IV.E. The Performance of Forecasting Groups in Computer Dialogue and Face-to-face Discussion

KLAUS BROCKHOFF*

1 The Problem

Advances in mathematical and statistical techniques, the availability of efficient computers as well as ideas and attempts to utilize certain organizational structures in the compilation of expertise are basic elements for an intensified discussion of the problems of forecasting specific future developments and events. As J. Wild has shown, those conditional forecasts which are derived from models by certain statistical techniques are based as much on empirical knowledge as on ad-hoc extrapolations, projections, or expert opinions.[1] However, the isolation of independent variables from the surrounding conditions and the intro-personal process of information processing often do not become clearly visible in the latter. Thus there is a danger that uncontrolled, or uncontrollably, misinterpretations and false judgments may occur.

We do not want to infringe upon the controversy on the superiority of "forecasts" as compared with "projections," which is being carried out in the theory of science[2] as well as on an empirical-pragmatic[3] level. It has by no means been settled for the forecasts. This seems particularly true when the comparison is drawn on the basis of a benefit-cost relationship.[4] If only for this reason we are interested in the question whether or not the utilization of the empirical knowledge of groups of experts in the derivation of statements about future developments or events can be improved upon by organizational arrangements. Improvement is meant as an increase in the accuracy of these statements. This is one reason for the development of the Delphi method.[5]

The conditions of group performance have been investigated in thousands of studies. Only few studies have been devoted to the question whether the peculiar organizational structure of the Delphi group[6] leads to higher group performance than the face-to-face

---

* In collaboration with D. Kaerger and H. Rehder

1 Wild, "Probleme der theoretischen Deduktion von Prognosen," *Zeitschrift fur die gesamle Staatswissenschaft (ZfgS) 126 (1970), pp. 553-75.*

2 Ibid. and E. v. Knorring, "Probleme der theoretischen Deduktion von Prognosen," ZfgS 128 *(1972), pp. 145-48; J.* Wild, "Zur prinzipiellen Uberlegenheit theoretisch deduzierter Prognosen," ZfgS 128 *(1972), pp. 149-55.*

3 E.g., R. M. Copeland, R. J. Marioni, "Executives' Forecasts of Earnings per Share versus Forecasts of Naive Models," *Journal of Business 45 (1972), pp. 497-512,* and the literature quoted there.

4 Thus the suggestion in H. A. Simon, D. W. Smithburg, V. A. Thompson, Public *Administration,* New York, 1961, p. *493.*

5 O. Helmer, N. Rescher, "On the Epistemology of the Inexact Sciences," *Management Science 6 (1959), pp. 25-52.*

6 Cf. below, section 2.3.1.

discussion in a group in which the each-toall pattern of the communication system can be activated.[7] Beyond this the unwieldy, nonhomogeneous, and inaccurate definition of types of tasks by which the performance of groups is judged[8] and thus the classification of concrete formulations of the question make it very difficult to derive statements as to the particular capacity of groups in forecasting.[9]

A large proportion of the statements as to the superiority of certain forms of group organization compared with others was obtained by observing group performance in solving certain kinds of problems and by assuming that the results would apply to tasks which appeared comparable. Thus references to the ability of groups to forecast particular future events were judged on the basis of their performance in responding to almanac-type questions. Martino has demonstrated that the answers to almanac-type questions observe the same type of distribution as the answers to forecasting questions.[10] However; it has not been investigated whether the parameters of the distribution vary significantly from one type of task to the other under conditions which are comparable otherwise. Thus it appears desirable to reconsider the original assumption.

In the following report we try to investigate some of these questions experimentally.

2 Initial Hypotheses on the Performance of Forecasting Groups

2.1 Group Performance and Group Size

2.1.1 Measurement of Variables

It has been shown in various studies that the performance of a group may depend on its size.[11]

The group size is determined by the number of members of a group. This measure refers solely to formal criteria. "Thus a person who does not contribute to the activity of the group, either because of his own reticence or because of a formal system of communication which does not accept his contributions, is still considered a member of the group.

---

[7] On the restriction of the each-to-all pattern cf. M. E. Shaw, "Some Effects of Varying Amounts of Information Exclusively Possessed by a Group Member upon His Behavior in the Group," ***Journal of General Psychology 68*** (1963), pp. 71-79.

[8] For the differentiation of types of tasks cf., e.g., M. E. Shaw, ***Group Dynamics: The Psychology of Small Group Behavior,*** New York, 1971, pp. 59, 403ff.; A. P. Hare, ***Handbook of Small Group Research,*** New York, 1962, pp. 246ff; C. G. Morris,"Task Effects on Group Interaction," ***Journal of Personality and Social Psychology 4,*** (1966), pp. 545-54. For problems of definition also R. Ziegler, ***Kommunikationsstruktur and Leistung sozialer Systeme,*** Meisenheim a. Glan, 1968, pp. 96ff.

[9] The limited choice of type of tasks and the strict assumptions on experimental group problem solving are most criticized. H. Franke, ***Gruppenproblemlosen, Problemlosen in odor durch Gruppen? Problem and Entscheidung,*** Heft 7, Miinchen, 1972, pp. 1-36, here p. 26 *et seq*

[10] J. P. Martino, "The Lognormality of Delphi Estimates," ***Technological Forecasting, 1,*** (1970) pp. 355-358.

[11] Cf., e.g., J. D. Steiner, "Models for Inferring Relationships between Group Size and Potential Group Productivity," ***Behavioral Science 11*** (1966) pp. 273-83; F. Frank and L. R. Anderson, "Effects of Task and Group Size upon Group Productivity and Member Satisfaction," ***Sociometry*** 34 (1971), pp. 135-49.

Group performance can be "synthesized" by a statistical aggregation of individual performances.[12] However, such groups lack the essential characteristic of communication which exists in natural groups. In the following we will report entirely about experiments with natural groups.[13]

It may seem natural to study group performance of groups with a considerable number of members. However, in the experiments to follow we have deliberately concentrated on small groups of four to eleven people. One reason for this is that very many small and medium-sized organizations are applying Delphi. They can call in only small groups of experts. Even so they may wonder about their performance, and how to measure it. With regard to the possible objection that the results from the observation of small groups may be subject to considerable "noise," we may say that basic to most evaluations is the use of the median of individual responses. The median, however, is not sensitive to large dispersions, even if they are one-sided. On the other hand, it goes without question that it would be desirable to repeat our experiments in order to check on the reliability of the results.

Very different things can be understood by the "performance" of a group. The tripel, number of pieces of information exchanged, time needed for solving a problem and number of mistakes, can be considered a "classical" yardstick of performance. Ziegler ascribes the origin of this tripel to a paper written by Bavelas in 1950[14]. As Barnard's definition of performance-"the accomplishment of the recognized objectives of cooperative action"[15] – makes clear, however, this classical tripel is not compulsory. Indeed, it generally remains unclear whether performance refers to the goals (recognized objectives) of the members of the group, to those of the group, or to those presented to the group.

The task given to the groups is to find an answer A to a question which deviates as little as possible from the answer A' which can be verified now or in the future. Increasing performance then means that | A-A' | approaches 0. In order to make comparisons between different questions or different groups a standardization is necessary. Thus the relative deviation of an estimate from the correct answer is used:

$$\frac{|A - A'|}{A'}$$

[12] For a chronology of the publications on statistically "synthesized" performance, cf. J. Lorge, D. Fox, J. Davitz, and M. Brenner, "A Survey of Studies Contrasting the Quality of Group Performance and Individual Performance, 1920-1957," Psychological Bulletin 55 (1958), pp. 337-72, here pp. 367f.

[13] Here natural group does not mean only a natural group with an each-to-all pattern of the communication system. With this I depart from the narrower definition in my earlier publication. See K. Brockhoff, "Zur Erfassung der Ungewissheit bei der Planung von Forschungsprojekten (zugleich cin Ansatz zur Bildung optimaler Cutachtergruppen)," in H. Hax, Entscheidung bet unsicheren Erwartungen, Koln, 1970, pp. 159-88, here pp. 167f.

[14] R. Ziegler, op. cit. p. 18; see also p. 55.

[15] C. J. Barnard, The Functions of the Executive, Cambridge, Mass., 1962, p. 55.

We call this expression the "error." If the error refers to a person, we speak of an individual error. If the error refers to' group performance, we speak of a group error. If A is the median of the estimates of all members of a group, we speak of a median group error (MGE).

The "mean group error" as used by Dalkey[16] is not identical with the MGE as given here. The basic difference is that Dalkey uses the logarithm of the quotient A|A' in order to test hypotheses about the distribution of his "mean group error." The distribution is of secondary importance for our present considerations regarding performance. For this reason we do not use logarithms here.

The MGE is used here directly as a measure of performance. It is not put into relation to the expenditures made for its derivation.

Further measures of group performance which are mentioned are the ability of the group to survive in a changing environment, its satisfaction, and the habitual change of its members.[17] We do not intend to study group performance in such a broad context (although we have unsystematically collected remarks on member satisfaction).

2.1.2 The Relationship of Performance to Group Size

The study of hypotheses about a relationship between group size and group performance occupies a prominent position in small-group research. A brief survey of the diversity of empirical results was compiled by Turk.[18] A uniform result cannot truly be expected, because the individual studies were carried out under different conditions (types of tasks, performance measures, etc.). With respect to forecasting it has been hypothesized that the mean group error decreases with increasing group size.[19] It should be taken into account, however, that this statement has been formulated only for synthetic groups, i.e., a statistical aggregation of individual judgments, and with reference to the performance of the group in answering fact-finding questions.[20]

The assumption of a decreasing error with increasing group size: is based on the probability model of search.[21] From this model one deduces the possibility of compensating individual errors by calculating the mean for the group.

In natural groups, however, the rigid conditions on which alone the statements of the probability model are valid cannot always be fulfilled. Since, however, a consistent system

---

[16] Cf. N. C. Dalkey, "'The Delphi Method: An Experimental Study of Group Opinion," Rand Corp., RM 5888 PR, 1969. Also H. Albach, "Informationsgewinnung durch strukturierte Gruppenbefragung – Die Delphi-Methode" *Zeitschrift fur Betriebswirtschaft (ZfB), Suppl.* 40, Yr. 1970, pp. 11-26, here p. 20.

[17] Summarized, e.g., by M. Deutsch, "Group Behavior," in D. L. Sills (ed., International Ency*clopedia of the Social Sciences 6,* New York, 1968, pp. 265–75, here p. 274.

[18] K. Turk, "Gruppenentscheidungen. Sozialpsychologische Aspekte der Organization kollektiver Entscheidungsprozesse," ZfB 43. (1973), pp. 295-322, here p. 302.

[19] N. C. Dalkey, op. cit. pp. 9f.

[20] These were questions where the experimenters knew the answer but the subjects did not": N. C. Dalkey, op. cit., p. 10, fn.

[21] Cf., e. g., P. R. Hofstitter, Gruppendynamik. Die Kritik der Massenpsychologie, 11th ed., Reinbek, 1970, pp. 35ff., 160ff.

of other factors that influence group performance as well (as the direction of their influence cannot be given), we may formulate:

> *Hypothesis G: With increasing group* size, *the group performance increases ceteris paribus.*

It becomes clear in section 2.5 to what extent the restriction *ceteris paribus* can be repealed in our experiments.

2.2 Group Performance and Expertise

2.2.1 The Measuring of Expertise

H. A. Simon represents "expertise" as a possible basis of authority or as a form of authority.[22] By expertise we mean expert knowledge upon which professional authority can be' founded. This expert knowledge can be "proven"[23] by demonstration' or by recourse to confirmation through third parties. A "proof" by recourse to third parties can hardly confirm more than a refutable conjecture as to the expertise of a person. If influencing variables of group performance are being sought, the experiments which make these variables visible can hardly contain an analyzable test *of* expertise at the same time. It is necessary to measure expertise as an independent variable by some other means.

One could proceed by testing which persons demonstrate expert knowledge in solving fact-finding questions. When such' persons' have been found, they *can* be engaged in forecasting. This takes for granted that the answering' of both types of questions can be considered to be identical types of tasks.[24] Until now, however, no empirically tested statement to this effect exists.

Further, one could consider whether expertise ratings by third parties can be used. However, it may be of interest to maintain the anonymity of all persons who may possibly participate in a forecasting group.[25] Another problem that arises'is what criteria should be used in choosing those persons who are to judge the expertise of others.

Thus there remains the possibility of determining expertise by self-rating. For this purpose an ordinal scale is generally used, from which one value can be chosen to indicate expertise. We worked with a scale of real numbers graded from 1 to 5, in which low

---

[22] H. A. Simon, Administrative Behavior, 3rd ed., New York and London, 1965, p. 76.

[23] F. Landwehrmann, "Autoritit," in E. Grochla (ed.), Handworterbuch der Organisation, Stuttgart, 1969;" col. 269-73, here col. 270, refers to H. Hartmann, Funktionale Autoritat, Stuttgart, 1964

[24] For a procedure oriented thus; cf. M. A. Jolson, G.-L. Rossow, "The Delphi Process in Marketing Decision Making," Journal of Marketing Research, 8 (1971), pp. 443-48. Another procedure, based on the solution of test questions and a test of the understanding of :professional terminology, is described by A. J. Lipinski, H. °M. Lipinski, R. H. Randolph, "Computer-Assisted Expert Interrogation: A Report on Current' Methods Development," Technological Forecasting and Social Change 5 (1973), pp. 3-18, here pp. 9f: (The same *in* S. Winkler [ed.], Computer Communications, Impact *and* Implications, New York, 1973, pp. 147-54. The authors also test the "quality of respondents' comments" [presumably on factual questions], the degree of attention and the degree of optimism [with the aid, of a price list for old phonograph records] and inquire from this a rank order of expertise.)

[25] C.f. Section 2.3.1.

numbers must be used to express a low degree of expertise while high numbers may be used to express a high degree of expertise. Such a determination of expertise is employed already in some forecasting groups by their management.[26] These results of individual forecasts are weighted according to the self-ratings when a group judgment is derived.

Measurements of expertise which are obtained for individuals should also permit statements as to the expertise of the total group. Since self-ratings are measured on an ordinal scale it is not' permissible to form an arithmetic mean of all the self-ratings of the members of the group. We therefore characterize the expertise of a group by the median of the individual self-ratings.

Whether such self-ratings have a high positive correlation with ratings by third parties has not yet been studied in realistic situations. For this reason it remains an open question whether corresponding confirmatory results of psychological tests[27] can be applied to real situations, which generally are not free of conflicting interests.

Even if no significant positive correlation exists between ratings by third parties and self-ratings concerning expert knowledge, it is not determined which of the ratings is more correct. Thus in the approach used here, which involves self-ratings, the question remains whether the participants in the experiment rate themselves correctly when compared with an (inapplicable) objective standard.

2.2.2 The Relationship between Expertise and Group Performance

We assume that groups with high self-ratings of expertise perform better than groups whose members rate themselves as less qualified. With this assumption we follow Dalkey, Brown, and Cochran[28]. Their results must, however, be examined with care insofar as they were obtained from answering fact-finding questions. Furthermore, the subjects were able to compare all questions to one another before rating their expertise with respect to each question. In many applications it is not possible to present all questions at once. We shall therefore follow a different procedure by presenting tasks in a sequential manner. Even so, we assume that the basic relationship is still valid. Thus, we arrive at

> *Hypothesis E: With increasing expertise, group performance increases ceteris paribus.*

---

[26] D.L. P-1-; "A Practical Aporoach to Delphi: TRW's Probe II," *Futures 2* (1970), pp. 143-52; H. P. North, D. L. Pyke, "Probes of the Technological Future," *Harvard Business Review 3 (1969), pp. 68-76; A. J. Lipinski, L. M. Lipinski, R. H. Randolpk, op. cit., pp. I* 1ff.

[27] M. A. Wallach, N. Kogan, J. Bem, "Group Influence on Individual Risk Taking," *Journal of Abnormal and Social Psychology 65* (1962), pp. 75-86, here p. 83.

[28] N. C: Dalkey, B. Brown, S. Cochran, "The Use of Self-Ratings to Improve Group Estimates," *Technological Forecasting, 1* (1970) pp. 283-292.

2.3 Group Performance and Communication System

2.3.1 The Characteristics of Face-to-Face Discussion Groups vs. Delphi Groups

The question whether decentralized or centralized organizations exhibit higher performance is another of the classical questions in organization research. The attempt to set up a universally applicable rule of organization to answer this question had to be dropped because little by little conditions become known on which first the one organizational structure and then the other seemed more advantageous for accomplishing specific tasks. Fundamentally, it is assumed in these studies that all organizational structures considered are capable of accomplishing certain tasks. A formal organization is characterized essentially by its communication system and the distribution of competence among its members.[29] It is further assumed that fact-finding questions as well as forecasting questions can be answered more accurately by groups than by individuals, if the (expected) error[30] is taken as the measure of accuracy.

Fact-finding questions and forecasting questions can be discussed in natural groups with an each-to-all pattern of communication. If it is desired to have a group judgment, this task can be left up to the group, or a rule for aggregating the group judgment from the individual judgments of the participants can be given. Depending on the situation, the use of such a rule can be restricted to the case where the group does not agree on a group judgment within a given period of time.[31]

Particularly Carzo's results indicate that in natural groups in which communication is not limited, solutions to complex tasks are reached rapidly, with few errors, and to the satisfaction of the group members.[32] One objection to this is that this form of organization may also produce dysfunctional effects.

A first dysfunctionality may arise as certain group members consciously or unconsciously influence the group result to a greater degree than their expertise warrants.[33] A further dysfunctionality arises if the exchange of information is interfered with by "noise."[34] A more far-reaching possibility for the occurrence of dysfunctional effects is that the transmission of information necessary for task accomplishment from some group

---

[29] This is made particularly clear by H. Albach, "Organisation, betriebliche," in ***Handworterbuch d*** Sozialwiss. 8, Stuttgart, Tubingen, Gottingen, 1961, pp. 111-17.

[30] Cf. H. Schollhammer, "Die Delphi-Methode als betricbliches Prognose-und Planungsverfahren," ***Zeitschrift*** fur ***betriebswirtschaftliche Forschung (ZfbF),*** N. S., 22nd Yr. (1970), pp. 128-37, here particularly fn. 8.

[31] Cf. the experiments by B. Contini, "The Value of Time in Bargaining Negotiations -- Some Empirical Evidence," ***American Economic Review 58*** (1968), pp. 374-93.

[32] Cf. R. Carzo, Jr., "Some Effects of Organization Structure on Group Effectiveness," ***Administrative Science Quarterly 7*** (1963), pp. 393-424.

[33] This very abridged presentation must be viewed on the basis of the entire discussion concerning the question: which conditions promote behavioral conformity in individuals in groups; cf. A. P. Hare, *Handbook ..., op, cit.,* chapters, 2, 13 (there in reference to status rivalry); L. Festinger, E. Aronson, "The Arousal and Reduction of Dissonance in Social Contexts," in D. Cartwright, A. Zander (eds.), Group *Dynamics: Research and Theory,* Evanston, Ill., 1960, pp. 214-31.

[34] A. P. Hare, *op. cit.,* chapter 10, pp. 272ff.

members to others is blocked by somebody interested[35] or that the transmission of information from some group members in the time period given for accomplishing the task is altogether impossible. In the latter case the participation of the individual group members in the exchange of 'information may be independent of the degree of expertise. Then participants with a high degree of expertise cannot necessarily influence group performance.

One can summarize that possible dysfunctionalities in a natural group with an each-to-all pattern of communication result from utilizing the communication system and the system of competence in a manner which does not correlate positively with the degree of expertise. Assuming that these effects often interfere .with group performance, rules should be set up which reduce the effects of dysfunctionality or prevent their appearance. A set of such rules was suggested and introduced by Helmer and Dalkey[36] and , given the name "Delphi." In spite of diverse variations in procedure, the applications known to date; .have as -their primary objective: "...the establishment of a meaningful group communication structure."[37]

2.3.2 The Relationships

According to Dalkey's studies, Delphi groups demonstrate a certain, though not significant superiority when compared with certain other groups in solving fact-finding questions.[38] We apply this perception to our

> *Hypothesis D: The Performance of Delphi groups is ceteris paribus higher than the performance* of *natural groups with an each-to-all pattern of communication.*

The diverse arrangements of Delphi experiments make it necessary to investigate some of their special features as well. Of particular importance is the question whether the performance o€ Delphi groups is the same in each round, or whether it increases with increasing number of rounds, at least up to the fourth round.[39] This leads to

> *Hypothesis R': The performance of Delphi groups increases ceteris paribus with increasing number of rounds.*

We test this hypothesis here for up to five rounds. It cannot be expected that hypothesis R' is valid for an unlimited number of rounds, since growing dissatisfaction of

---

[35] On this broad field, see H. H. Kelley, J. W. Thibaut, "Group Problem Solving," in G. Lindzey, E. Aronson (eds.), *Handbook of Social Psychology* 4, Reading, Mass., 1968, pp. 1-101, here pp. 6ff., 26ff.

[36] N. C. Dalkey, O. Helmer, "An Experimental Application of the DelphiMethod to the Use of Exerts," *Management Science* 9 (1963), pp. 458-67.

[37] M. Turoff, "Delphi and its Potential Impact on Information Systems," AFIPS *Conference Proceedings,* Fall Joint Computer Conference (Fall 1971), 39, pp. 317-26, here p. 317.

[38] N. C. Dalkey, *op. cit. passim,* particularly p. 22.

[39] Cf. also J. B. Martino, "An Experiment with the Delphi Procedure for Long-Range Forecasting," IEEE Trans. on Engineering Management 15 (1965), pp. 138-44.

the participants and increasing time requirements - make it seem senseless to continue the consultations indefinitely. Therefore, we modify hypothesis R' to

> *Hypothesis R: The performance of the Delphi groups increases ceteris paribus with increasing number of rounds only at first. Finally, the increase in performance can be reduced and inverted. The question in which round the performance inversion" begins must be answered empirically.*

Finally, Helmer and Dalkey[40] showed an interest in the observation that the variance of the responses around the median decreases with increasing number of rounds. The reduction of variance is not in itself a criterion for increased performance. One must view this observation on the basis of hypothesis *R:* together with it, variance reduction gains importance in the sense that it means increasing certainty and accuracy of the answers.[41] We therefore also test this question by investigating

> *Hypothesis V: The variance of answers around the median decreases ceteris paribus with increasing number of rounds.*

Two statistical measures of variance are at our disposal for testing this hypothesis: average quartile difference and average variance from the median. The latter measure offers certain advantages for a comparison between groups of different sizes. For this reason it is given preference here.

2.4 Group Performance and Type of Task: Fact-finding Questions and Forecasting

Only a few of the generally available results of forecasts by Delphi groups can be tested against reality, because they mainly refer to events which are expected to take place in the distant future. For this reason, the comparison of performance of face-to-face discussion groups and Delphi groups is made by observing tasks which appear similar.[42] The problem of solving fact-finding or almanac-type questions is assumed to be similar to forecasting. The answers of such questions are, as a matter of principle, unknown to the participants but known to the experimenters. These two applications of Delphi, its use in forecasting and its simulation with only subjectively unknown bits of knowledge, exist as yet side-by-side without comparison. Since the complexity of a task is an important determinant of group performance, but the criteria for determining tasks of varying degrees of complexity are not clear enough, we want to test directly

---

[40] O. Helmer, Social Technology, New York and London, 1966, pp. 101ff.; N. C. Dalkey, op. cit., p. 20.

[41] They are based according to the Delphi method, on renewed intrapersonal conflict solution and problem solving, after being provided with additional data. On the interpersonal process which is to be eliminated here, cf. J. Hall and M. S. Williams, "A Comparison of Decision-making Performances in Established and Ad Hoc Groups," Journal of Personality and Social Psychology 3 (1966), pp. 214-22. On the practical organization of the elimination of dysfunctionality and pressure to conform, cf. 3.2, below.

[42] N. C. Dalkey, op. cit., pp 9f. Dalkey cites (p. 23) a paper by Campbell, in which similar methods evidently were used to the ones planned here. The original publication was not available.

Hypothesis F: The performance of a group in answering fact-finding questions is ceteris paribus equal to that in forecasting.

The hypothesis is deduced from the assumption that both types of tasks exhibit the same degree of complexity. The tests should be carried out separately for face-to-face groups and Delphi groups of the same size. If the hypothesis is refuted, many of the statements about Delphi-groups which we rederived using fact-finding questions cannot be maintained. In order to test hypothesis F, we chose only facts referring to events that had occurred, on the average, six months before the experiments. The forecasts, on the other hand, refer mainly to a period of time which did not exceed six months after the tests were carried out.

2.5 The Relationship between the Hypotheses

The repeated use of *ceteris paribus* in the hypotheses leads us to assume that they are in fact related to each other. It seems senseless to describe the great variety of possible combinations. The rejection of certain hypotheses can reduce the test program considerably as it leaves only a small number of relationships which need to be tested. Hypothesis E, e.g., can be tested for a given group size, a given type of question, a given type of group, and, in Delphi groups, with regard to the results of any round. No matter what the result is, it is irrelevant for the further experiments, if expertise should be distributed equally in the different groups. Since the distribution of the actual expertise becomes known from the results of the experiments, it is advisable to bring about the necessary clarifications at first.

The situation is similar when discussion groups and Delphi groups are compared with each other. If hypothesis R is not tested for the latter, it cannot be determined from which round the results should be taken if compared with the performance of the face-to-face discussion groups.

The presentation of our results in Section 4 is organized according to such reasoning.

3 The Experiments

3.1 Participants, Group Formation, Place, and Time

All experiments were carried out as part of a lab. course listed in the University of Kiel catalogue. It was planned to have "students" and practitioners work in separate groups and to compare the results. However, since we could not give credits for the course, too few students registered to be able to form even *one* small group.

Practitioners were designated and chosen from the permanent staffs of the local banks with the assistance of the bank managers. Bank employees were chosen because hardly any other line of business is represented in the area by enough individual organizations with personnel trained in economics and with relatively uniform fields of business. At the same time, the size of the participating organizations is generally so large that persons who perform specialized functions (long-term credits, short-term credits, investment brokerage, etc.) and who differ with respect to the lengths of their employment could be chosen. This

seemed desirable in order that definite differences could show up in the self-ratings of expertise in reference to the individual tasks.

The thirty-two participants were randomly assigned to four groups, having five, seven, nine, and eleven participants, respectively. At the face-to-face discussions, however, registered participants were absent for various personal reasons, so that the groups had four, seven, eight, and ten participants only.

The experiments began with an introductory lecture about forecasting methods and an exercise in the use of the displays which were used in the experiments with the Delphi groups. After that, each subject participated in a session in which the members were organized as a Delphi group. At a later session a face-to-face discussion took place. After the experiments were concluded, an opportunity for criticism was given. Eight months later, the results were communicated. We have tried to motivate our participants to cooperate well in the preparatory lecture and demonstration. Besides, we offered book prizes for outstanding performances with regard to different types of questions and the two basic group structures.

The experiments were conducted in May and July 1973. With three exceptions they were scheduled for Thursdays to conform with late closing time of banks. The Delphi groups worked in the computer center of the University of Kiel; the face-to-face discussions were carried out in a library room.

3.2 The Organization of the Delphi Groups and the Face-to-Face Groups

The Delphi groups were set up so that the participants received all information from the experimenters on a computer-generated display.[43]

The participants in a group were not supposed to establish immediate contact with each other. They responded to all questions by writing an alpha-numeric text in their normal language. The responses can be divided into three classes: (1) responses that were known only to the experimenters; (2) responses which, after the responses of all participants had been received by the experimenter, became objects of computing procedures, the results of which were made known to all participants; (3) responses which were recorded and made known to all participants without any changes. The first class includes the name of the participant and the degree of expertise that he expresses with regard to each question. This is handled differently in the face-to-face discussions groups. A response of the second category is an individual estimate. After the computation of the median of the responses of the group members, this figure is made known to all participants. The third category includes all arguments for divergent opinions of those whose responses lay outside of the lower or upper quartiles.

Computer communication has been praised as a means to enable experts to communicate with each other even though they are separated from each other by large distances. In the real world this could mean savings in travel expenses and in the efforts expended in coordinating dates for groups of experts. Beyond this it is of importance for the experiments that the computation of quartiles, and the preparation, distribution, collection,

[43] The programming was done by D. Kaerger, who will report separately on problems that arose herewith. The program had to be in FORTRAN IV. It was run on a PDP 10.

reproduction, and renewed distribution of questionnaires do not have to be carried out by hand during the sessions. This gives one the chance to shorten the experiments considerably.[44]

The entire exchange of information between the participants as well as between the experimenter and the participants during the sessions, with the exception of certain recurrent standard formulations, was stored on tape as a record of the experiments.

Figure 1 shows part of a record. A separate data file, an abstract from the records, is kept on tape. It serves as the data base for the diverse computations. The abstract from the record shows the beginning of a session of the Delphi group with seven participants. Vertical lines on the left edge of the text signal those portions of the texts which appear in the same form on the display of each participant. In section 1 the names of the participants are given to the experimenter. Section 2 contains the first fact-finding question for the group. In section 3 you see information which is considered relevant for the judgment of the problem and which is given to each participant. Additional information of this sort is not given to participants when forecasting questions are asked. This difference is justified by the assumption that the subjects need some information to refresh their memories with respect to judging "facts" which are about six months old. It is expected that they do not need help in evaluating present-day facts as a basis far their' forecasts.

In the following section, 4, participants are asked to give their degree of expertise. It is given only once for each question. In section 5, we enter the response portion-of the first round.

The numbering of the participants in sections 4 and 5 serves only as an internal identification. It begins with "3", because "l" and "2" are reserved for the experimenter, and the tape on which we store the record.

After the estimates have been made (in section 5) the 0.25, 0.5 and 0.75 quartiles are calculated. The participants whose estimates lie outside the 0.25 and 0.75 quartiles are shown the quartile values which they exceeded or fell short of, and are asked to give reasons for this divergence, in case they think they possess particular additional information. The text of the questions, which is repeated at this place, is not put out in the record shown here. In section 6, data necessary for the analysis are recorded.

The second round begins with section 7. First, the question is repeated, then the relevant information. In section 9, additional data are given, namely, group responses' from the previous round and any additional information which was collected in section 6 of the previous round. This information is presented in the following order: first, additional information from those participants whose estimates fell short of the lower quartile; then, information from those whose estimates exceeded the upper quartile. In the present case, there is only one item of additional information. The original text accompanying this information, which does not refer to its sources; is not reproduced here.

Beginning with section 10, the process which was` described for section 5 and the succeeding sections is repeated. Five rounds are carried out for each question. This scope was chosen in compliance with the observation that after the fourth round generally the

[44] For a brief discussion of the advantages and disadvantages of "computer communication" compared with direct communication, cf. A. I. Lipinski, H. M. Lipinski, R. H. Randolph, op. cit., particularly pp. 11-12.

results do not improve (see section 2.3.2). An additional round is added here to test this statement:

> *After the five rounds are completed the next question is asked. It is a forecasting question. The two types of questions are asked alternately so that possible effects of learning or fatigue do not influence only one type of question.*

In face-to-face discussion groups, the group members are asked to introduce themselves to each other by their name, field of employment, official position, and the number of years spent in banking. The idea of this was to provide each participant with a basis for judging the experience of the discussion partners in the following discussions. To what extent this information was taken into account in the formation of the group judgments could not be registered explicitly. Furthermore, the participants were asked to specify their degree of expertise for each question on a record. They noted their personal estimates for each question before any discussion took place. A discussion of the problem was expected to follow and a unanimous group estimate was demanded. A discussion leader was not appointed.

3.3 The Questions

The fact-finding questions and the forecasting questions refer to finance, banking, stock quotations, and foreign trade. We could choose from a list of ninety fact-finding questions and thirty forecasting questions that were kept on a separate tape. The questions were chosen at random from this stock. Each question of the two different types had the same chance of being chosen. However, no question appeared twice in a group.

All questions refer to items which are reported in the monthly statistics of the German Federal Bank (Deutsche Bundesbank), the daily stock market quotations of the Frankfurt Stock Exchange, and the market reports of the big banks. Only very few of the fact-finding questions refer to facts which are reported in the foreign trade statistics.

In all cases the correct responses can be verified objectively at the time of the experiments: or at a later date. In the opening lecture it was called to the attention of the participants that, for example, the questions about certain past or future interest rates did not refer to the rates of the respective local institutions; which may depend largely on effects of local competition. Rather, they refer to the rates which are listed as averages in the statistics of the Federal Bank.

In five **cases** the ,wording of the questions was unclear to the participants[45] This resulted partly from an inexact formulation on our part and partly from imperfect knowledge of the definitions as used by the Federal Bank on the part of the participants. In the face-to-face discussions, clarifications could be made immediately. In two cases in the Delphi groups, the "correct answer" in the sense in which it was understood by the

[45] On the significance of the wording of questions and its influence on the level of estimates, cf. J. R. Salancik, W. Wenger, and E. Helfer, "The Construction of Delphi Event Statements," Technological Forecasting and Social Change 3 No. 1"(1971), pp. 65-73.

participants was carried on rather than the answer to the original formulation of the question. Further cases of general misunderstanding did not become evident.

In a final discussion many of the participants expressed the feeling that the fact-finding questions were rather irrelevant and annoying: all the facts could be looked up with no trouble. This point of view was not expressed with regard to the forecasting questions. It should be recorded, however, that both sets of questions refer to the same objects, although at different points in time. (Thus, for example, one question asks for the price of a share of RWE common stocks six months before the experiments and another for the same quotation six months afterward).

3.4 The Volume of the Tests

It was originally planned that each group of different size and organization should be asked to give ten forecasts and to answer ten fact-finding questions. This plan was carried out with one exception. The Delphi group with eleven participants was able to handle only eight questions of each kind.

The time spent on the discussions amounts to between 140 and 200 minutes. In the Delphi rounds between 200 and 240 minutes of connect time were spent per participant.[46] (This length of time does not correspond to the CPU time, however.[47] The following points can be considered as possible explanations for the greater length of time spent on the Delphi rounds: (1) Participants write more slowly than they speak. (2) Communications between the experimenter and the participants takes place "sequentially," i.e., if participants j and k are involved, $j < k$, the message of participant k to the experimenter or back can be exchanged only after the same kind of message has been exchanged between participant j and the experimenter. This pattern of sequential communication is determined by the available technology. (3) Communication among the participants and between the participants and the experimenter can take place only during the periods of time in which the computer, which operates in a time-sharing mode, is available for the job. Although the CPU was not busy with batch operation during the experiments, the demand for memory space for other jobs which were also initiated at remote terminals was noticeable during the experiments. The participants considered such delays very disturbing. Finally let us point out that since the available teletype terminals type more slowly than the displays, preliminary experiments indicated that the former were not suitable for the experiments.

The operating system allowed for the connection of 13 displays at one time. This determined a possible maximum of group size. However, as each participant was supposed to join each type of group only once, we have limited maximum size to 11. The minimum size of groups was determined by the consideration that we wanted to have two clearly determinable participants whose estimates lay outside the quartile values. This can be

[46] With 20 questions, that is 10 to 12 minutes; with 16 questions, 12 to 15 minutes. Lipinski, Lipinski, and Randolph, op. cit., report 15 minutes per question, on comparable hardware.

[47] D. Kaerger will contribute to the further analysis. See also, Institute for the Future, "Development of a Computer-Based System to Improve Interaction among Experts," First Annual Report to the National Service Foundation, 8/1/73, p. 6, Table 2. The relationship of CPU time to connect time varies from 1:110 to 1:135.

achieved with five people. The fact that we had one man less in the discussion group did not interfere with this principle.

4 The Results[48]

4.1 The Distribution of Expertise

We want to investigate whether the expertise of the group members is distributed evenly or unevenly in the groups. It can be assumed that the distribution is even, because the participants and the questions were assigned to the groups at random.

However, the preliminary question whether expertise is rated differently with regard to fact-finding questions and forecasts must be clarified. Only if the hypothesis of an uneven distribution is rejected can a comparison between the groups be made with aggregated data from fact-finding questions and forecasts. Therefore, we first test "auxiliary hypothesis 1": *expertise* is *rated differently in each group with regard to fact finding questions and forecasts.*

Table 1
Quartiles of Expertise in Fact-Finding Questions and Forecasting Questions

Type of Group	Group Size	Quartiles					
		Lower Quartile		Median		Upper Quartile	
		Fact-finding questions	Fore-casting questions	Fact-finding questions	Fore-casting questions	Fact-finding questions	Fore-casting questions
Face-to-Face Dis-cussion Group	7	1	1	1	2	2	3
	4	2	2	2	2	3	3
	8	1	1	2	2	2	3
	10	1	1	2	1	2	2
Delphi Group	5	1	1	2	2	3	3
	7	2	2	2	3	3	3
	9	1	1	2	2	3	3
	11	1	1	2	2	3	3

The comparison of the quartiles given in Table 1 reveals different results in only five out of twenty-four cases. However, the narrow limits of a scale from 1 to

[48] The tests quoted in the following are described, e.g., by S. Siegel, *Non-Parametric Statistics for the Behavioral Sciences,* New York, and London, 1956; G. A. Lienert, *Verteilungsfreie Methoden in der Biostatistik,* Meisenheim a. Glan, 1962.

5 does not allow this result to appear sufficiently reliable.[49] We therefore compare the differences between the cumulative relative frequencies of the expertise ratings within each group for fact-finding questions and forecasts. The Kolmogorov-Smirnov Test shows no significant difference of the distributions on the 5 percent level.

Thus, auxiliary hypothesis 1 is rejected. We can use the entire set of data to investigate auxiliary hypothesis 2. It says that *between the groups no differences occur in the relative frequencies with which the different scale values of expertise are chosen.* Table 2 shows the relative frequencies of the distributions of expertise.

Table 2
Distribution of Expertise (%)

Type of Group	Group Size	Degree of Expertise				
		1	2	3	4	5
Face-to-Face Discussion Groups	4	44	32	13	5	6
	7	21	38	34	6	1
	8	30	44	19	4	3
	10	45	33	15	6	1
Delphi Groups	5	39	32	16	9	4
	7	15	35	37	9	4
	9	29	28	25	13	5
	11	39	30	18	9	4

Within each type of group comparable results appear, as expected.[50] The distribution of expertise is in both cases indistinguishable between the smallest and the largest groups; the distribution varies greatly between the group with seven participants and the largest and smallest groups, respectively (cf. Table 3).

[49] This conjecture is based on general reflections and empirical results on the correct construction of a scale. On the inferiority of the scale with five divisions compared with scales with more graduations in estimates of decisionmaking groups, see G. Huber and A. Delbecq, "Guidelines for Combining, the judgments of Individual Group Members in Decision Conferences," manuscript, TIMS-meeting, Detroit, 1971, pp. 5ff. It could not be investigated whether these statements can be applied to our results, because of the difference in the task and because American subjects may be less familiar with a scale using five divisions than are Germans.

[50] A comparison with the groups with the same rank of size within the type of group shows no significant differences at $p > 5$ percent, neither with the Kolmogorov-Smirnov Test nor with the sign test. The latter test was carried out in compliance with the possible objection that the samples were interrelated.

Table 3
Maximum Absolute Difference between Cumulative Distributions of the Relative Frequencies of the Degree of Expertise between Pairs of Groups (in %)

Face-to-Face Discussion Groups		Delphi Groups	
Group Size	7 8 10	Group Size	7 9 11
4	23 14 5	5	24 14 2
7	15 24	7	14 24
8	15	9	12

An investigation of the observations for all groups of a given type leads us to reject auxiliary hypothesis 2. If one were to formulate auxiliary hypothesis 2 for a pairwise comparison between groups it would be refuted in all cases except when the smallest group is compared with the largest group or the second largest group respectively (Kolmogorov-Smirnov Test on the 5 percent level). A consideration of the data in Table 2 now gains increased importance. The significant differences are due to the fact that in the middle-sized groups, and particularly in. the groups with seven participants, the ratings of expertise seem higher than in the smallest or largest groups. It is obvious, and is not examined more closely here, that the lower degrees of expertise are chosen much more frequently than the higher degrees. Whether this is an illustration of the often assumed pyramid of qualifications and abilities, or whether it only reflects a fear of using the higher values on the scale, cannot be determined definitively. The second assumption is supported by the observation that in the Delphi groups, where greater anonymity is guaranteed, 9.6 percent of the self-ratings fall into the categories four and five, whereas in the face-to-face discussion groups, where the self-ratings were occasionally asked for directly by other members of a group, only 5.5 percent fall into these categories. The results of the next section contribute to the extension of these reflections.

4.2 The Significance of the Self-Ratings

A first indication as to the validity of hypothesis E can be gained by observing individual errors and self-ratings. We test whether the lowest individual errors in each group and in relation to each question are attained by those persons who rate their expertise highest. Individual errors (see section 2.1.1) are taken as absolute values. In the Delphi groups we can determine this error in every round. We restrict ourselves here to the first and the last round. In the face-to-face discussion groups the only data for judging individual errors is from the questionnaires filled out before entering the dis cussion of each question. We test auxiliary hypothesis 3:

The distributions of the highest self-ratings with regard to each question and the self-ratings of those who attain the highest level of performance (i.e., the lowest individual error taken as an absolute value) coincide.

Auxiliary hypothesis 3 is tested separately for fact-finding questions and forecasting questions in face-to-face discussion groups and Delphi groups. In the latter case it is also tested separately for the first and the last rounds. In all cases, auxiliary hypothesis 3 is refuted at a high level of significance in a X^2 test (0.01). Thus, one must assume that in the situations investigated here, selfratings with regard to expertise do not give enough information as to which persons actually possess expertise. For this reason either the ability to give a self-rating which corresponds to actual expertise must be studied more closely and, if possible, promoted or other methods of determining expertise before the beginning of the questioning must be tested for their effectiveness. The Institute for-the Future emphasizes the latter problem[51] in its recent studies.

Auxiliary hypothesis 3 was based on individual performance, Hypothesis E, on the contrary, refers to the performance of the entire group. We attempt to operationalize this viewpoint by testing the rank correlation[52] between group performance with respect to each question and the average expertise of the group with respect to the same question, separately for each type of group, each group size, and each of the two types of questions. The skew distribution of the degrees of expertise already lets us expect that important information cannot be gained from such a test. Indeed, we do not find any significant rank correlation coefficient in the relationships tested for fact-finding questions. A classification of the data in a 2 X 2 contingency table according to the criteria: low and high degree of expertise vs. upper and lower half of the scale of the rank figures for group performance, does not lead to significant relationships. With regard to forecasts, significant relationships (at the 5 percent level) show up in Delphi groups with five and nine participants in the third as well as in the fifth round. However, as is shown in Table 6, these groups are not noted for particularly good overall results. In this respect the correlation seems unimportant.

Since the conclusion that objectively existent expertise is not an essential factor of individual or group performance contradicts the definition of expertise and thus is not tolerable as an explanation of the results, it can only be concluded that the self-rating of expertise by practitioners for tasks of the present type does not coincide with their objective expertise.

After the experiments were concluded, this unsatisfactory result led to the question whether explanations can be found for the choice of the rank figures of expertise by the individuals. We attempt to explain the behavior of our subjects by the following auxiliary hypothesis 4:

The degree of expertise with regard to a question is related to the number of years that a subject spent in banking. Furthermore, it is higher whenever the

---

[51] Cf. Institute for the Future,"Development of a Computer-Based System to Improve Interaction among Experts," op. cit.

[52] Spearman rank correlation with correction for ties.

subject matter of *the question coincides with one of the fields handled during the years in the profession.*

The necessary data were collected by questionnaire. Answers from up to twenty-eight participants were available. In the analyses a high positive correlation showed up between age and number of years in the profession, so that separate hypotheses for these two variables were not tested. Further, a positive correlation showed up between the number of fields one had experience in (a list of possible fields was presented which could, however, be supplemented) and the number of years in the profession. Thus the two components of hypothesis 4 can no longer serve as mutually independent variables for explaining the degree of expertise. Therefore, we tested the hypothesis in a simplified form, once for the influence of the number of years in the profession and once for the influence of the fields of employment on the choice of the degree of expertise. In neither case, a Kolmogorov-Smirnov Test refutes the hypothesis (0.001 level):

Thus an observation that was gained in the face-to-face discussion groups is confirmed. Once in a while during the discussions the number of years of "banking experience" was brought into play to decide points of controversy, obviously with the view that this is a criterion for measuring expertise objectively. The same is true for the present field of employment of the persons in question. Obviously, however, these criteria for judging on the expertise are not sufficient in the light of very special questions. It would probably be better to consider, for example, the regular observation of special sections of the bank statistics.

4.3 The Performance of the Delphi Groups

We formulated two hypotheses regarding the performance of the Delphi groups. We first test hypothesis V. To do so, we determine how frequently the measure of variance for the last round is smaller than that for the first round (cf. Table 4).

Hypothesis V cannot be refuted. When up to five rounds are carried out in Delphi groups a reduction of variance of the estimates takes place.

A closer examination shows that it cannot be rejected that variance reduction appears with equal frequency in all groups, and that variance reduction occurs independent of the type of question (chi-square test, 5 percent level).

Table 4
Frequency of Variance Reduction in Delphi Groups, Round Five compared with Round One (% of all possible cases)

Group Size	Fact-Finding Questions	Forecasting Questions
5	100	80
7	90	80
9	90	100
11	100	100

Hypothesis R will be tested now for judging performance. In order to represent the performance of a group for a certain type of question, individual performances must be aggregated. Since the individual performances are "index numbers" only the geometric mean can be chosen for this. At first we turn to the fact-finding questions. If considered individually, it becomes evident that the lowest and the highest median group error (taken as an absolute value) lie with approximately equal frequency in the first two rounds (cf. Table 5). In case of identical figures for the observed variable, its first appearance was considered.

Table 5
Relative Frequency of the Lowest (and Highest) Median Group Error by Number of Rounds, Group Size, and Type of Question

Group Size	Round (Fact-Finding Questions)				
	1	2	3	4	5
5	0.70(0.40)	0.20(0.40)	0.10(0.20)	0:0(0.0)	0.0(0.0)
7	0.40(0.70)	0.30(0.20)	0.10(0.0)	0.10(0.10)	0.10(0.0)
9	0.40(0.80)	0.30(0.0)	0.30(0.0)	0.0(0.20)	0.0(0.0)
11	0.635(0.40)	0.125(0.20)	0.125(0.10)	0.0(0.10)	0.125(0.0)

Group Size	Round (Forecasting Questions)				
	1	2	3	4	5
5	0.50(0.70)	0.40(0.20)	0.0(4.10)	0.0(0.0)	0.10(0.10)
7	0.70(0.60)	0.10(0.20)	0.0(0.20)	0.10(0.0)	0.10(0.0)
9	0.70(0.70)	0.30(0.10)	0.0(0:0)	:0(0.10)	0.0(0.10)
11	0.375(0.865)	0.50(0.0)	0.0(0.125)	0.0(0.0)	0.125(0:0)

The median values for each group, which are easily read from Table 5, all lie in the, first or second round. If we consider the lowest median group errors (taken absolutely), we observe that round two evidently is of greater importance concerning forecasting questions than concerning fact-finding questions, whereas rounds one and three are of much greater importance for fact-finding questions than for forecasting questions. On the other hand, the highest median group errors (taken absolutely) are much more heavily concentrated in the first round for forecasts than for fact-finding questions.

This analysis does not, however, take into consideration the degree to which the medians' deviate from the correct values. This may be evaluated by looking at the

geometric mean of the individual errors for each round, each group, and each type of question. For it could be possible that good results deteriorate not at all or only very little, while poor results improve greatly with an increasing number of rounds.

Data to judge on this question are presented in Table 6.

Table 6
Geometric Mean of the Individual Errors (taken as absolute values)

Group Size	Round (Fact-Finding Questions)				
	1	2	3	4	5
5	0.20	0.19	0.18	0.23	0.27
7	0.13	0.09	0.07	0.10	0.11
9	0.22	0.14	0.09	0.17	0.16
11	0.24	0.29	0.23	0.28	0.22

Group Size	Round (Forecasting Questions)				
	1	2	3	4	5
5	0.28	0.27	0:28	0.35	0.35
7	0.44	0.42	0.43	0.36	0.36
9	0.25	0.20	0.16	0.16	0.16
11	0.12	0.09	0.10	0.10	0.10

Data are analyzed by Friedman's two-way analysis of variance by ranks.

A significant difference for the entire set of data for fact-finding questions barely fails to be demonstrated at the 10 percent level ($Xr^2 = 7.6$ compared with the tabulated value of 7.78). The computed value of Xr^2 for forecasting questions is considerably lower (5.75) and fails the 20 percent level. The phenomenon that for fact-finding questions in all groups except the one with eleven participants the highest performance is attained in the third round does not influence the tests significantly. This is even less so for forecasting questions, where the best performance is achieved twice in the second round and otherwise in the fourth or fifth rounds.

If one assumes that people get bored with answering the same questions over *and* over again and if one therefore cuts out the results of the last,round, we observe a minor increase in the test statistics. The calculated value for factfinding questions exceeds the significance *level* of 10 percent. For forecasts the increase is so slight ($Xr^2 = 5.77$) that no further consequences can be drawn from it. Taking everything together, our results seem to

indicate that it is not reasonable to extend the number of rounds in Delphi groups beyond the third round.

This would support hypothesis *R* while it refutes hypothesis R'. Since in all cases investigated it is not assumed that the self-rating of expertise varies with the number of rounds, the result cannot be tested further in this respect.

The observations of group performance are not supported by the results of the best individual performance. In no case, i.e., neither when all rounds are considered nor when the number of rounds taken *into* consideration is limited, can a significant difference in the best individual performance be observed which would vary with the number of rounds. The best individual performance is very often maintained over consecutive rounds. So, if one would have objective criteria by which one could pick real experts, one might expect a greater stability of their judgments as compared with that of all members of our present groups.

4.4 The Size of the Groups

We test hypothesis G directly, separately for face-to-face discussion groups and Delphi groups. We do not correct the hypothesis to include the distribution of expertise (as determined subjectively by self-ratings) between groups, as it has practically a random influence.

When the data in Table 6 are analyzed by the rows, significant differences (on the 0.1 percent level) show up in Friedman's analysis of variance by ranks (Xr^2=12.84 for fact-finding questions and Xr^2 = 15 for forecasting questions). It is clear that the .group performance *in* all rounds may be rank ordered as follows:

Fact-finding questions: Group with seven participants on top, followed by the groups with nine, five, and eleven participants.

Forecasting questions: Group with eleven participants on top, followed by the groups with nine, five, and seven participants.

The groups of seven and eleven reverse their rank order of performance in fact-finding questions as compared with forecasting questions.

This observation cannot be explained by varying self-ratings of expertise, as can be read from the refutation of the hypothesis concerning a different distribution of expertise in fact-finding questions and forecasting questions (see section 4.1).

If one considers the best individual performances directly, the result for the groups is confirmed, except for a shift in the rank orders of the groups with five and eleven participants for fact-finding questions, and of the groups with seven and five participants for forecasting questions. The individual estimates of the participants can thus be considered to be one factor which influences the result.

The quality of the estimates could be determined by the frequency of information exchange between the participants.[53] The frequency of information exchange is shown in Table 7.

Table 7
Frequency of Information Exchange between the Participants in Delphi Groups

Group Size	Absolute Number of Information Exchanges				% of All Possible Exchanges of Possible			
	After rounds				After rounds			
	1	2	3	4	1	2	3	4
5	7	9	4	2	17.5	22.5	10.0	5.0
7	10	13	11	15	25.0	32.5	26.5	37.5
9	43	33	17	20	53.8	41.3	21.3	25.0
11	26	17	11	11	40.7	26.5	17.2	17.2

We have aggregated data for fact-finding questions and forecasting questions as we have found that the frequency of information exchange does not vary significantly with the type of question (binomial test, 5 percent level).

We find that the absolute frequency of information exchange between the participants varies significantly between the individual groups (Xr^2 =19.8 is significant on the 5 percent level). Nevertheless the group with nine participants clearly leads the sequence, followed by those groups with eleven, seven, and five participants. If the frequency of information exchange is related to the number of possibilities for information exchange (which depends on the number of questions as well as on the number of participants outside of the quartiles, of course) a significant difference between the groups can likewise be determined ($Xr^2$15.0): In this case only the groups with seven and eleven participants change their positions in the rank order as compared with the rank order of absolute frequencies of information exchange. One could assume that the participants themselves, as keepers of information, channel varying amounts of information into the group. If so, it should be possible to find a rank correlation between group performance and the frequency of information exchange with regard to each question within each group. It turns out, however, that a significant result (5 percent level) can be identified only for the group with nine participants.

Low, and in part negative, values for the rank correlation, particularly in the group with eleven participants, can probably be explained by the fact that the opportunities for

[53] See I. Lorge *et al.,* "Solutions by Teams and Individuals to a Field Problem at Different Levels of Reliability," *Journal of Educational Psychology 46 (1955), pp. 17-24.*

information exchange were used to transmit signs of impatience toward the end of each session. These were not considered to contribute to group performance in the stipulated sense, and thus were not counted in selecting the data of Table 7.

In the face-to-face discussions group performance is distributed differently (see Table 8). For the fact-finding questions we discover the following:

If performance is measured again by a median group error which is composed of individual estimates given before entering discussion, we find that the group with seven participants attains the highest level of performance. The groups with ten, four, and eight participants follow in that order. It should be noted that these data are only approximately comparable to those for round one in Table 7 because the initial data used here are given before the start of the exchange of information. If we take the geometric mean of the absolute values of the differences between the group estimate and the correct values, the group with ten participants appears at the top of the scale of performance. The groups with seven, four, and eight participants follow.

Let us now turn to the forecasting questions.

Table 8
Geometric Mean of the Median of Individual Relative Errors (taken as absolute values) before Discussion and the Relative Error of the Group Estimate in Face-to-Face Discussion Groups

Group Size	Fact-Finding Questions		Forecasting Questions	
	Geometric Mean Individual Errors	Group Errors	Geometric Mean Individual Errors	Group Errors
4	0.171	0.163	0.179	0.184
7	0.116	0.154	0.103	0.147
8	0.257	0.220	0.072	0.121
10	0.141	0.139	0.187	0.212

As before, a corresponding rank order of group performance cannot be demonstrated for the two types of group structure. Here the group with eight participants leads, and the lowest level of performance is attained by the group with ten participants (see Table 8).

When these results are interpreted it should be noted that the fact that in several cases one member of the Delphi groups did not participate in the face-to-face discussion groups does not affect the results of the latter in a uniform fashion.

Moreover, the direct observation of the group activity suggests that the number of members present in a discussion group cannot be considered the decisive variable. It is more important to consider the number of members of a face-to-face discussion group who actively take part in the discussion, as the nonparticipants do not influence the group judgment in any way. This situation occurs in our face-to-face discussion groups.

According to our observations, in the group with eight participants, one member of the group only very rarely took part in the discussions. In the group with ten members two persons did not express any' opinions during the entire experiment. When the group average was determined by voting, they conformed to the majority opinion, which was apparent. Another person only rarely joined in the consultations. Thus the "active" part of the group is almost always reduced to seven persons.[54]

Since in the two smaller groups the group itself required that each member participate on each question, there the "active" group corresponds in size with the entire group.[55] Thus our observation material is reduced practically to "active" discussion groups with four, seven, and possibly eight participants.

After these reservations as to the interpretation of the results, it is not surprising that the rank order of performances, if partitioned with respect to fact-finding questions and forecasting questions, as well as with respect to both types of group structures, does not generally exhibit comparable results in groups of varying size. Results coincide in the smallest group only. This, however, cannot be considered significant.

4.5 Hypothesis D

The problems of drawing a comparison between the results from the Delphi groups and the face-to-face discussion groups have already been mentioned. Besides, one has to find a generally accepted criterion on which to base the comparison. Dalkey's statements are based on the number of cases of superior performance. However, the interest in group performance can be based with at least the same right on the amount of the errors that occur. We take up this latter point.

The geometric average of all group performances for fact-finding questions in face-to-face discussion groups (0.167) is higher than the corresponding value for the third round of Delphi groups, which is as low as 0.127. However, the result *is* reversed for forecasts. Here `the corresponding value of 0.209 for the third round in Delphi groups is higher than the result of 0.162 for the face-to-face discussion groups. Thus an unequivocal relationship cannot be established, The performance of the only group with identical size under different organizational structures also differs greatly.

It is further noteworthy that with regard to the forecasts, the discussion in the discussion groups shows no progress in performance if the results are compared with the geometric means of individual errors before discussion. *On* the other hand, the result of the discussion in all the Delphi group's is that the mean error *is* reduced *up* to the third round. Furthermore, we find that the performance *level* of the ,face-to-face discussion groups *is* approximately equal for fact-finding questions and for forecasts, whereas the performance

---

[54] Unfortunately, only short notices about the impressions of the author after each session of the face-to-face discussion groups exist as a proof of these statements. Video tapes of the discussions do not exist; they would certainly have contributed to the interpretation of the results.

[55] Our observations coincide with the statements of Alter et al., who confirm clique formation with disturbing internal discussion, a higher absolute proportion of inactive group members, and poorer agreements with dominant participants as dysfunctional effects in large groups in brainstorming sessions. Cf. U. Alter, H. Geschka, H. Schlicksupp, "Methoden and Organisation der Ideenfindung," report on a group project of Battelle Institute, Frankfurt (July 1972), Methodological Appendix, p. 20.

level of the Delphi *groups* with regard to forecasts is much lower as compared with fact-finding questions.

4.6 Hypothesis F

The preceding statements have already made it clear that different results are definitely produced when group performance with regard to fact-finding questions *is* differentiated from group performance with regard to forecasts (cf. Tables 6, 8). A general confirmation or refutation of hypothesis F on the. basis of, group performance is not possible. We therefore formulate and test the auxiliary hypothesis 5 in addition to what we have found until now. It says:

> *There exists a positive relationship between the rank order of performance of individuals in each group with regard to fact-finding questions and forecasts. Performance is defined as the geometric mean of the individual errors (taken as absolute values) in answering fact-finding questions and forecasting questions.*

The auxiliary hypothesis is refuted in each of the groups. With increasing group *size* we calculate rank correlations of -0.300, -0.391, +0.357 and -0.150. In the majority of cases not even the sign corresponds to the expectations of the auxiliary hypothesis:

5 Summary

Although one should not overestimate the results from the very few experiments presented here, they do lead to doubts as to the efficiency of the Delphi method. Of course, it must be admitted that the attempt to use the Delphi method for short-term forecasting is a comparatively tough test, for it was originally designed for long-range forecasting. At another occasion (sales estimates) it was also found that the errors of short-term forecasts can be very much higher than those of long-term forecasts.[56] If one assumes that the results of the forecasts could be interpreted as an attempt to estimate an unknown status quo at the time when the experiments took place as well, then they should be corrected by the average value of the relative difference between the status quo and the realization of each topic. These corrections vary between 0.042 and 0.098 in the individual groups, according to the choice of questions. Their application does not alter the rank order of the results as compiled in Tables 6 and 8. Therefore we may refer directly to the text with our summary:

(1) It cannot be discerned that fact-finding questions are suitable test material for recognizing expertise or appropriate organizational structures for forecasting groups.
(2) A general positive relationship between group size and group performance cannot be recognized.

---

[56] J. Berthel, D. Moews, Information *and* Planung *in industriellen Unternehmungen,* Berlin, 1970, pp. 158 ff. (with data from fifteen firms).

(3) In face-to-face discussion groups the measure of the group size must be determined by the number of active participants. Appropriate precautions should be developed.

(4) Variance reduction almost always occurs in Delphi groups between the first and the fifth rounds, but the best results are as a rule already known in the third round. Further rounds may impair the results.

(5) Self-ratings of expertise show a positive relationship to the performance of the persons questioned in only two of four Delphi groups. They tend to be lower in face-to-face discussion groups than in the Delphi groups, and are determined substantially by the extent of professional experience rather than being set with regard td the questions in case. It is important to employ and develop better methods for the determination of expertise.

(6) A direct comparison of Delphi groups and face-to-face discussion groups was not possible because several participants dropped out. However, the results, if separated for fact-finding questions and forecasts, do not point in one direction.

(7) Proponents of the Delphi method will point out that our subjects, being banking experts, are better able to express themselves in a face-to-face discussion than on a display, even if its use has been explained to them. This should apply particularly when the space for exchanging information among participants is limited. The fear of making mistakes in the operation of the display could lead to exaggerated caution. However, if one agrees with this argument, the first point of this summary has to be explained also, as the results are not uniform in this respect. Anyhow, it appears to be important to avoid a "new Taylorism," on which some mis-concepts are founded.[57] However, it must be granted that the originators of the Delphi method did not say that it has to be operated as a computer dialogue.

(8) Only in the Delphi group with the greatest exchange of information did we observe a positive relationship to group performance. The results indicate that in small Delphi groups more opportunities for information exchange should be given. However, it probably must be tested whether the information given by the participants does coincide with what others would want to know, i.e., whether it adds to their knowledge.[58] How can the "confusion effect" in the majority of the discussions, which is recognized when the "reference values" mentioned there are compared (cf. Table 8), be explained without this distinction and the assumption of a difference between information supplied and information demanded?

(9) It must be admitted that in our strongly discipline-oriented group there has been relatively little opportunity for improving estimate by sharing information as compared to interdisciplinary groups concerned with other tasks. However, this affects all our groups in the same way. This criticism would be valid if it could be demonstrated that different groups react differently to these two types of tasks.

---

[57] See W. Kirsch, "Auf dem Wege zu einem neuen Taylorismus?" in H. R. Hansen, M. P. Wahl (eds.), *Probleme bein Aufbau betrieblicher Informationssysteme,* Miichen, 1973, pp. 338-48.

[58] E. Witte (ed.), *Das Informationsverhalten in Entscheidungsprozessen,* Tiibingen, 1972, especially, pp. 44ff.; R. Bronner, E. Witte, B. R. Wossidlo, "Betriebswirtschaftliche Experimente zum Informations Verhalten in Entscheidungsprozessen," in E. Witte, op. *cit., pp.* 186ff.