proved or disproved, in general, because the

propositions are amorphous stereotypes and are not amenable to scientific testing unless they are operationally defined. Once such definition is applied to limited concrete situations, one approach may prove more effective than another, both approaches may be more powerful than either alone, or the two approaches may be so close as to not make much practical difference. Investigators should be more interested in a flexible eclectic approach that freely capitalizes on the best of both worlds than in identifying with a ritualized approach on either side. In any case, the Delphi claim that pooled group "expert" opinion is more effective than individual opinion, and that anonymous interaction is more effective than direct confrontation, cannot be sustained.

Delphi Consensus

The goal of the Delphi procedure is to arrive at a meeting of the minds, consensus among the experts. The position taken here is that the Delphi procedure arrives at such consensus by feeding back the "correct" answer, by rewarding conformity and effectively penalizing individuality, and by proffering nonindependent iterative results as authentic expert consensus. Authentic consensus refers to group agreement reached as a result of mutual education through increased information and the adversary process, which leads to improved understanding and insight into the issues; it does not refer to changes of opinion associated primarily or exclusively with bandwagon statistical feedback.

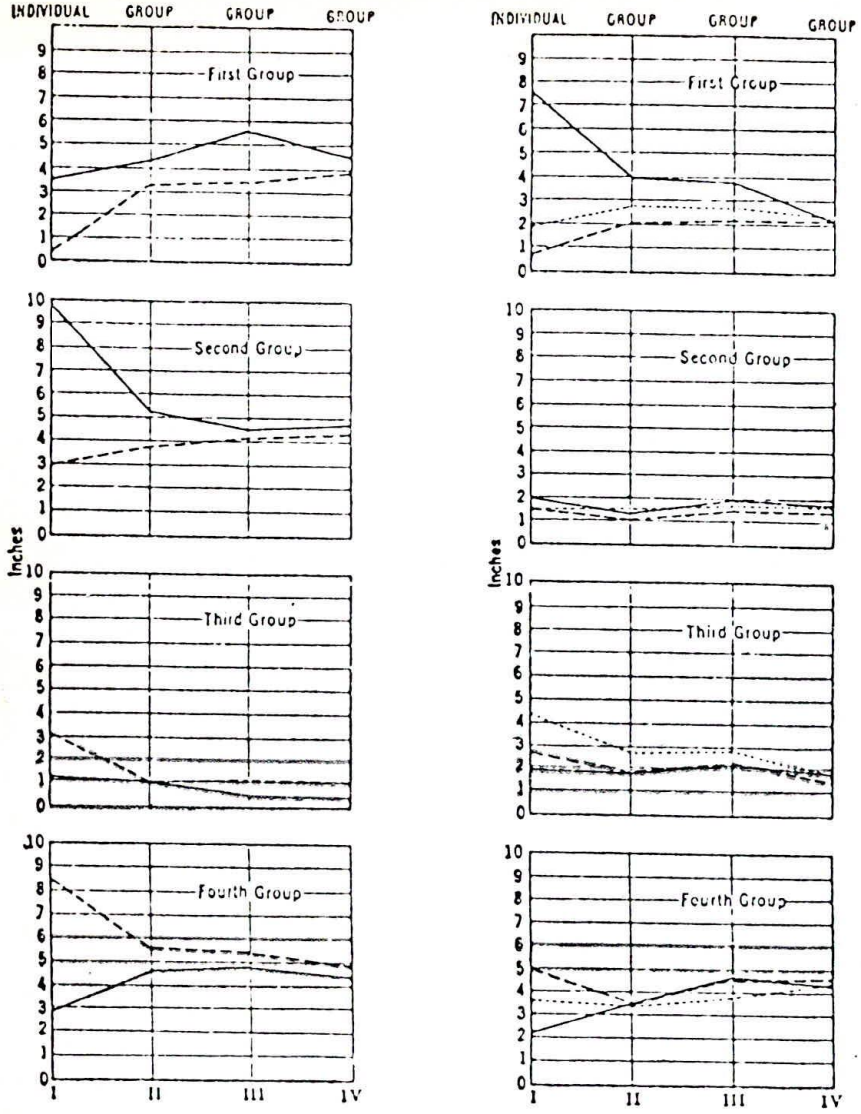
It was stated earlier, in connection with the APA professional standards for soliciting judgment, expert or otherwise, with a standardized instrument, that the judgment should be independent. The first Delphi round represents independent opinion, whereas succeeding rounds are strictly correlated. First-round results of "experts" may contain a range of responses up to four orders of magnitude for some types of quantitative estimates, such as Dalkey, (1969) and Baran, (1971), which are hardly publishable as "consensus." Raw-score frequency distributions are so highly skewed that logarithmic transformations are often required to approximate normal distributions. Perhaps this is why most Delphi investigators do not report first-round dispersions. Borko (1970) provides an exception to this rule.

An example of the logarithmic-range of first-round dispersions for some types of Delphi estimates is provided by Baran (1971) in an illustrative appendix of his report for "Cashless-Society Transactions." This item refers to cost and marketing estimates of hard copy recording of financial transactions with updated balance in computer memory. The first round showed a range of $.01 to $100 for average dollar value of a transaction (10,000:1), 5 percent to 90 percent market penetration five years after mass introduction of this service, and a range of 0 percent to 100 percent for percentage of this service that home subscribers would be expected to pay. The inkblot nature of such future projections speaks for itself.

Now, in succeeding rounds, do the panelists really think through their positions and work toward authentic consistency of opinion, or are they effectively pressured into conformity? Dalkey (1969) has indicated that statistical feedback alone (group medians for each item) is as effective in obtaining consensus as statistical feedback with adversary rationale for responses. Once the panelist knows the median for a problematic item, he has in a very real sense been given the "correct" answer to the item. Panelists are quite aware that median responses (or some other measure of central tendency) are offered as best estimates for questionnaire items in the final results.

Social psychologists have long been aware of powerful tendencies for individuals to conform to group opinion in relatively unstructured situations, particularly if the motivation level is not high (Stogdill 1959; and Berelson and Steiner 1964). The "autokinetic" effect is a striking example of this tendency (Sherif 1936). Place an individual or a group of people in a completely darkened room with a single, fixed point of light. The light will appear to drift randomly with a displacement as high as 20°, because of the absence of a visual frame of reference. (Astronomers were the first to notice and study the autokinetic effect.) Ask the subjects in such a room to estimate the direction and amount of perceived movement of the light. Initial random judgments soon converge closely around the group norm after a few rounds of group opinion. Group suggestion provides the "correct" answer to an inherently ambiguous situation. Consensus is specious.

Figure 6-1 shows some of Sherif's experimental results with the autokinetic effect. The first session involved individuals alone reporting observed deviations of the pinpoint of light in a completely darkened room. The ordinates in figure 6-1 represent median devia-



Source: Figure 10.3, p. 208 from *Social Psychology* by Muzafer Sherif and Carolyn W. Sherif (Harper & Row, 1969).

Figure 6-1. Specious Consensus: Autokinetic Convergence

tions in inches; the abscissas represent successive sessions (equivalent to Delphi rounds). The second, third, and fourth rounds were group sessions where each individual had an opportunity to hear the deviations reported by others. Note that the individual median deviations rapidly converge to a group norm by the fourth round for groups of two or three subjects.

The analogy with Delphi is startling. Convergence of medians is greatest with initial feedback of group opinion and is effectively achieved in three to four rounds. Delphi investigators typically reach the point of diminishing returns at about three or four rounds as far as measurable convergence of opinion is concerned. When we couple Sherif's results with Dalkey's assertion that statistical response alone is the most effective way to achieve consensus (without verbal feedback) we have the artifact of autokinetic consensus (group suggestion) explaining Delphi consensus. Sherif ran many variations of the autokinetic effect demonstrating easily *manipulated* shifts in subjects' opinions in *any* desired direction by suggestion from the experimenter (e.g., "you are underestimating light movement") or from other authority figures, such as group leaders. The uncontrolled, arbitrary introduction of selected verbal feedback by the Delphi director can with corresponding ease shift opinions in desired directions.

The Delphi technique thus deliberately manipulates responses toward minimum dispersion of opinion in the name of consensus. The presentation of median opinions (after the first round) and the coercion toward conformity are reassuringly represented to all as reasoned consensus. By the third or fourth round, the holdout individualist responses pose the threat of yet another tedious run through the same items, and even die-hards are inclined to yield to save everyone the dreary routine of another round. Martino (1972) states, "In many cases, there is no advantage in going beyond two rounds" (p. 27).

In passing, it should be noted that the term *panelist* is a misnomer in the Delphi context. Panelists usually communicate directly and exchange opinion with each other, primarily in a face-to-face setting. With Delphi, we have respondents, not panelists, because communication is strictly with the questionnaire, not with other people. Moreover, all responses are filtered through the intermediary of the Delphi director or his representative before reaching anyone else. There is no interactive discourse deserving of the name

panel in Delphi procedures. Respondents really represent a noncommunicating nongroup, linked primarily by remote statistical feedback.

Delphi consensus is suspect from yet another viewpoint. The first-round items are quite different when they are accompanied by statistical and verbal feedback provided by the director in succeeding rounds. Once the information accompanying an item is altered, it is literally a different item. Just as minor rewording can change a questionnaire item enormously, so does Delphi "feedback" change the item in uncontrolled and unknown ways. How can medians and dispersions be compared, and consensus claimed, if items are noncomparable from round to round?

The social implications of specious consensus are enormous. Variations of similar iterative query techniques, with conformist-reinforced feedback, provide almost unlimited possibilities for shaping and manipulating public opinion via the interactive communications media of the future.

Anonymity and Accountability

The anonymity of Delphi experts serves the dual purpose of attracting expert panelists by guaranteeing protection against individual accountability, and projecting an inviting image of a kind of permissive brainstorming where "anything goes" to help "cream off" the best the experts have to offer. The panelists are assured full protection against any invasion of privacy. When coupled with the blandishments of joining the inner circle of eminent experts, the combination is hard to resist. But few have realized that the price of such inducements is abandonment of accountability, and that it may promote elitist vested interests.

Under a "no disclosure of names" policy, no individual is accountable for either his own responses or for group Delphi results. As pointed out earlier, Delphi embodies circular buckpassing. The director reports group opinion following an objective ritual; the results are not his personal opinion. Each panelist is faceless in any of the results and can always blame nameless others for any findings he dislikes. The consumer of Delphi gets his low-cost preview of the future and can claim he had nothing to do with the final results.

Directors should be accountable for all flaws in the method, and for implicitly or explicitly overstating the value and significance of potentially misleading final results. Panelists should be accountable for unwittingly lending the authority of their reputations and their support to demonstrably unreliable and invalid short-cuts to the future. Individual and institutional users should be accountable for funding and popularizing such studies, and for accepting Delphi forecasts at face value.

Elitist tendencies are strongly reinforced not because of any diabolical plot on the part of Delphi investigators, but for the more mundane and more compelling reason that it is a lot easier and faster to assemble colleagues, acquaintances, or second-order recommended acquaintances for the expert panel.

A major attraction of Delphi for busy researchers of all callings is that it is cheap and easy, as well as a relatively painless and well-protected technique. A study can be conducted and a paper produced with relatively small effort. Martino (1972) claims, "A planning factor of two professional manhours per panelist per questionnaire is a fair approximation to the workload which will be required" (p. 60).

The questionnaire is quite likely to represent many aspects of the work done by these experts, almost by definition. Chances are that while such panelists will have much to disagree over, most will be interested in promoting the image, value, and particularly the future of their field. Big developments will then be perceived as occurring early and making large impacts on society. For example, Nanus, Wooten and Borko (1973), in their Delphi study on the social implications of multinational computer systems, admit that their sample of fifty-six "eminent" panelists, typically active in various aspects of this field, were probably biased to some extent toward promoting the importance and enhancing the image of multinational computer systems. There is no malevolent design or covert collusion in such opinions, merely self-aggrandizement and self-interest.

Controlled experiments soliciting opinions from contrasting or even antagonistic groups (expert or otherwise) are likely to produce quite different results. As mentioned earlier, Delphi opinion polls measure attitudes toward future events, not predictions of such events in their own right. As currently practiced, Delphi can easily slant results in the direction of aggrandizing vested interests. With

anonymous sampling of "experts", the burden of proof should be on the Delphi investigator to demonstrate that his panel does not represent a narrow elitist circle.

Kopkind (1967), in his widely cited article on "The Future Planners," expressed his concern over futurist elitism: "The danger is that Government and corporate elites will monopolize the business of question-asking, and so manipulate the attitudes of society they are pretending to serve as disinterested technicians" (p. 23).

Adversary Process

Most Delphi practitioners claim that Delphi is able to go where other investigators fear to tread. Opinion can peer into every nook and cranny, particularly those inaccessible to conventional techniques. Delphi thus has the advantage of being able to get "there" first, or among the first, and of making early pronouncements concerning new horizons far in the future. This capitalization on novelty is part of the dramatic appeal of Delphi. Plumbing the depths and climbing the heights of the future hold spills, thrills, chills, and some jolts of future shock for everyone.

It would seem plausible that at least until we learn a good deal about any new domain, it should be the object of free inquiry and of very active adversary proceedings. Delphi systematically inhibits the adversary process. This indictment is not in any sense original with the author; variations have been made by Bedford (1972), Milkovich, Annoni and Mahoney (1972), Turoff (1972), and Weaver (1970).

Delphi deliberately factors out face-to-face confrontation, and the adversary process associated with it, as one of its prime philosophical tenets justifying efficient consensus. Arguments are filtered, buffered, and effectively neutralized in Delphi. A panelist can participate without providing any justification for any of his opinions throughout the entire procedure. More conscientious panelists provide occasional brief commentaries.

The real payoff for the Delphi investigator is obtaining maximum consensus from the experts. Interquartile Delphi forecasting graphs, spreading from now to never, are the nemesis of Delphi practitioners. The smaller the spread the more powerful the impact. Real

adversary excitement over authentic controversial issues is plainly the enemy of consensus. Boredom and snap responses make for smaller differences and maximum consensus. In many cases, only the outliers have to justify their positions in Delphi iteration; directors make minimal demand on those occupying the middle ground.

By inhibiting the adversary process, Delphi also inhibits open exploration of new domains. Free exploration leads to adversary inquiry and generates new controversy. This can lead to polarization of opinion that undermines consensus in final Delphi results. But it is precisely the new domains that need free exploration and the adversary process the most. Delphi should be prodding conformers and rewarding outliers to maximize exploration, highlight controversy, and map out the unknown. When we are really ignorant, we need all the contrasting viewpoints we can get to encourage free and informed choice.

7 Delphi Evaluation: Validity of Results

This chapter treats the five remaining questions in the Delphi evaluation. These cover questionnaire items, questionnaire responses, results, Delphi epistemology, and the isolation of Delphi from the mainstream of social science. The chapter concludes with the results of the evaluative analysis in the form of answers to the ten questions posed earlier.

Delphi Questionnaire Items

The basic criticism leveled against Delphi questionnaire items is that they are, by and large, unavoidably amorphous. More specifically, complex future events (and value judgments) do not lend themselves to clear and unambiguous description in typical one-sentence Delphi questionnaire format (see the automation items from the Gordon-Helmer study in figure 5-1). What we get instead are vague, generalized descriptions of future events, permitting the respondent to project any one of a large number of possible scenarios as his particular interpretation of that event. Delphi asks panelists about event-stereotypes, and panelists respond with stereotype estimates. Delphi verbal responses, when they occur, are typically vague and sweeping descriptions, slogans, or simplistic statements.

The more thoughtful and careful Delphi investigators attempt to qualify forecasts by identifying percentages of specific respondent populations and by associating probability estimates with predictions. Such attempts, although in the right direction, are no substitute for precisely defined, detailed scenarios for each item where a host of assumptions specifying the "event" are made explicit. The questionnaire format does not lend itself to such presentation.

For example, the Delphi inquiry might be concerned, as in Baran's study (1971), with the "Potential Market Demand for Two-Way Information Service to the Home." Baran had to leave vast areas unspecified in asking panelists when such services were

likely to be available and how much they would cost the consumer. These unspecified areas included the configuration of hardware, software, and communications; the nature of federal, state, and local regulation of such mass computer services; the mix of public and private support of the information services considered; very brief general descriptions of the thirty information services (typically one paragraph); no indication of how the public will be taught to use such services; and many other socioeconomic-technological areas impacting directly on these services. Baran's study is probably one of the best available in the Delphi literature, featuring extensive use of computer support, and a rational quantitative and probabilistic cost format for couching questionnaire items. But even with all these precautions, considerably more than are encountered in the typical Delphi study, the items incorporate vast areas of ambiguity and represent an array of possible specific events "fitting" into each item. (Recall the "cashless society transaction" item cited previously.) As presently practiced, Delphi is, in many respects, a psychological projective technique for future inkblots.

In his recent experimentation with Delphi procedures in the field of drug abuse, Thompson (1973) underscores the top-priority need for extensive pretesting, and the great difficulties encountered in developing reliable and useful Delphi questionnaires.

> The most challenging aspect of future applications of Delphi techniques to the drug field will almost certainly be the design of a cohesive set of questionnaire items that are both well-posed and useful to the decision-maker. On the one hand, it became apparent during the study that developing concise questions which will be given similar interpretations by all respondents will inevitably involve substantial pre-testing. The usual difficulties in questionnaire design are compounded in the context of drug abuse by disagreement over underlying assumptions, and by the absence of an agreed-upon vocabulary.

The psychological literature on attitude and opinion testing has described an instructive historical process that appears to have gone unnoticed by the Delphi community. After an initial era of free-wheeling, broad-gauged questionnaires on attitudes, covering almost anything of interest, the evolutionary trend has been toward highly specialized attitude and opinion instruments concerned with investigation of specific issues in depth (e.g., Anastasi 1968). In the Delphi context, this means that single items are often of sufficient

complexity to warrant construction of a complete questionnaire dedicated exclusively to that item, exploring major implications and aspects, to better reveal the constellation of opinions to which it gives rise. This permits the development and test of theory to explain and enhance understanding of the item or issue in question.

Delphi Responses

If Delphi questions are ambiguous, then Delphi responses are also ambiguous. Several factors in the structure and dynamics of Delphi responses compound the ambiguity. Among the most important pitfalls is inviting snap answers to amorphous questions.

Delphi investigators rarely analyze and report the effort panelists put into responding to their questionnaires. Solicitation requests and instructions accompanying Delphi questionnaires typically assure the panelist that the forms can be quickly and easily filled out. If not, the investigator runs the risk of massive dropout rates, as occurred with a 19 percent first-round response in Kochman's study (1968). Assurances are often provided that forms should not normally require more than about an hour of the panelists' time for each round. Martino (1972), for example, recommends an upper limit of twenty-five items for Delphi questionnaires.

In the absence of information on panelist effort, the author timed his own responses for two Delphi studies in which he was a panelist. The results showed great variation from item to item, with an average of one minute per item where few comments were written, to an average of about two minutes per item for heavily annotated justifications of responses. The typical sequence would be to read the item; think quickly about key critical factors influencing the forecast; peg the crucial factor, if any, or fall back on a general stereotype, if available; get a crude estimate of its occurrence; and fit the gross estimate into the questionnaire scale. The average one minute pass per item was armchair, top-of-the-head opinion, for strictly ballpark estimates. The average two minute per item session involved spending almost three hours on a lengthy form, with many annotations, which was as much as the author was willing to contribute. Although this experience is not necessarily representative, it is difficult to conceive average speeds very much faster than a fraction

of one minute per item or, at the other extreme, Delphi questionnaires taking more than half a day of the respondent's time for a single round—even if the data are collected in the costly form of a personal interview.

The author is participating in a Delphi study being conducted by Bell Canada to assess future home communication service trends. Response times for the first round have been carefully recorded. The Delphi director indicated in the instructions that the 207-item questionnaire should not require more than an hour to answer. The author took 65 minutes. The average time per question was 19 seconds. The range of response times, for different groups of items (usually one page per group), varied from 40 seconds per question to 9 seconds. The pacing mechanism was "fastest possible reading speed for comprehension and instant response." The author indicated to the Delphi director that he had no confidence in such free-association judgments.

An "analysis" averaging one minute or less for complex forecasts is merely a snap judgement, experts notwithstanding. The results are free-association attitudes toward the future, not analyses of future events. We also tend to get order-of-magnitude responses, particularly for quantitative data. As mentioned earlier, this is particularly apparent in first-round results.

Responses tend to represent stereotyped thinking, as illustrated by the following comments taken from the Nanus, Wooten, Borko study (1973): "This technology is essentially here already, so I'll forecast early"; or "utopian dreamwork, so I'll forecast never"; or "costs are much too high—appear later"; or "no one cares, the public won't buy it"; or "this is a trivial advance"; or "this will kill scientific progress"; or "people will rebel against this invasion of privacy." This is not to deprecate the talent and experience of experts, but most human beings, when placed in a situation where they are regarded as experts, accountable to no one and expected to provide quick answers to complicated questions, are quite likely to lean very hard on stereotypes.

The hypothesis has been advanced in various contexts in previous sections that Delphi forecasting is a form of psychological projection of inkblots of the future. Anyone familiar with psychological projective techniques, such as the Rorschach inkblot test and the Thematic Apperception Test, will appreciate the fundamental basis of such techniques; there are as many "correct" answers as there are respondents. The respondent projects his own emotions, needs, attitudes, imagination, experience, stereotypes, and personal problems into the amorphous stimulus situation, modulated by distinguishable cultural factors related to age and education (Sackman 1952). We saw that group conformity factors prevailed in the autokinetic situation studied by Sherif (1969). It has been pointed out earlier that the typical single-sentence questionnaire format for Delphi is such an unstructured stimulus that it amounts to an inkblot scenario for the future. We have noted that one minute per response is typical of Delphi exercises; it is also characteristic of psychological inkblot tests, where subjects are urged to free-associate to amorphous stimuli. Investigators have collected thousands of responses to the standardized set of ten inkblots originated by Hermann Rorschach, and have tallied responses and published statistical norms of popular and unusual responses. They do not assert that the most popular responses (e.g., butterfly, dancing girls) are "true" or "accurate" responses. By the same token, Delphi investigators have no basis for equating popularity with validity for their "inkblot" results.

Delphi proponents object to this characterization and insist that the statement of Delphi questions in objective, quantitative format yields objective, quantitative results, not amorphous personality projections on arbitrary inkblots. We have already cited order-of-magnitude, log-normal dispersions possible for first-round quantitative estimates. At this point, additional experimental evidence as to the underlying dynamics of such dispersion is presented in support of the hypothesis that Delphi forecasts are often no more than inkblot projections of the future.

McGregor conducted a large-scale study (1938) of psychological determinants of individual predictions of social events. One part of his study was concerned with the impact of the type of information given to respondents when they were asked to make their forecasts. Table 7-1, reproduced from his study, shows results obtained under three conditions in response to the request to estimate the size of the Communist party in the United States for the next year (1936). The three conditions include (1) no information, (2) correct information (e.g., 35,100 members in 1935 with official figures for prior years), and (3) incorrect information where the true figures were multiplied

Table 7-1
Estimates of the Membership of the Communist Party of the United States Under Three Conditions

(1) Without knowledge of the membership for past years, (2) with correct knowledge of the membership for the past five years, and (3) with incorrect knowledge (figures five times too large) of the membership for the past five years.

Interquartile Range	*Without Knowledge N = 246* *100,000 to 1,000,000*	*With Correct Knowledge N = 376* *33,000 to 38,000*	*With Incorrect Knowledge N = 246* *160,000 to 180,000*
Percent Predicting Between 30,000 and 40,000	5	83	0.8
Percent Predicting 50,000 or Less	21	97	2
Percent Predicting Between 150,000 and 200,000	7	0.3	76
Percent Predicting 1,000,000 or More	30	0	0
Mean Prediction	?[a]	35,100	172,000

[a]The calculation of a mean from these estimates would have been a meaningless operation because the distribution revealed no central tendency. There were "clusters" of estimates: (1) below 25,000, (2) around 100,000 (3) around 500,000, and (4) between 1 and 5 million.

Source: Reproduced from McGregor 1938

by five. There were two groups of subjects, 246 in the first and 376 in the second. The first group effectively went through two rounds of this question, initially with no information (first column of results in table 7-1), and later with incorrect information (third column in table 7-1).

The data in table 7-1 reveal several notable results. First, the estimates of the uninformed group were too high by an order of magnitude (first column showing an interquartile range in the hundreds of thousands). Popularity had no relation to accuracy. Note the tendency toward order-of-magnitude clusters at tens of thousands, hundreds of thousands, and millions. With accurate baseline statistics, as expected, the forecasts in the second column of table 7-1 were less variable and far more accurate, more like simple short-term trend extrapolation. The third column in table 7-1,

roughly analogous to a second-round Delphi with feedback, shows how easy it is to manipulate quantitative individual and group opinion to cluster closely around erroneous or misleading data if the situation is sufficiently unstructured. The point of this example is that the inkblot hypothesis applies to quantitative as well as qualitative data for unstructured situations such as quantitative Delphi forecasts of complex social phenomena.

We have already discussed the contamination of opinion with feedback in second and successive rounds, and we need not dwell any longer on the well-established finding that individuals tend to shift their expectations to conform to overt group norms, such as a Delphi median issuing from experts. The iterated expert response to each Delphi item is thus built on snap judgment on the first round followed by various forms of overt and covert conformist pressure in succeeding rounds.

Delphi Results

Delphi group results are merely collections of results for individual questionnaire items. The items are rarely linked together with theoretical or systematic constructs; this potpourri contributes to a mixed bag of findings. As mentioned earlier, item reliability and item validity are typically ignored, making it easier for the uninformed user to accept results at face value. Standard errors of estimates for forecasts and ratings are usually absent in the Delphi tradition, thus giving the final results an aura of precision. Interquartile graphs knock out 50 percent of the sample and the embarrassingly long tails of the extreme nonconformists (see figure 5-1). The final report may include a few anecdotal comments on selected items, but rarely any connected discourse on controversial interpretations. There may also be a few caveats on the limitations of the study.

The presentation of raw frequency distributions of aggregate opinion generates serious problems for the user in the interpretation of the results. Many forecasts may not differ significantly from one another with respect to the null hypothesis for mean or median differences. Many items may be highly redundant, with similar or indistinguishable results, reflecting a pervasive halo effect. The antidote is to test for differences between items in a systematic analysis of variance for items, subjects, and rounds, as mentioned

earlier, to determine main and interaction effects. Redundant items can be discovered through this technique, or through factor analysis of items, as is routine in conventional analysis of questionnaire items. In an unusual exception to standard Delphi neglect of statistical analysis, Dalkey and Rourke (1971) used a type of cluster analysis for quality of life indicators which reduced a very large number of initial raw items to a much smaller number of relatively independent composites or factors. We have no idea how rampant item redundancy and associated halo effects are in the results of the Delphi literature at large, especially with the characteristic absence of techniques equivalent to item factor analyses. It is easier, cheaper, and perhaps more impressive to present the naive user with unprocessed raw data resting on face validity.

After perfunctory qualifications, the investigator makes it quite clear that the experts have pronounced concurred judgment. This is the trump card in the Delphi game. With the apparent tacit agreement not to criticize other Delphi investigations, the results tend to remain unchallenged.

Delphi Epistemology

A fundamental epistemological confusion exists between Delphi method and Delphi results. Practitioners claim that the end result of a Delphi study is a series of expert forecasts of future events or, more broadly, concurred estimates of whatever social attribute is under study. Prior discussion has provided grounds for a very different interpretation of Delphi results.

Delphi items are typically broad, amorphous classes of events, not precisely defined empirical occurrences. Delphi forecasts are opinions about such broad classes of events, not systematic, documented predictions of such events. These opinions are typically snap judgments frequently based on free-association stereotypes. Consensus for such opinion tends to be manipulated consensus to minimize dispersion of opinion. Further, the universe from which items are sampled is typically disregarded and unknown, as are the identity and qualifications of the expert panelists.

Orthodox Delphi epistemology holds that this type of polling procedure produces reasonable and useful forecasts of object events. This worthy goal is not attained. The Delphi process produces manipulated convergence of opinion reflecting ephemeral attitudes of very small samples of unknown individuals. More precisely, Delphi produces transient attitudes about the future, which is quite different from systematic predictions of the future. The epistemological confusion arises from focusing on Delphi results and naively taking them at face value as expert predictions of the future, rather than looking at the underpinning method, which reveals Delphi as an attitude-polling technique dealing in snap judgments of ill-defined issues.

There is a closely related epistemological issue concerned with Delphi validity: the so-called accuracy of Delphi predictions. Observers continue to ask, "How accurate is Delphi? Prove that its accuracy is better or worse than other techniques." These questions presuppose a scientifically replicable calendar/stopwatch concept of forecasting validity where an impartial observer with a stopwatch waits for the objective event to happen, clocks it, and records the time and date of occurrence. This is fine for simple, unambiguous, factual items such as "When will man first land and walk on the moon?" The calendar/stopwatch concept can not be applied to such items as "widespread use of robot services, automated rapid transit, use of computers in tax collection, automated legal information retrieval, etc." (see fig. 5-1, once more). There are as many scenarios for each of these items as there are respondents. How can anyone validate the truth or falsity of an inkblot of the future?

The way out for some Delphi investigators is to ask the experts at a later date whether the forecasts have materialized (Martino 1972). However, this results in another opinion poll, or opinion validated by opinion, not an objective assessment of external events. This amounts to bootstrap validation, Delphi validating itself. Such studies, if conducted rigorously, would provide an indication of longitudinal test-retest reliability (correlation of Delphi with itself over time), not an indication of application validity, which requires correlation against an external criterion.

In limited application areas, such as immediate or very short-range forecasts (excluding the questionable applications to almanac items), Delphi accuracy can be measured. Farquhar's study (1970) of the estimation of software manpower requirements, previously mentioned, is one example. Delphi performed very poorly when compared with face-to-face groups in this case. Delphi forecasting of well-defined short-term economic indices, based on Campbell's

(1966) doctoral dissertation at UCLA, was not shown to differ substantively from simple extrapolation of short-term time-series data. In 1952, Helmer published the results of a Delphi study predicting the results of the 1952 presidential contest between Eisenhower and Stevenson. After four rounds, the seven panelists converged on Stevenson as the winner.

Even this very limited and inconclusive sample of studies indicates that Delphi accuracy will often be untrustworthy and will vary enormously between, and even within, object problems or application areas, reflecting differences in experimenters, "experts" selected, particularly with the ground rules and baseline data made available to them, and numerous other methodological issues. If these frequently untrustworthy and highly variable results over various application areas are characteristic of relatively immediate or short-range estimates, it can be fairly confidently inferred that medium and longer-range results will, by and large, be even more variable in reliability and validity. Recall, for example, Martino's findings cited previously, that the standard deviations of forecasts for independent studies are consistently highly and positively correlated with median expected values of occurrence. (For example, scan the widths of the interquartile "houses" in figure 5-1 with increasing estimated median of occurrence.) Put simply, the farther in the future an event is expected to occur, the more uncertain the prediction is likely to be.

A concrete example illustrates the scope and magnitude of the inkblot problem for Delphi accuracy or validity. Suppose the Delphi questionnaire asks: "When will mass information utilities become commonplace?" The range of "correct" answers for this item, depending upon the scenario projected by the respondent, can literally vary from the Western Renaissance to beyond the year 2000. If "mass information utility" is interpreted to mean mass-produced books, then the answer is somewhere in the sixteenth century, after the introduction and spread of Gutenberg's printing press. If interactive long-distance conversation is the preferred scenario of the respondent, then the advent of the telephone in the late nineteenth century is the answer. If the expert interprets the item to mean mass electronic broadcasting, he would identify the radio as the source and opt for the early 1920s. Another expert might interpret the item as meaning audiovisual broadcasting and list the 1950s for the mass use of television. Another respondent might interpret the item as

involving mass computerized transmission of information and indicate the mid-1970s as the point where computerized information may greatly exceed noncomputerized information over various transmission media. If the item were interpreted as two-way, interactive computer services in the home, as in Baran's (1971) study, the respondent might pick the 1980s. A cosmopolitan expert, accepting the same scenario, but thinking of popular use throughout the entire industrialized world, would place his prediction in the next century. Although this illustration is deliberately extreme, the central point should be quite clear: the Delphi questionnaire format does not lend itself to scientifically objective and externally verifiable statements of future events.

In contrast to the above example, consider more conventional technological forecasting studies under the sponsorship of NASA (Feldman 1965) and the Air Force in Project Forecast (Amsler and Newton 1963). In both of these studies, the authors assembled extensive data on engineering characteristics for specialized forecasting targets: communication satellite output devices (Feldman) and multipurpose long-endurance aircraft (Amsler and Newton). Qualifying specifications and assumptions were spelled out, technical baselines were carefully defined and established, and most likely technological developments were projected. Most results were expressed quantitatively, often in graphic format. The key difference between these results and conventional Delphi results lies in the rigorous technical framework in which the forecasts were embedded. These NASA and Air Force examples illustrate initial steps in the direction of operationally defined predictions essential for scientifically verifiable forecasts.

Thus, when someone asks: "How accurate are Delphi results?" the answer should be: "Accuracy can not be measured for most Delphi items, because changing attitudes and opinions on amorphous issues are not true or false and do not have specific dates at which they occur." Asking for proof or disproof of Delphi accuracy amounts to giving Delphi credit for generating results capable of proof, a property that conventional Delphi, as currently practiced, does not possess.

There is nothing inherently wrong with studying and learning more about opinions concerning the future. Such knowledge is crucial to any intelligent appraisal of the future. But we should not confuse such opinion with seriously considered, qualified, and

documented predictions of well-defined future developments. Attitudes and opinions change, and fresh sampling in real time is needed to track such changes. And the sampling must be explicit in terms of subject populations if any systematic inferences are to be made.

Delphi Isolationism

The history of Delphi reveals a highly exploratory and tentative technique that was never validated. Delphi was obviously full of problematic issues and potentially serious flaws and was treated with some measure of caution and skepticism by its Rand originators before the Gordon-Helmer study (1964) catapulted the technique into international prominence. After that point, the shaky hypotheses on which Delphi rested were apparently transformed into axioms, and Delphi was promoted as an established, proven technique.

Only relatively recently have Dalkey and some of his co-workers made attempts to demonstrate the validity of Delphi, as reviewed in this report, primarily with almanac-type items and nonexpert panelists such as college students. These efforts, and spotty returns from a small number of other studies mentioned in this review, provide no scientific validation of Delphi. This history of early experimentation and tardy efforts to assess validity reflects a pattern of isolationism from the mainstream of behavioral research.

Delphi has led a protected existence for the decade it has been actively pursued. From exploratory and tentative beginnings at Rand, it has spread from government to industry and academia, and diversified from scientific and technological forecasting to policy studies and planning, to quality of life assessment, and is being touted as the emerging nexus for human communication and decision making (Turoff 1972). Droves of eminent people and experts from all callings have lent their name, time, and effort to hundreds of Delphi investigations. All this, and undoubtedly more to come. Why?

In part, because there has been virtually no critical literature. The roots of this criticism-free development of Delphi are found in two sources: the isolation of Delphi from the mainstream of relevant behavioral science, and the rapid concurrent emergence and growth

of futurism. For various reasons, Delphi originators and subsequent Delphi practitioners have shied away from psychometric and opinion survey specialists who could have professionalized Delphi as an opinion-polling technique along the lines previously suggested in connection with the discussion of the APA test manual and social science standards. Isolation is attested by the fact that there are virtually no listings of Delphi studies in the *Psychological Abstracts*, as our literature review revealed. The proof of this isolation is the disregard and unconcern for professional questionnaire standards in Delphi practice that has been heavily documented in this study.

The reasons for such isolation are not hard to find. The professional standards would immediately transform Delphi from a cheap and easy shortcut technique to a far more difficult, expensive, and time-consuming procedure. Unprepared and untrained Delphi investigators would have to develop new skills in psychometrics, opinion sampling and polling, and experimental design with human subjects, and would lose considerable control over the technique if experts in these skill areas were taken seriously.

Delphi practitioners and many futurists, broadly considered, identify themselves as interdisciplinarians. They sought to enlist the necessary diversity of skills to assessments of the future. This is most commendable, if taken seriously. The place to begin, however, is with the disciplines vital to the method. This was never done with Delphi.

Neither the originators of Delphi nor subsequent practitioners have been willing to attempt to establish rigorous standards and to police the Delphi literature by discriminating between better and poorer work. This has contributed to the spate of crude Delphi studies generated by neophytes.

This lack of standards is characteristic of new disciplines going through early growth. Futurism has not been heavily pursued for much more than a decade. Delphi played no small part in getting futurists on the map by dignifying forecasting with its seemingly impressive ritual for obtaining expert consensus. Other methods, such as brainstorming, scenarios, gaming, input-output analyses, contextual mapping, simulation, and morphological analyses also experienced rapid growth during this period, contributing largely undisciplined exploratory techniques to futures forecasting and planning. Each technique needs adversary checks and balances for healthy growth, and futurism as a whole needs to develop minimal

professional standards and a vigorous critical literature representing more authentic interdisciplinary work.

Results of the Analysis

This portion of the study, concerned with analysis of the specific and unique assumptions and principles of Delphi, as distinct from opinion questionnaires and human experimentation broadly considered, was organized under ten key questions formulated at the outset. The analysis suggests the following answers to the ten questions for conventional Delphi:

1. The Delphi concept of the expert, and its claim to represent valid expert opinion, is scientifically untenable and overstated. As summarized by Professor Haythorn, an external technical reviewer of this report:

> The procedure by which the selection of subjects occurs is not properly explicated, the exact nature of the panel of experts is often left unspecified, and the implicit assumption that results obtained using conventional Delphi with a panel of experts is better than or different from results that would be obtained using another population has not been empirically established.

2. Delphi claims of the superiority of group over individual opinion, and of the superiority of remote and private opinion over face-to-face encounter, as well as their counterstatements, are unproven generalizations.

3. Delphi consensus is specious consensus. As succinctly stated by Professor Haythorn:

> The group process used in Delphi rounds is quite similar to the techniques used in social psychological research to study group conformity, rejection of deviant opinion, and deindividualization, all of which have been shown to be counterproductive with regard to the quality of group decisions.

4. Delphi questions are likely to be vague.
5. Delphi responses are likely to be ambiguous.
6. Delphi results probably represent compounded ambiguity.
7. Delphi is primarily concerned with transient collections of snap-judgment opinions of polled individuals from unknown samples, which should not be confused or equated with coherent predictions, analyses, or forecasts of operationally defined and systematically studied behaviors or events.
8. Delphi anonymity reinforces unaccountability in method and findings.
9. Delphi systematically discourages adversary process and inhibits exploratory thinking.
10. Delphi has been characterized by isolation from the mainstream of scientific questionnaire development and behavioral experimentation, and has set an undesirable precedent for interdisciplinary science in the professional planning and policy studies community.

8 Epilogue

The following sixteen conclusions sum up the evaluation of conventional Delphi in regard to its method and application. This report finds conventional Delphi:

1. Often characterized by crude questionnaire design.
2. Lacking in minimal professional standards for opinion-item analyses and pilot testing.
3. Highly vulnerable on its concept of "expert" with unaccountable sampling, and in the selection of panelists, expert or otherwise.
4. Abdicating responsibility for item population sampling in relation to theoretical constructs for the object area of inquiry.
5. Virtually oblivious to reliability measurement and scientific validation of findings.
6. Capitalizing on the fallacy of the expert halo effect.
7. Typically generating snap answers to ambiguous questions, representing inkblots of the future.
8. Seriously confusing aggregations of raw opinion with systematic prediction.
9. Capitalizing on forced consensus based on group suggestion.
10. Unwittingly inhibiting individuality and any adversary process by overtly and covertly encouraging conformity and penalizing the dissident.
11. Reinforcing and institutionalizing premature closure, using a highly questionable ritual for conducting opinion studies that tends to inhibit more scientific approaches.
12. Giving an exaggerated illusion of precision, misleading uninformed users of results.
13. Indifferent to and unaware of related techniques and findings in behavioral science in such areas as projective techniques, psychometrics, group problem solving, and experimental design.

14. Producing virtually no serious critical literature to test basic assumptions and alternative hypotheses.
15. Denigrating group and face-to-face discussion, and claiming superiority of anonymous group opinion over competing approaches without supporting proof.
16. Encouraging a shortcut social science method that is lacking in minimum standards of professional accountability.

Final Evaluation

Two alternative final recommendations were considered as conclusions of this evaluation. One was to seek to upgrade Delphi by recommending higher standards, more consistent with scientific method in the collection, analysis, and use of questionnaire data. The other was to conclude that the assumptions and principles on which conventional Delphi is based are so unscientific and inherently misleading that they preclude any attempts to improve the technique. This second alternative was tantamount to a recommendation to drop Delphi completely.

The evidence adduced in this study clearly indicates that the massive liabilities of Delphi, in principle and in practice, outweigh its highly doubtful assets.

As the preferred alternative to conventional Delphi, professionals, funding agencies, and users are urged to work with psychometrically trained social scientists who can apply rigorous questionnaire techniques and scientific human experimentation procedures tailored to their particular needs. It is recommended that conventional Delphi be dropped from institutional, corporate, and government use until its principles, methods, and fundamental applications can be experimentally established as scientifically tenable.

Beyond Delphi

Some will grant the very shaky opinionative structure of Delphi and insist that Delphi was never really put forth as science, but merely as a heuristic vehicle for exploring vague and unknown future issues otherwise inaccessible. They might insist that Delphi as an exercise has generated many insights and has been well received.

Jolson and Rossow (1971) have commented on the heuristic value of Delphi in facilitating communication in the corporate environment. Reisman et al. (1969) have noted the communication potentials of Delphi for community participation in evaluating alternative social services. Even as a heuristic exercise, it would be highly advisable to mix iterative polling with varying forms of quantitative and qualitative feedback, personal confrontation where feasible, cultivated development of adversary positions as opposed to consensus, and controlled variations in the type and level of anonymity. As we have seen, there is nothing sacred in the Delphi process; all basic assumptions, particularly in informal exercises, should be systematically challenged, examined, and tested with other eclectic approaches and tailored to the unique mission and needs of the object problem.

Brainstorming, if done properly, is fun, generates many insights, and can be well received. Advocates of brainstorming no longer present their results as finished products. Practitioners of Delphi publish results in journals, as master's and doctoral dissertations (e.g. Kochman 1968; Campbell 1966; and Weaver 1969) as major corporate reports, (e.g., North and Pyke 1968), as significant social indicators for national and international planning (e.g., Dalkey, Rourke, Lewis and Snyder 1972; Bjerrum 1968), as results worthy of weighty consideration, the embodiment of balanced expert opinion.

There is a vast difference between Delphi as an informal forecasting exercise among questionnaire respondents, and Delphi as the authentic embodiment of thoughtfully concurred expert opinion, wherever it is applied. Nanus, Wooten, and Borko (1973), in a relatively ambitious Delphi study on the social impact of the multinational computer, make it clear that no claims are made for the reliability or validity of their Delphi results. Delphi was used for strictly exploratory purposes in an uncharted domain; "The authors chose to use Delphi with realization that the results would be more in the nature of a structured 'brainstorming' session with noted thinkers than a scientific exercise in prediction" (p. 11).

The rejection of conventional Delphi recommended here should not in any way be construed as denying the growing and urgent need of society to learn and understand more about the future. Perhaps the greatest of all human rights is the right to help shape and determine one's own and society's future. We need to know far more about human attitudes toward future developments.

It is to be hoped that forthcoming opinion polls will systematically sample attitudes toward the future from all segments of the population for more effective and more humanistically informed social planning. Delphi, with its exclusive reliance on small coteries of "experts," has unwittingly fostered another form of elitism to set the pace and formulate the pattern for attitudes toward the future.

The originators of Delphi had the right instincts in responding to growing and pressing needs to enlist the active participation of geographically distributed professionals to work in concert assessing unknown and complex problems. Perhaps their most significant insight was the concept of physically distributed teams building a cumulative base of knowledge through the mechanism of temporally spaced interaction and feedback. Although this concept responds to a strongly felt social need, the implementation has been counterproductive. The originators arrived at premature closure along the lines of an iterative ritual producing ambiguous results.

Instead of testing a great variety of flexible alternatives, the method zeroed in on iterative statistical group response. The alternatives could have branched out into structured adversary procedures, including dialectical planning (Mason 1969), adversary polling between groups with vested interests, as in SPRITE (Bedford 1972), iterative online teleconferences (Sackman and Citrenbaum 1972), and eclectic mixtures of confrontation and isolated responses (Heller 1969; Weaver 1972). All of these areas need vigorous experimental work.

It is beyond the scope of this analysis to enter into a systematic review of areas of inquiry related to Delphi and possible offshoots that might lead to useful advances in method and findings. Suffice it to say that many research opportunities exist for teleconferencing, iterative polling, the analysis of human attitudes toward the future, cooperative problem solving among geographically dispersed individuals, and the social dynamics of real, not specious consensus, which should be based on a profound understanding of the adversary process in its own right.

Consumers of information on the future need far better advice and protection from contributing professionals than they have gotten to date. The future is far too important for the human species to be left to fortune-tellers using new versions of old crystal balls. It is time for the oracles to move out and for science to move in.

Appendix
Semiannotated Delphi Bibliography

This bibliography includes standard and annotated citations to the Delphi literature. Much of this material was assembled by Barbara Quint of the Rand Library staff. Annotations are included as available from our sources; entries are arranged alphabetically by author.

We were greatly aided by Delphi listings made available to us from the following sources:

1. *Delphi and Long-Range Forecasting*, SB-1019. The Rand Corporation, Santa Monica, California, 1972. Annotations from this source are indicated by (R) at the end of the listing.
2. *Selected Bibliography on Delphi Literature*, Institute for the Future, Menlo Park, California, 1972.
3. Pill, Juri, "The Delphi Method: Substance, Context, a Critique and an Annotated Bibliography," *Socioeconomic Planning Science*, vol. 5 (1971), pp. 57-71, Annotations from this source are indicated by (P) at the end of the listing.
4. Turoff, Murray, "Delphi and Its Potential Impact on Information Systems," *AFIPS Conference Proceedings*, vol. 39, AFIPS Press, Montvale, New Jersey, pp. 317-326, 1971.
5. *Annotated Delphi Bibliography*, provided by L.H. Day and Michael T. Bedford, Bell Canada Business Planning Group, Montreal, Canada, 1973. Annotations from this source are indicated by (B) for Bell Canada at the end of the listing.
6. A search through various standard indexes in the Rand Library.

The Bell Canada bibliography and the Rand bibliography provided the most extensive annotated listings. The primary focus of the Bell Canada entries is on corporate applications of Delphi. These entries include listings and cross-references for corporations using Delphi which are retained for the convenience of the reader. The Rand entries primarily cover the historical and methodological literature. The accumulated sources should enable the reader to obtain a reasonably balanced picture of the Delphi technique with numerous applications over many areas in a single alphabetical listing.

ADELSON, M., M. ALKIN, C. CAREY, and O. HELMER
"Planning Education for the Future: Comments on a Pilot Study"
American Behavioral Scientist, vol. 10, no. 7, (1967), pp. 1-31.

The character of American education is determined by many related decisions. Improving it will require a broad base of participation within and outside of school systems. This requirement implies a need for generating and disseminating information about education, and for devising procedures for bringing informed judgment to bear on the decision process in a regularized way. It may be as important to improve the decision process in education as to modify any of the specific features of contemporary schooling. The trend toward systematizing or rationalizing the decision process seems promising, although there is a need to avoid centralized control of the process of developing new citizens who are to live in a democratic society. The future role of the federal government in American education is one of the deep residual issues.

ADELSON, M., M. ALKIN, C. CAREY, and O. HELMER
"The Education Innovation Study"
American Behavioral Scientist, vol. 10, no. 7 (1967), pp. 8-12, 21-27.

Most of the contents of this paper are discussed also in Rand P-3499. It is a description of the process and output of a seminar group that studied innovation in education and came up with some recommendations and proposed priorities. The Delphi technique was used, but the emphasis in the discussions is on the results rather than the methodology. There is an interesting work flow chart which would be useful for any studies of this type. (P)

AIL

See paper by PACKARD that describes the use of Delphi at AIL for forecasting the development in the LSIchip industry.

ALDERSON, R.C., and W.C. SPROULL
"Requirement Analysis, Need Forecasting, and Technology Planning Using the Honeywell PATTERN Technique"

Technological Forecasting and Social Change, vol. 3 (1972), pp. 255-65.

The authors describe the development and use of the Honeywell PATTERN technique—Planning Assistance Through Technical Evaluation of Relevance Numbers. This technique may be regarded as a distant cousin to Delphi, since groups of experts are used to develop consensus on the relevance numbers for the projects under consideration. Although the technique was developed for military purposes, the article uses examples of a personal transportation decision and a biomedical study conducted by Honeywell (B)

ALLPORT, GORDON
Becoming
Yale University Press, New Haven, 1955.

AMARA, R.C., A.J. LIPINSKI
"Some Views on the Use of Expert Judgment"
Technological Forecasting and Social Change, vol. 3, no. 3, 1972.

AMENT, R.H.
"Comparison of Delphi Forecasting Studies in 1964 and 1969"
Futures, vol. 2, no. 1, March 1970. (Also Institute for the Future, P-9.)

AMERICAN PSYCHOLOGICAL ASSOCIATION
Standards for Educational and Psychological Tests and Manuals
Washington, D.C., 1966. (Revised in 1974.)

AMSLER, R.C., and J.S. NEWTON
Multipurpose Long Endurance Aircraft (MPLE) Airplane Design Analysis
Northrop Corporation, NOR-63-109, June 1963.

ANASTASI, A.
Psychological Testing
Macmillan, New York, 1968.

AT&T.
The Future of the Telephone Industry
Sponsor of the Institute for the Future Study R-20.
(See Baran and Lipinski.)

Index

About the Author

Harold Sackman is Senior Information Scientist with The Rand Corporation, working in the areas of policy studies, telecommunications, and man-computer problem solving. He has also served as Professor and Head of the Department of Computer Science at Kansas State University and as Senior Research Leader at System Development Corporation. Dr. Sackman has published nine books concerned with planning and policy studies, computers and society, man-computer problem solving, and mass information utilities. A pioneer in the scientific study of the human use of computers, he has been Chairman of the AFIPS (American Federation of Information Processing Societies) Committee on Social Implications of Computers since 1970. Dr. Sackman received the Ph.D. in psychology from Fordham University in 1953. He is a member of the American Psychological Association, the Association for Computing Machinery, the American Association for the Advancement of Science, and Phi Beta Kappa.

Technological Forecasting for Decision Making

Second Edition

Joseph P. Martino
University of Dayton Research Institute, Ohio

North-Holland
New York • Amsterdam • Oxford

Contents

Preface to the Second Edition

Since the first edition of this book there have been significant changes in the state of the art of technological forecasting. These include refinements and improvements on older techniques, as well as some completely new techniques. In addition, there has been a change in emphasis among techniques; for instance, a decade ago computer models were hardly used, whereas their use is now widespread. This new edition brings together the new techniques and the changed emphases and integrates them with the older techniques.

Another important change since the first edition is the increased use of technological forecasting for a variety of applications. It has become widely accepted in industry, government, and universities, and the chapters on applications have been updated to reflect this.

Finally, this second edition has benefited from considerable feedback from users, both individual readers and those who have used it as a text in formal classes. Some of the background material and historical illustrations have been shortened or eliminated to sharpen the focus on the more important points. In addition, lengthy derivations have been omitted where they did not contribute to an understanding of the techniques presented. Readers can refer to the original literature to find these derivations, and those interested only in applications of techniques will find them presented more compactly here.

Joseph P. Martino

Chapter 2

Delphi

1. Introduction

Formal forecasting methods are intended to replacement subjective opinion with objective data and replicable methods. However, there are three types of circumstances under which expert opinion will always be needed. (Note, however, that the selection of a particular "objective" method may involve some subjectivity and implicit assumptions. The importance of these assumptions will be discussed in Chapter 17.)

The first type is when no historical data exist. In technological forecasting this usually involves new technologies. Despite the absence of historical data, a forecast must often be prepared. Expert opinion is then the only possible source of a forecast.

The second type is when the impact of external factors is more important than the factors that governed the previous development of the technology. These external factors may include decisions of sponsors and opponents of the technology, or changes in public opinion. In such a case data about the past may be irrelevant. Expert opinion may be the only possible source of a forecast.

The third type is when ethical or moral considerations may dominate the economic and technical considerations that usually govern the development of technology. These issues are inherently subjective, and expert opinion may be the only possible source of a forecast.

Given that expert opinion is needed, how is it to be obtained? The problems of expert opinion may be overcome to some extent by using several experts—two heads are better than one. In a group of experts individual biases may be canceled out, and the knowledge of one member may compensate for another's lack of it.

On the other hand, "a camel is a horse designed by a committee." A forecast designed by a committee might be equally grotesque. What is needed is some way to obtain the benefits of a committee while minimizing the disadvantages.

2. Advantages of Committees

The first major advantage of a committee is that the sum of the information available to a group is at least as great as that available to any individual member. Adding members to a group does not destroy information. Even if one member knows more than the rest put together, this does not reduce the total information available to the group; the others may still make useful contributions. If the group has been chosen to contain only people who are experts in the subject, the total information available to the group is probably many times that possessed by any single member.

The second major advantage is that the number of factors that can be considered by a group is at least as great as the number which can be considered by a single member. This point is at least as important as the first. Studies of forecasts that have gone wrong show that one very common cause of failure is neglecting to take into account important factors outside the technology being forecast, which in the long run turned out to be more significant than those internal to the technology. This advantage of a group is therefore very important.

3. Disadvantages of Committees

The first major disadvantage of a group is that there is at least as much misinformation available to the group as there is to any single member. One reason for using a group is the hope that the misinformation held by one member may be canceled out by valid information held by another. However, there is no guarantee that this will take place.

The second major disadvantage is the social pressure a group places on its members—pressure to agree with the majority even when the individual feels that the majority is wrong. This is especially true in the production of group forecasts. One member may well give up presenting certain relevant factors if the remainder of the group persists in taking a contrary view.

The third major disadvantage is that a group often takes on a life of its own. Reaching agreement becomes a goal in itself, of greater importance than producing a well thought out and useful forecast. Group forecasts may thus be only a watered-down least common denominator that offends no one, even though no one agrees strongly either.

A fourth major disadvantage is the influence that the repetition of ar-

guments can have. Experiments with small groups show that often it is not the validity but the number of comments for or against a position that carries the day. A strong vocal minority may overwhelm the majority by pushing its views vigorously, even though the arguments may have little objective merit.

A fifth major disadvantage of groups is their vulnerability to the influence of dominant individuals. One individual, by active participation in debate, by putting ideas forward with a great deal of vigor, or through a persuasive personality, may have an undue influence on the group's deliberations. Such an individual may get his or her way simply by wearing down the opposition with persistent argument.

A sixth disadvantage of groups is that members of a group may come to have vested interests in certain points of view, especially if they have presented them strongly at the outset. Their objective becomes one of winning the remainder of the group over, rather than reaching a more valid conclusion. Such members may be impervious to the facts and logic of the remainder of the group. They will concentrate only on winning the argument.

A seventh disadvantage of groups is that the entire group may share a common bias. This often arises from a common culture shared by the members—especially a subculture peculiar to the technology in which the members are experts. The presence of a common bias nullifies the advantage of a group in canceling biases.

4. The Delphi Procedure

Delphi is intended to gain the advantages of groups while overcoming the disadvantages. It was originally developed at the Rand Corporation as a means of extracting opinion from a group of experts. Its first public presentation in a Rand report dealt with a series of technological forecasts, which led to the misunderstanding that Delphi was primarily a technological forecasting method. It is not. It can be used for any purpose for which a committee can be used. While the emphasis here is on technological forecasting, other uses of Delphi are discussed in Linstone and Turoff (1975).

Delphi has three characteristics that distinguish it from conventional face-to-face group interaction: (1) anonymity, (2) iteration with controlled feedback, and (3) statistical group response.

Anonymity. During a Delphi sequence the group members usually do not know who else is in the group. The interaction of the group members is handled in a completely anonymous manner through the use of questionnaires. This avoids the possibility of identifying a specific opinion with a particular person. The originator can therefore change his mind without publicly admitting he has done so. In addition, each idea can be considered on its merits, regardless of whether the group members have high or low opinions of the originator.

Iteration with Controlled Feedback. Group interaction is carried out through answers to questionnaires. The group moderator extracts from the questionnaires only those pieces of information that are relevant to the issue and presents these to the group. Each group member is informed only of the current status of the group's collective opinion and the arguments for and against each point of view. Group members are not subjected to a harangue or an endless restatement of the same arguments. Any viewpoint can be presented to the group, but not in such a manner as to overwhelm the opposition by sheer repetition. The primary effect of this controlled feedback is to prevent the group from taking a life of its own. It permits the group to concentrate on its original objectives rather than self-chosen goals such as winning the argument or reaching agreement for its own sake.

Statistical Group Response. Typically a group will produce a forecast that contains only a majority opinion; it will represent simply that viewpoint on which a majority of the group could agree. At most there may be a minority report. There is unlikely to be any indication of the degree of difference of opinion that existed within the group. Delphi presents instead a statistical response that includes the opinions of the entire group. On a single item, for instance, group responses are presented in statistics that describe both the "center" of the group opinion and the degree of spread about that center.

5. Conducting a Delphi Sequence

The following description is of "classical" Delphi as originated at Rand. Variations from this base line will be taken up in a later section.

Before describing Delphi, some definitions are required. A Delphi sequence is carried out by interrogating a group of experts with a series of questionnaires. Each successive questionnaire is a "round." The term *questionnaire* may be misleading, however. The questionnaires not only ask questions, but provide information to the group members about the degree of group consensus and the arguments presented by the group members for and against various positions. The questionnaire is the medium for group interaction. The set of experts taking part in the Delphi is usually referred to as a "panel." In large Delphis there may be subgroups devoted to specific specialties. These subgroups may be identified by subject, such as the "electronics panel." Either the entire set of experts or a subgroup may be referred to as a panel. Context usually

provides a guide as to which use is meant. The person responsible for collecting the panel responses and preparing the questionnaires is called the "moderator."

Delphi will be described in terms of rounds. Each round calls for somewhat different activities on the part of the panelists and moderator. Before the first round there must be preliminary activities such as clarifying the subject and explaining the methods. After these preliminaries the first round can begin.

Round One. The first questionnaire is completely unstructured. The panelists are asked to forecast events or trends in the area for which the panel was assembled. This has some disadvantages, which will be discussed later, but it also has some significant advantages. The panelists have been selected because of their expertise in the area to be forecast. They should know much more than the moderator does about that area. If the first questionnaire were too structured, it might prevent the panelists from forecasting some important events of which the moderator might not be aware.

The questionnaires are returned to the moderator, who then consolidates the forecasts into a single set. Similar items must be combined; items of lesser importance must be dropped to keep the list at a reasonable length; and each event must be stated as clearly as possible. The list of events then becomes the questionnaire for the second round.

Round Two. The panelists receive the consolidated list of events and are asked to estimate the time of occurrence for each event. The estimate may be a date; it may be "never," if they think that the event is impossible; or it may be "later" if some time horizon has been specified for forecasts and they believe that it will occur later than that horizon.

The moderator collects the forecasts from the panel and prepares a statistical summary of the forecasts for each event. This usually consists of the median date and the upper and lower quartile dates for each event (for definitions see Appendix I). The third questionnaire consists of the set of events and the statistical summary of the forecasts.

Round Three. The panelists receive the questionnaire with events, medians, and quartiles. They are asked to prepare new forecasts for each event, either sticking with their previous forecast or making a new one. If their forecasts fall in either the upper or lower quartile (that is, if they are "outliers"), they must present reasons why they believe they are correct and the majority of the panel incorrect. Their reasons may include references to specific factors the other panelists may be overlooking, facts the other panelists may not be considering, and so on. The panelists

are just as free to advance arguments and objections as they would be in a face-to-face group; the only difference is that their arguments are written and anonymous.

When the moderator receives the third-round responses, he or she prepares a statistical summary of the forecasts, as well as a consolidated summary of the panel's reasons for advancing or delaying the forecasts. Similar arguments are combined and lengthy arguments summarized. (Fortunately the need to write the arguments often forces the panelists to be concise.) The questionnaire for the fourth round consists of the list of events, the medians and quartiles for the third round, and the summary of arguments for changing the forecasts of each event.

Round Four. The panelists receive the events and dates and the reasons given for changing their estimates. They are asked to take the reasons into account and make new forecasts for each event. Depending upon the needs of the moderator, they may be asked to justify their position if their forecasts fall in the upper or lower quartile. In addition, the moderator may invite comments from all panelists on the arguments given during the third round.

Upon receiving the forecasts from the panelists, the moderator again computes medians and quartiles and, if comments were requested, consolidates and summarizes them. (If the moderator does not plan to analyze the arguments, there is no point in asking for them.) In some cases, when the panel has not been able to reach a consensus, the moderator may well be interested in the arguments on both sides. In such cases the moderator should ask for comments and be prepared to analyze them.

The date forecast for each event is the median date on the fourth round. In addition, the moderator can determine the amount of disagreement in the panel on the final round from the difference between the quartile dates. The comments on each event provide a summary of those factors the panelists believe are important and that may affect the forecast. The output of a Delphi thus contains a great deal more information than is usually obtained from a committee. In addition, the nature of Delphi focuses this information on the topics of interest to the moderator and organizes it in a readily understandable manner.

Ordinary committees are judged as successes if they reach agreement or consensus. Indeed, committee action is designed to achieve consensus and may force a false one. Delphi is intended to display disagreement where it exists and search for the causes. Delphi sequences are judged as successes when they reach stability, that is, no further change of opinion, with the reasons for divergence clearly displayed. Thus if a particular item reaches stability before the fourth round, it may be dropped. In some cases, however, it may be necessary to restate an event,

to split an event into distinct subevents, or to combine separate events in order to obtain agreement on what is really at issue, and thereby reach stability.

General experience is that there is convergence of the panel estimates during the sequence of rounds. The panel members will usually have widely varying estimates on each event on the second round. However, as the panelists offer their reasons for shifting the estimates, the subsequent estimates tend to cluster near preferred dates. This convergence results from actual transfer of information and interaction among the panel members.

Panel members do not always shift their opinions under the influence of the arguments of other panelists. Delphi panelists have just as much opportunity to stick with their original views as do members of a face-to-face group. The advantage of Delphi is that panel members can shift position without losing face when they see convincing reasons from other panel members for a shift of their estimates.

6. Variations on Delphi

Since Delphi was first publicly announced there have been numerous variations on the basic procedure. Some of these are described briefly below.

Providing an Initial List of Event. Classical Delphi has been described as "starting with a blank sheet of paper." While this has advantages, it also seems to bother some panelists, who find themselves confused by the unstructured situation. Some users of Delphi have started with an initial list of events generated by some process before the start of the Delphi. The panelists may be asked to make forecasts for these, effectively going immediately to round two; alternatively, they may be asked to suggest additional events. The augmented event set then becomes round two.

Beginning with a Context. The exact course of the development of a technology will depend upon external political and economic conditions. When these are important, the forecasts will depend upon the assumptions made about these external conditions. If the panel is composed of technology experts, they should not be expected to forecast these economic and political conditions as well. Hence it may be desirable to obtain a political and economic forecast and present this to the panelists prior to round one. This provides the panelists with a common context for their forecasts of technology. If the economic and political forecasts are in error, the resulting technological forecast will also be in error. However, this problem cannot be avoided by failing to provide a context. Doing so

simply means that the technology experts will make their own political and economic forecasts. Providing a context can be especially useful in industrial applications of Delphi when a panel of experts has been chosen from the company's technical staff. A context provided by the company's sales, marketing, and top management personnel can provide a helpful guide to the technical specialists on the panel.

Number of Rounds. Classical Delphi includes four rounds. Some Delphis have taken as many as five rounds. Experience indicates that four rounds is usually sufficient. Round four can be deleted if the moderator sees no need to obtain rebuttals to the arguments presented in round three. Round one can be omitted if the panel is started off with a list of events. Thus in some cases two rounds may be sufficient. Since Delphi provides advantages over face-to-face groups, it should be used if at all possible, even when a full four rounds cannot be used. Even two rounds may be better than the use of a single expert or a face-to-face panel.

Multiple Dates. In the classical Delphi each panelist provides one forecast for the date of an event. In some cases this is specified as the date by which the event is 50% likely. In other applications of Delphi, however, panelists may be asked to provide three dates: In addition to the 50% date, they may be asked to provide "barely possible" and "virtually certain" dates. These may be quantified as 10, 50, and 90% probability estimates or some other suitably chosen probabilities. The statistical group response is then obtained by taking the median date for the 50% estimates. The degree of disagreement in the panel is represented by the spread between the median dates for the low-likelihood and high-likelihood dates.

Computerization. Computerized analyses of Delphi results are quite common, especially for Delphis with many people and several panels. However, computerization can go well beyond processing the Delphi responses. In some Delphi sequences panelists have used remote computer terminals to participate. The terminals are connected to a central computer that keeps track of the current status of each event and the last estimate made by each panelist. A panelist participates by "logging on" to the computer via a terminal; the computer displays the median and quartiles of the current estimates of the panelists, reminds the panelist of his or her last estimate and asks whether this estimate should be changed. This approach does away with the round structure. Panelists may log on as often as they choose. Some will do so more frequently than others; some panelists will change their estimates frequently, while others will permit theirs to stand for a longer time. This "real-time, on-line Delphi" can allow participants to achieve a consensus much more rapidly

than via written questionnaires sent through the mail. Application of this approach is currently limited by the availability of computer terminals. As terminals become more widely available, this approach to Delphi is likely to become much more widespread.

Delphi with Partial Anonymity. Delphi is sometimes used in face-to-face groups: Arguments are thus made publicly, while estimates are still made anonymously through secret voting. The panelists discuss an event and then make their forecasts. This may go through several rounds as panelists offer reasons why the others should change their forecasts. Paper ballots are often used; however, an electronic device known as a "Consensor" is sometimes employed instead. The Consensor consists of a small computer, a TV-like display screen, and a dozen or so control units connected to the computer by cables. The control units consist of a numbered scale and a knob that can be rotated to one of the numbers. Each participant can "vote" by setting the knob on his control to the number representing his estimate. When all participants have voted, the computer prepares a statistical analysis of the estimates and displays it on the screen. The display may be a bar graph or some other suitable picture that shows the "center" of the estimates and the dispersion about that center. With the Consensor votes may be taken quickly at any point in the discussion. Participants can quickly see how much consensus has been reached; they can decide whether further discussion is worthwhile. The discussion is public, but the cables to the control units can be scrambled so that the voting is untraceable, and the control units can be concealed from the rest of the participants to maintain the anonymity of individual estimates.

7. Delphi as a Group Process

People frequently ask about the accuracy of Delphi; however, this question is misdirected. Delphi is based solely on expert opinion. The accuracy of the forecasts is only as good as the opinions that go into the forecasts. Since Delphi is used when expert opinion is the best forecast available, the proper issue is whether Delphi is a better method for extracting opinion from a group of experts than is any other method.

Much of the work on the accuracy of Delphi, as a group process, goes back to Dalkey's experiments with "almanac" questions. Dalkey asked his subjects questions to which there was a known numerical answer; moreover, the questions were ones to which the subjects were unlikely to know the answer but about which they could make informed judgments. A typical question was, How many telephones were there in Africa in 1965? In Dalkey's experiments each participant made an estimate for each question. Then the participants either received anonymous feedback, as in Delphi, or participated in a face-to-face discussion. The feedback, whether anonymous or face to face, was then followed by another set of individual estimates. The findings were that, more often than not, the anonymous feedback made the median of the second round better. The face-to-face discussion, more often than not, made the median of the second round worse.

Another view of Delphi as a group process comes from an experiment reported by Salancik (1973). The panelists took part in a Delphi to forecast applications for computers. For instance, the panelists were asked to forecast the date by which it was 50% likely that one-half of all physicians would be using computers in a particular application. In addition to his forecast date, each panelist was asked to give reasons for his estimate.

The reasons given by the panelists were categorized as dealing with benefits, costs, or feasibility. Whether a response gave a positive or negative view of the benefits, for instance, it was categorized as a statement of benefits. The net number of benefit statements (number stating positive benefit minus the number stating negative benefit) was computed for each application. The same was done to obtain a net number of statements regarding feasibility and (low or acceptable) cost. The median date forecast for each application on the second round was then regressed on the net number of statements in each of the three categories. The regression equation explained 85% of the variance in median dates. This analysis showed that Delphi panels do assimilate the comments from panel members into their aggregate estimates. Delphi is not simply a repeated poll; group interaction does take place.

Another important finding is that first-round estimates have a log-normal distribution; this is true for both almanac-type questions and actual forecasts. A typical result is shown in Figure 2.1, which is based on a total of 19,000 separate forecasts. For the first-round responses the mean and standard deviation were computed for each event. The response to each event was then standardized by first subtracting the mean for that event and dividing by the standard deviation. The standardized responses were then pooled and stratified. The cumulative frequency was plotted against the response on log-normal paper. The implication of this finding is that people tend to think in ratios; so an estimate half the true value is considered to be of the same size error as an estimate twice the true value. Thus the logarithms of the ratios of estimates to mean values are normally distributed.

The log-normality of first-round estimates shows that estimation is a "lawful" behavior governed by rules that produce regularity in the estimates. The relationship between net numbers of arguments and the median date of the forecast shows that the initial estimates are subjected to a genuine group interaction. Dalkey's findings show that the group process in Delphi does an efficient job of extracting information from a

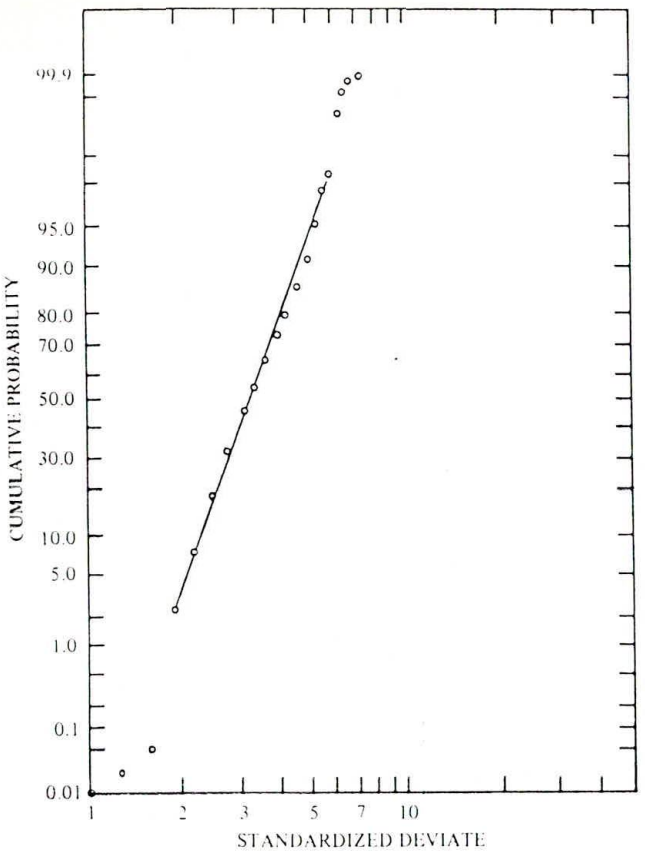


Figure 2.1. Cumulative probability distribution of standardized deviates for 50% likelihood estimates. A value of 3.7 was added to each deviate to make all the results positive.

panel as compared with face-to-face interaction. While none of these findings can "validate" a Delphi forecast, taken together they indicate that when it is necessary to use expert opinion, Delphi is a good way of getting it.

8. The Precision of Delphi

The uncertainty in a Delphi forecast is measured by the interquartile range of the panel's responses or, in some cases, the difference between the dates forecast for very low and very high likelihoods of occurrence. In general, the precision of Delphi estimates varies with the length of the forecast. A typical result is shown in Figure 2.2, which plots the time

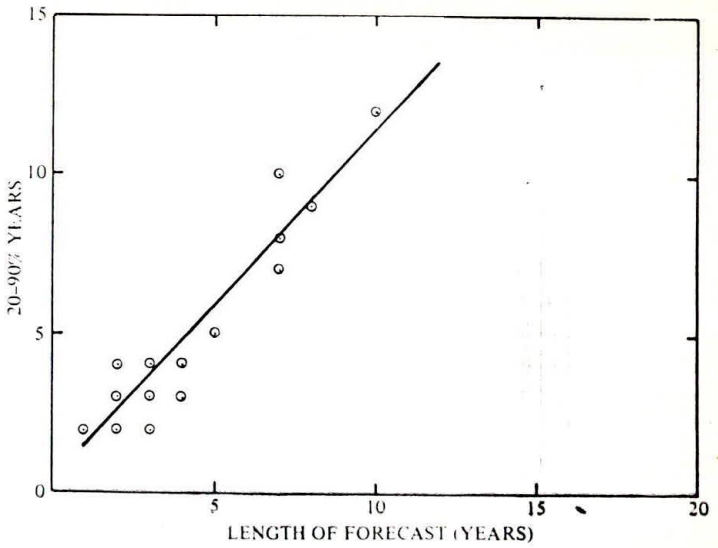


Figure 2.2. Spread of estimates versus length of forecast.

between the dates for the 20 and 90% likelihoods of occurrence against the length of time from the present to the 50% likelihood of occurrence for several forecasts by the same panel. The farther away the event, the greater the uncertainty of the panel, and the less the implied precision of the forecast. The linear growth of uncertainty is one more indication of the "lawful" behavior of Delphi estimates.

9. The Reliability of Delphi

How likely is it that two equally expert Delphi panels will give significantly different forecasts for the same event? Since experts do not always agree, this is a possibility; but if it happened often, Delphi would be a useless method of forecasting.

Dalkey investigated this in his work with almanac-type questions. He took first-round responses and treated them as a population from which he drew samples of various sizes. For each sample he obtained the median and for each sample size he obtained the correlation between the median and the true answer. The results are shown in Figure 2.3, which shows the mean correlation coefficients, over all questions, for several sample sizes. The mean correlation between the median and the true answer increases with increasing sample size. For panels of as few as 11 members the correlation exceeds 0.7. These results indicate that a panel of 15 members, if truly representative of the "expert community" on some

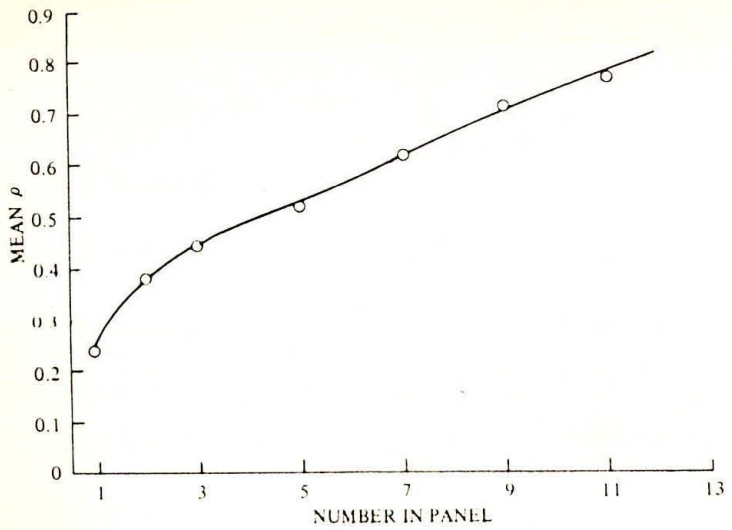


Figure 2.3. Panel reliability versus panel size.

topic, is unlikely to produce forecasts that differ markedly from those of another equally expert panel of the same size.

10. Selecting Delphi Panel Members

Some people today challenge the notion that expert opinion is needed when objective data is lacking. This has been debated in our society for years. One side insists that society is so complex that it can be understood only by the "experts," who should be given control of it; the other side denigrates expertise, claiming that the experts do not know any more about society than the rest of us. Following this latter argument, many people reject Delphi as one more attempt to put the experts in charge.

Both sides in the debate are making the same mistake: They think that there is some small group of people who are experts while everyone else is a nonexpert. This is completely wrong. An expert is someone who has special knowledge about a specific subject. Each one of us is an expert on *something*, and all of us are nonexperts on most things. There is no subset of society that can be called experts in contrast with everyone else.

This point is particularly relevant in selecting experts for Delphi panels. The panelists should be experts in the sense that they know more about the topic to be forecast than do most people. On all other topics, however, the panelists may know even less than most people. A panel member

should be selected for expertise *with regard to the topic to be forecast.* Expertise in other areas is irrelevant and is certainly not implied by selection for the panel.

How can the forecaster identify an expert? Here we will focus on identifying experts in technology. Delphis run for other purposes may require other methods for identifying experts in their subject matter.

There are two aspects to selecting experts for a Delphi panel: First, how does one identify an expert? Second, of the experts identified, which should be selected for the panel? A related issue is whether to select experts from inside the organization or from outside.

The question of whether to use inside or outside experts depends primarily on the type of forecast needed and, in some cases, the uses to be made of the results. If the preparation of the forecast requires intimate knowledge of the organization, its history, policies, and so on, then there is little alternative to the use of experts from within the organization. If, however, the forecast does not depend on knowledge of the organization but more on familiarity with some area of technology, then it is probably better to obtain the best people available, and, in general, these will come from outside the organization. Except for organizations like large universities, in general no organization can afford to have on its own staff more than one or two people of the caliber desired for this type of Delphi panel.

If the forecast is intended to be used in some manner that requires that it remain secret to be effective, then again there is little choice but to use experts from within the organization. The federal government, when obtaining a forecast in an area touching on national security, probably would have little difficulty in maintaining the desired degree of secrecy, even if outsiders were employed to help prepare the forecast. A business firm, however, that hopes to gain an advantage over its competitors through the effective use of a forecast, probably cannot count too heavily on maintaining a proprietary status for the forecast if really high-caliber outsiders are to serve on the Delphi panel. Some of the desired people may not be willing to serve if the results are to be maintained in a proprietary status. In such cases the firm is most likely better off using its own people; the employees of the firm may well make up for their lack of expertise, as compared with the best experts available anywhere, with their knowledge of the firm's interests, strengths, and weaknesses.

If the decision is made to use experts from within the organization, the identification of such experts is very much simplified. This is especially true if part of the required expertise is knowledge of the organization. The panel director will look for people in responsible technical or managerial positions who have been with the organization long enough to have acquired the desired knowledge of its special or unusual features. Evaluation of the level of technical expertise can usually be obtained

from supervisors, records of merit promotions and pay increases, and so on. In some cases the organization chart will be a sufficient guide.

Once the experts within the organization have been identified, there remains the problem of selecting among them. The biggest problem in this regard is that experts are busy people. This will be more true the higher they are placed within the management structure. This means that they may not have time to give the Delphi questionnaires adequate attention. In practice, a tradeoff must usually be made between getting panelists whose organizational position gives them a sufficiently broad view, and getting panelists who will be able to spend adequate time filling out the questionnaires. There is always a temptation for the panelist to make an estimate coincide with the panel median, simply to avoid the problem of justifying a different viewpoint. If the panelist is a busy executive, trying to fill out the questionnaire in his or her spare time, the temptation may be overwhelming, despite a sincere desire to provide a responsive and useful answer. The hasty opinion of a vice-president is probably not worth as much as the considered opinion of someone two or three levels lower in the organization.

If the decision is made to use outside experts, then the problem of identification is much more difficult. Peer judgment is usually the best criterion for identifying an expert. If the organization has on its own staff one or more specialists in the desired field, they can be asked to nominate outside experts; the outside experts can themselves be asked to nominate others. A good rule of thumb is to select those who have been nominated by at least two other people. In addition to these nominations, there are other selection criteria that have at least the appearance of being objective and which are in any case useful aids to judgment: honors by professional societies, the number of papers published, the number and importance of patents held, citation rates of published papers, and other signs of professional eminence such as holding office in a professional society.

With outside experts, assuring that they will have adequate time to answer questionnaires is not a serious problem. Outside experts are usually chosen from among university faculty members, private consultants, and others who have a significant degree of control over their own time. Their agreement to serve on a Delphi panel can be construed as a commitment to devote adequate time to preparing the forecast. The most serious problem is finding a panel who will not only agree to serve but also be available for the full sequence of questionnaires. University faculty members, for instance, tend to do a great deal of traveling during the summer; thus if the panel is to be staffed mainly with university faculty, the sequence should be timed so that it can be completed during the academic year.

Given that a set of experts has been identified, which of them should be asked to serve on the panel? Or, viewing it from a more practical standpoint, which should be asked first in the hope that they will agree to serve and that it will not be necessary to contact others? How can the panel moderator establish a hierarchy among the potential panelists? Degree of expertness, as determined during the initial search, is probably the most important single consideration. The forecast should represent the best opinion available; hence the panel should be composed of the most knowledgeable experts available. After that, considerations such as likely availability and probable willingness to serve can be taken into account.

There is another factor that must be given consideration during the selection of the panel. As pointed out earlier, one of the difficulties with any forecast prepared by a group is the problem of common or cultural bias. If the members of the panel share some set of biases, these will almost inevitably show up in the forecast. The panelists themselves are unlikely to be aware of them. There is no absolute guarantee that this problem can be eliminated; it can only be minimized by selecting representatives of every major school of thought in the subject area. If there are people within the organization who are sufficiently familiar with the field, they may be asked to identify the major schools of thought and to indicate which experts belong to which schools. The panel moderator can also make use of the various *Who's Who* publications, rosters of professional societies, and so on, to determine the background of each expert. Facts such as previous employers, schools attended, identity of thesis advisor(s), and so on can be used to help assure that a panel is not inadvertently chosen which has a one-sided outlook. If this kind of information is not readily available, then the panel should be chosen to include members with widely varying ages and representing a variety of institutions with as wide a geographical spread as possible.

It cannot be emphasized too strongly that choosing the panel is the most important decision the panel moderator will make, and considerable effort in making a good selection is fully justified.

11. Guidelines for Conducting a Delphi Sequence

It is a mistake to say, "We don't have time to get a forecast by any other methods, so let's do a Delphi." Probably more people have had bad experiences with Delphi for this reason than for any other. Delphi cannot be done "on the cheap"; Delphi takes as much time, effort, and expense as does preparing an equivalent forecast by other means.

Even though Delphi is neither cheap nor easy, it can be done with reasonable cost and effort if the more common mistakes are avoided. The following guidelines can help the user avoid such mistakes. They should not be taken as an indication that Delphi is either cheap or easy.

Obtain an Agreement to Serve on the Panel. If questionnaires are simply sent out to a list of names, without making sure that these people are willing to serve on the panel, the moderator runs the risk of not getting enough answers to be meaningful, especially if the list of names is a very short one. A few attempts to run Delphi sequences have begun by sending the first questionnaire to 200 or 300 names. Response rates typically run to 50% or less, and six to eight weeks are sometimes required to get even that many responses. In addition to the delay involved, there is no assurance that the same people will respond to every round. The moderator may well be putting in a lot of effort and not gaining any of the advantages of Delphi, in fact simply running a poll by mail, and of a very poorly selected group at that. As emphasized in the section on panel selection, choosing the panelists is the most important decision the moderator makes during the course of a Delphi sequence. The moderator must not only select the right people, but also make certain that they will in fact serve. The panel selected should also be slightly larger than the moderator thinks will be necessary (panelists have been known to die during the course of a Delphi sequence). In addition, if the panel includes the best people available, the moderator must expect that from time to time some of the panelists will have to miss a round because of higher-priority demands on their time. If the original panel is just big enough, any losses such as these may seriously reduce the effectiveness of the resulting forecast.

Explain the Delphi Procedure Completely. Delphi is not yet so well known that the moderator can be confident that the experts selected are familiar with or have even heard of it. Even if they are aware of it, they may have only a distorted picture of what is involved and what will be expected of them. It is especially important that they understand the iterative nature of the sequence. Several Delphi sequences have run into problems because some of the panelists did not understand the purpose of the successive questionnaires.

Make the Questionnaire Easy. The format of the questionnaire should be designed to help, not hinder, the panelist, who should be spending his or her time thinking about the forecast, not wrestling with a complicated or confusing questionnaire. A good way of doing this is to make use of "check the block" of "fill in the blank" questions. This is not always possible, especially in the case of events surrounded by considerable debate as to whether they will occur at all. However, it should be done whenever possible. In addition, the arguments for and against each event should be summarized and presented in a compact form that makes it easy for the panelists to follow the arguments and connect them with the question. Finally, there should be ample space on the questionnaire for the panelists to write in their own comments and arguments. In short,

the questionnaire should be designed for the convenience of the panelists and not that of the moderator. Efforts in making the questionnaire easier to answer will directly improve the quality of the responses.

The Number of Questions. There is a practical upper limit to the number of questions to which a panelist can give adequate consideration. This number will vary with the type of question. If each question is fairly simple, requiring only a single number in response to a simple event statement, the limit will be higher. If, on the other hand, each question requires considerable thought, with the weighing of conflicting arguments and the balancing of opposing trends, the limit will be lower. As a rule of thumb, 25 questions should be considered a practical upper limit. In special circumstances the number of questions may be higher; however, if the number of questions rises to 50, the moderator should examine them carefully to be sure they are focusing on the points of real interest and not diluting the efforts of the panel on minor matters.

Contradictory Forecasts. When the set of questions is generated by the panelists during the first round, it is entirely possible that contradictory forecasts will appear. These might be, for instance, pairs of events that are both possible but mutually exclusive. In principle, there is no reason why both such events should not be included in the questionnaire, especially if the outcome is of considerable interest to the moderator. However, it should be made clear to the panelists that both events are included because of the responses to the first round. The panelists should not be left with the feeling that the moderator is including contradictory events for the purpose of trapping them in an inconsistency.

Injection of the Moderator's Opinions. From time to time during a Delphi sequence it will appear to the moderator that the two sides in a debate on some event are not effectively meeting each other's arguments or that there is some obvious (to the moderator) argument or fact which both sides are overlooking. Under these circumstances the moderator may be tempted to include his or her own opinions in the feedback on the next round. This temptation must be resisted without fail. Under no circumstances should a moderator inject personal opinions into the feedback. This advice may seem harsh, but there is no alternative. Once the moderator has violated this rule, there is no recognizable place to draw the line. If a little bit of meddling is permissible, why not a little more? And this can continue until the entire forecast is distorted to conform to the views of the moderator. If the moderator's own opinions are injected into the feedback, there is a risk of converting the Delphi sequence into an elaborate and expensive means of fooling the moderator (or worse yet, fooling the clients, who may be impressed by the names of the panelists).

The moderator has gone through considerable trouble picking a panel of experts, people who presumably know a lot more about the subject than anyone else. Their deliberations should not be meddled with. If a moderator becomes convinced that the panelists are overlooking some significant elements of the problem, it should be recognized that somehow the panel selected is unqualified, and the only solution is to discard the forecast produced and repeat the work with another panel. This advice is particularly important, since considerable research (e.g., see Bradley, 1978) has shown that Delphi results can be manipulated by the injection of false or distorted information into the genuine feedback of the participants.

Payment to Panelists. Originally most Delphi forecasts were prepared by unpaid panelists; it was considered almost an honor to be asked to participate. However, those days are over. The moderator of a Delphi panel is asking for time and expert advice from the panelists and should be prepared to pay for these valuable commodities at market rates. The forecast is presumed to be valuable to the organization asking for it; a bad forecast may cost much more than the cost of preparing it. Thus the panelists should be paid at customary consulting rates.

Professional societies and charitable institutions may still be able to obtain unpaid Delphi panelists. Experts may be as willing to lend their time and knowledge to these organizations as they are to donate money or other kinds of effort. Nevertheless, moderators for such Delphis should remember that they are asking for something valuable and are depending on the good will of the panelists, which should not be abused.

Workload Inolved in a Delphi Sequence. During the Delphi the main task of the moderator is to receive and analyze responses from the panelists and prepare the questions for the next round. Experience shows that this will require about two person-hours per panelist per round. The clerical workload in preparing the questionnaire is about the same, but the timing is different.

For large panels, computerizing the analysis is almost essential. Even for panels of 50 the manual-processing workload is so heavy that there is no time for adequate analysis, and the turnaround time becomes excessive. Even for small panels, computerizing the computation of medians and quartiles is often worth the effort.

Turnaround Time Between Questionnaires. Delphis run using the mail usually take about a month between successive questionnaires. When Delphis are carried out within organizations located in a small area (plant, laboratory, university campus, etc.), turnaround can be much shorter. For panels of 10 to 15 members, using interoffice mail or couriers, two weeks has often been sufficient to carry out four full rounds. However, the panelists must be motivated to respond promptly, or the advantages of internal communication can be lost.

12. Constructing Delphi Event Statements

A Delphi questionnaire is neither a public opinion poll nor a psychological test. Many critics have failed to understand this and have complained that Delphis do not follow the rules developed for questionnaires in these fields; of course, there is no reason why Delphis should. However, there are some rules that must be followed if Delphi questionnaires are to obtain the information the moderator wants. Some of the most important rules are the following.

Avoid Compound Events. If the event statement contains one part with which a panelist agrees and another part with which he or she disagrees, there can be no meaningful response. Consider the following event: A commercial nuclear fusion plant for generating electricity using deuterium from sea water will begin operation in the year _____. The panelist who thinks that nuclear fusion will be based on the use of tritium cannot respond to this event: If he or she believes that fusion power will be available commercially at a certain date and gives this date, the response may be interpreted as supporting the use of deuterium from sea water. If he or she responds "never," it may be interpreted as doubting that fusion power will ever become commercial. In general, it is best to avoid event statements of the form, "Capability A will be achieved by method B in the year _____."

The moderator can never be certain to have eliminated all compound events. Despite one's best efforts, some panelists may find two distinct parts to what was intended to be a single event. In such a case the feedback between rounds can help the moderator improve the question. Clarifying an event statement on the basis of feedback may be as important as the forecast itself if it uncovers alternatives that were not apparent at first.

Avoid the Ambiguous Statement of Events. Ambiguity can arise from the use of technical jargon or from terms that "everyone knows." Most ambiguity comes from the use of terms that are not well defined. Consider the following event: By the year _____, remote-access computer terminals will be common in private homes. How common is "common?" Ten percent of all homes? Fifty percent? Ninety percent? If 70% of all homes with incomes over $20,000 have terminals, but only 10% of homes where the income is less than that figure do, is this "common"? Descriptive terms such as *common, widely used, normal, in general use, will*

become a reality, a significant segment of, and so on are ambiguous and should not be used.

Ambiguity can often be eliminated by using quantitative statements of events. However, consider the statement, "By 19____, the per capita electric power consumption in Africa will be 25% of the U.S. per capita power consumption." Does this mean 25% of today's U.S. consumption or 25% of the U.S. consumption in the same year? Even though the statement is quantitative, it is not clear. Consider the statement, "By 19____, a majority of all foods sold in supermarkets will be radiation sterilized and will not require refrigeration." Does this mean over 50% of each kind of food? Or does it mean over 50% of the total, but some foods not at all? If the latter, is it 50% by weight, volume, or dollar sales?

Avoid Too Little or Too Much Information in Event Statements. Some research by Salancik et al. (1971) shows that it is just as bad for a statement to have too much information as too little.

The researchers related the degree of consensus in the forecast to the complexity of the statement. They measured complexity by the number of words, which is a crude but objective measure of complexity. They also used a measure of consensus more sophisticated than the interquartile range: They borrowed a concept from Information Theory, where the information content of a message is measured in "bits." One bit is the information contained in a single "yes" or "no" when both are equally likely; that is, one bit of information is just enough to answer a binary question. So the degree of consensus was measured in bits by comparing the actual distribution of forecasts with a condition of complete uncertainty, that is, a uniform distribution of forecasts over the entire possible range of years to the time horizon.

Complete consensus would have provided 2.58 bits; however, the actual degree of consensus provided only an average of 0.6 bits per event, about one quarter of the maximum possible. On the average, the greatest consensus was achieved for event statements about 25 words long, which provided about 0.85 bits. The degree of consensus declined for either longer or shorter statements. The number of bits provided was only about 0.45 for both 10- and 35-word statements.

Some of the event statements dealt with technology, while others dealt with applications. For the technology events the shorter the event statement, the higher the degree of consensus. For application statements the reverse was true. Salancik et al. tested the possibility that the panelists were more familiar with the technology than with applications, which would mean that fewer words were needed to adequately define technology statements. They divided the applications statements into three categories on the basis of degree of use, from common to unusual. For the more common applications the most consensus was reached for the shortest statements. For the unusual applications the most consensus was reached for the longest statements.

The panelists were asked to rate their expertise on each question. For the nonexperts the longer a statement, the greater the consensus reached. For the experts the most consensus was reached at intermediate-length statements, with consensus declining for both longer and shorter statements.

The conclusion is that if an event is unfamiliar to the panelists, the more description given, the greater the degree of understanding and the greater the degree of consensus. If an event is familiar, the more description given, the more confusing the statement appears and the less the degree of consensus. Event statements should therefore be chosen to provide neither too much nor too little information. If the panel has trouble reaching consensus, the problem may be an event statement that provides either too much or too little information. The moderator should attempt to clarify this and reword the statement.

13. Summary

In the years since Gordon and Helmer brought the Delphi procedure to public notice, hundreds of Delphi sequences have been run by a variety of organizations and groups, for a variety of purposes. Descriptions of many of these sequences have been published in report form, as well as in the form of articles in journals devoted to management, planning, and forecasting. On the basis of these studies some conclusions can be drawn that should be of value to those considering the use of Delphi.

Delphi does permit an effective interaction between members of the panel, even though this interaction is highly filtered by the summarization of arguments made by the moderator. Several experiments in which the panelists were asked to give reasons as to why they changed their estimates showed that the panelists were, in fact, reacting to the views of their fellow "experts." However, this cannot be viewed as weakness of will. (In one such experiment, on the contrary, one of the panelists claimed that it made him even more "stubborn" to know that "only I had the right answer.") Panelists do shift their estimates when the arguments of their fellow panelists are convincing; otherwise they will hold tenaciously on to their differing opinions.

At the same time, however, there is ample evidence from a number of experiments that if the panelists feel that the questionnaire is an imposition on them, or if they feel rushed and do not have time to give adequate thought to the questions, they will agree with the majority simply to avoid having to explain their differences. In this respect, therefore, the Delphi procedure is not an absolute guarantee against the degrading influences of the "bandwagon effect" and fatigue. However, in a Delphi these problems are to some extent under the control of the moderator, whereas they

are virtually uncontrollable in a face-to-face committee or problem-solving group.

The Delphi procedure is thus a feasible and effective method of obtaining the benefits of group participation in the preparation of a forecast while at the same time minimizing or eliminating most of the problems of committee action. It can take longer to complete than a face-to-face committee, especially if the deliberations are carried out by mail. Since it is unlikely that a long-range forecast would be prepared in a hurry, this delay need not be a disadvantage. Even if a forecast must be obtained by a certain deadline, sufficiently advanced planning can usually make the use of Delphi possible. Thus, whenever adequate time is available, Delphi should be considered as a practical approach to obtaining the required forecast.

References

Linstone, Harold A., and Murray Turoff (1975). *The Delphi Method* (Reading, MA: Addison-Wesley).

Nelson, Bradley W. (1978). "Statistical Manipulation of Delphi Statements: Its Success and Effects on Convergence and Stability," *Technological Forecasting and Social Change* 12 (1), 41–60.

Salancik, J. R. (1973). "Assimilation of Aggregated Inputs into Delphi Forecasts: A Regression Analysis," *Technological Forecasting and Social Change* 5, 243–247.

Salancik, J. R., William Wenger, and Ellen Helfer (1971). "The Construction of Delphi Event Statements," *Technological Forecasting and Social Change* 3, 65–73.

For Further Reference

Dajani, Jarir S., and Michael Z. Sincoff (1979). "Stability and Agreement Criteria for the Termination of Delphi Studies," *Technological Forecasting and Social Change* 13 (1), 83–90.

Kendall, John W. (1978). "Variations of Delphi," *Technological Forecasting and Social Change* 11 (1), 75–86.

Linstone, Harold A., and Murray Turoff (1975). *The Delphi Method: Techniques and Applications* (Reading, MA: Addison-Wesley).

Technological Forecasting and Social Change 7 (2) (1975). Entire issue devoted to Delphi.

Problems

1. Which of the following items are likely to require expert judgment to determine, and which would be better obtained by some objective means?
   a. The likelihood of the Supreme Court upholding a patent on a technological innovation.
   b. The profitability of a new device, as compared with the device it will replace.
   c. The likelihood of new technology being rejected on moral or ethical grounds.
   d. The willingness of the public to accept a specific alternative to the automobile for personal transportation.
   e. The Federal Government's probable response to a new technological advance.

2. Assume that laboratory feasibility of a radically new technological device has just been demonstrated. This device is based on new principles and is largely the work of one man, who therefore knows much more about it than anyone else in the world. You want to obtain a forecast of its future level of functional capability and degree of acceptance by potential users. Is it worth supplementing the judgment of the inventor by organizing a committee, with him as member, to prepare the forecast? If your answer is yes, what characteristics would you look for in selecting other members of the committee?

3. If a forecast is being prepared by a committee, would you insist that the committee forecast only those things on which a majority of the members agree? Would you even insist on unanimous agreement? Is insistence on agreement likely to produce a better forecast?

4. What are the relative advantages and disadvantages of asking the panel to suggest events whose times of occurrence are to be forecast in subsequent rounds?

5. When would it be desirable to provide a panel with economic, political, and so on, context for the forecast it is to produce?

6. Your company wants a forecast of technological advances that may supplant its current products or provide it with new ones. It is decided to obtain the forecast using a Delphi panel. What are the relative advantages and disadvantages of the following panel types?
   a. A panel of experts from outside the company.
   b. A panel of experts from within the company.
   c. A panel combining both company experts and outside experts.
   d. Two panels, one of company experts and one of outside experts.

7. You are an official of a charitable organization that has in the past supplied funds for a great deal of medical research on a particular class of diseases. Cures or satisfactory preventives for these diseases are expected to be available within the next few years. You need to determine the avenues of medical research toward which your organization should shift its support. What kind of a panel (or panels) would you select to provide forecasts useful in this situation?

8. Your company manufactures a type of device that, traditionally, has been

Problems *(continued)*

bought and installed by the Federal Government for widespread public use. Technological progress in the field has been rapid, with successive devices being rendered obsolete by improvements within a few years. As a guide to your company's long-range planning, you wish to obtain a forecast of likely technological progress in the field over the next 20 years. What type of members would you include on a Delphi panel?

9. Correct the following questions so that they will not cause confusion if used in a Delphi questionnaire.
   a. Computer-controlled education for self-teaching will be available in the home by the year _____.
   b. The teaching-machine market will be a significant part of the total market for educational materials and equipment by the year _____.
   c. Power from nuclear fusion will be a reality by the year _____.
   d. Electric automobiles will be in common use as "second cars" by the year _____.
   e. A majority of office clerical operations now handled manually will be done by computer by the year _____.

Chapter 3

Forecasting by Analogy

1. Introduction

New technological projects are often compared to older projects in terms such as, "This is as big as the Manhattan Project was for its time." The idea is to convey the relative difficulty of the project with respect to the conditions of the time. This is an analogy.

The use of analogies in forecasting simply builds on this notion. It involves a systematic comparison of the technology to be forecast with some earlier technology that is believed to have been similar in all or most important respects.

But what does it mean to be "similar"? And which respects are "important"? Answering these questions is the whole point behind the idea of systematic comparison. This chapter presents a method for identifying those respects that are important and estimating their degree of similarity.

2. Problems of Analogies

The use of analogies is subject to several problems. These must be understood before a suitable method can be devised to overcome them.

The first problem is the lack of inherent necessity in the outcome of historical situations. A forecaster may discover a "model" historical situation, which is then compared with the situation to be forecast. If the two are sufficiently similar, the forecast would be that the current situation will turn out as the model situation did. However, the current situation will not necessarily follow the pattern of the model situation; only in Greek tragedy is the outcome inevitable. Moreover, a study of historical