DOCUMENT RESUME

ED 064 302 TM 001 420

AUTHOR Macmillan, Thomas T.
TITLE ["The Delphi Technique."]
INSTITUTION Santa Barbara City Schools, Calif.
PUB DATE May 71
NOTE 24p.; Paper presented at the annual meeting of the California Junior Colleges Associations Committee on Research and Development (Monterey, Calif., May 3-5, 1971)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS College Students; Comparative Analysis; *Data Analysis; *Evaluation Techniques; *Feedback; *Group Discussion; *Information Processing
IDENTIFIERS *Delphi Technique

ABSTRACT

The Delphi technique is a method of eliciting and refining group judgments. The procedures used have three features: anonymous response, iteration and controlled feedback, and statistical group response. A series of experiments were initiated at RAND to evaluate the procedures. Upper-class and graduate students were used as subjects and general information of the almanac type as subject matter. The two basic issues examined were: (1) a comparison of face-to-face discussion with the controlled-feedback interaction, and (2) a thorough evaluation of controlled feedback as a technique of improving group estimates. The results indicated that face-to-face discussion tended to make the group estimates less accurate, whereas, the anonymous controlled feedback made the group estimates more accurate. Other results include: (1) the insight gained into the nature of the group information processes, (2) the fact that a meaningful estimate of the accuracy of a group response to a given question can be obtained by combining individual self-ratings of competence on that question into a group rating. It is concluded that the experiments represent a beginning of a field of research that could be called "opinion technology." (Author/CK)

The Delphi technique is a method of eliciting and refining group judgments. The rationale for the procedures is primarily the age-old adage "Two heads are better than one," when the issue is one where exact knowledge is not available. The procedures have three features: (1) Anonymous response—opinions of members of the group are obtained by formal questionnaire. (2) Iteration and controlled feedback—interaction is effected by a systematic exercise conducted in several iterations, with carefully controlled feedback between rounds. (3) Statistical group response—the group opinion is defined as an appropriate aggregate of individual opinions on the final round. These features are designed to minimize the biasing effects of dominant individuals, of irrelevant communications, and of group pressure toward conformity.

In the spring of 1968, a series of experiments were initiated at RAND to evaluate the procedures. The experiments were also designed to explore the nature of the information processes occurring in the Delphi interaction. The experiments were conducted using upper-class and graduate students from UCLA as subjects and general information of the almanac type as subject matter. Ten experiments, involving 14 groups ranging in size from 11 to 30 members, were conducted. About 13,000 answers to some 350 questions were obtained.

The two basic issues being examined were (1) a comparison of face-to-face discussion with the controlled-feedback interaction, and (2) a thorough evaluation of controlled feedback as a technique of improving group estimates. The results indicated that, more often than not, face-to-face discussion tended to make the group estimates less accurate, whereas, more often than not, the anonymous controlled feedback procedure made the group estimates more accurate. The experiments thus put the application of Delphi techniques in areas of partial information on much firmer ground.

Of greater long-range significance is the insight gained into the nature of the group information processes. Delphi procedures create a well-defined process that can be described quantitatively. In particular, the average error on round one is a linear function of the dispersion of the answers. The average amount of change of opinion between round one and round two is a well-behaved function of two parameters—the distance of the first-round answer from the group median, and the distance from the true answer.

Another result of major significance is that a meaningful estimate of the accuracy of a group response to a given question can be obtained by combining individual self-ratings of competence on that question into a group rating. This result, when combined with the relationship between accuracy and standard deviation mentioned above, opens the possibility of attaching accuracy scores to the products of a Delphi exercise.

A number of supplementary analyses—including the effect of time-to-answer on accuracy, the comparison of performance as a function of college major, and the effect of different question format--have added useful elements to the overall picture, giving additional weight to the presumption that information-handling procedures that are appropriate for well-confirmed material are not suitable for the less well confirmed area of expert opinion.

Although the experiments conducted to date have been informative beyond initial expectations, they represent only a small beginning in a field of research that could be called "opinion technology."

CONTENTS

1. THE SPECTRUM OF DECISION INPUTS

One of the thorniest problems facing the policy analyst is posed by the situation where, for a significant segment of his study, there is unsatisfactory information. The deficiency can be with respect to data—incomplete or faulty—or more seriously with respect to the model or theory—again either incomplete or insufficiently verified. This situation is probably the norm rather than a rare occurrence.

The usual way of handling this problem is by what could be called "deferred consideration." That is, the analyst carries out his study using whatever good data and confirmed models he has and leaves the "intangibles" to the step called "interpretation of results."* In some cases the deferment is more drastic. The analyst presents his study, for what it is worth, to a decisionmaker, who is expected to conduct the interpretation and "inclusion in the total picture."

In describing the interpretation-of-results step, interesting words are likely to appear. These include terms like "judgment," "insight," "experience," and especially as applied to decision-makers, "wisdom" or "broad understanding." These terms contrast with the presumed precision, scientific care, and dependence on data that characterize operations research. Above all, there is a slightly mystical quality about the notions. They are never explained. Standards of excellence are lacking. And there is more than a hint that the capabilities involved somehow go beyond the more mundane procedures of analysis.

---

*The not infrequent case where the analyst "makes do" with faulty data or shaky models has been sufficiently excoriated in the manuals of operations research methodology.

Taking a look at the kinds of information that can play a role in decisionmaking, there are roughly three types (see Fig. 1). On the one hand, there are assertions that are highly confirmed--assertions for which there is a great deal of evidence backing them up. This kind of information can be called knowledge. At the other end of the scale is material that has little or no evidential backing. Such material is usually called speculation. In between is a broad area of material for which there is some basis for belief but that is not sufficiently confirmed to warrant being called knowledge. There is no good name for this middling area. I call it opinion. The dividing lines between these three are very fuzzy, and the gross trichotomy smears over the large differences that exist within types. However, the three-way split has many advantages over the more common tendency to dismiss whatever is not knowledge as mere speculation.

Where in this scale do the products of judgment, wisdom, insight, and similar intellectual processes, lie? Not in speculation, we hope. And, almost by definition, not in knowledge. The most reasonable interpretation would be that these are flattering names for kinds of opinion.*

Unfortunately, there is no practical, objective measure for the dimension of evidence sketched in Fig. 1. The best we have is an intuitive and rough feeling for the scale.** The prototype of knowledge may be found in the systematized, experimentally confirmed propositions of the natural sciences. But many of the assertions in the area that is called "common sense" have an equal solidity; e.g., the gross features of

---

*One might say, "Wisdom is opinion with charisma."

**A Delphi approach for locating assertions on the evidence scale will be discussed in Section 9.8, p. 68ff.

SPECTRUM OF INPUTS

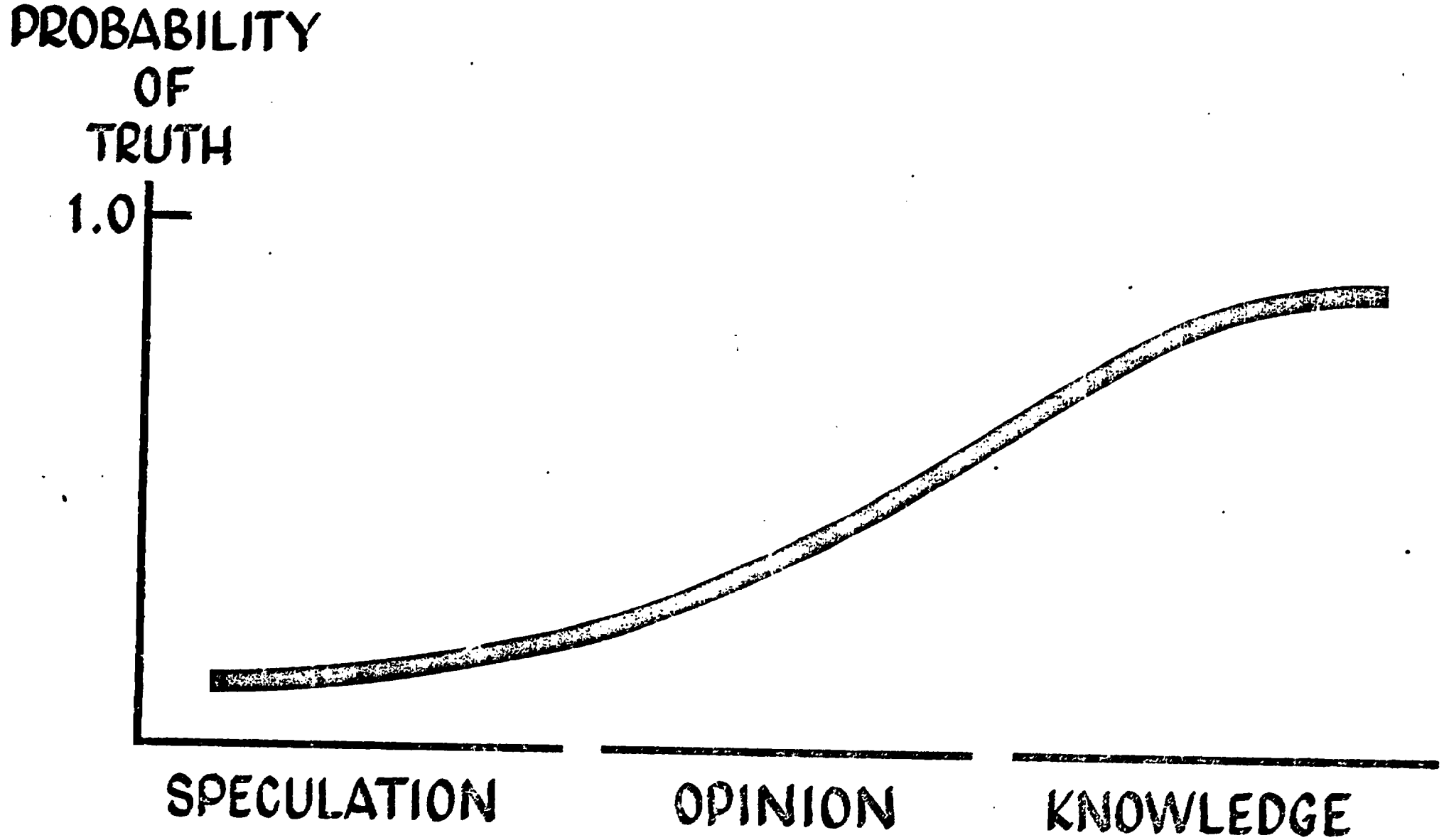

Fig. 1

gravity--"unsupported objects fall to the surface of the earth", the permanence of objects, and the like. A large part of the empirical generalizations of common technology are equally well confirmed. The technologist's criterion--does it work?--is at least as effective in eliminating unfounded notions as the scientist's is it confirmed by laboratory experiment?

In the following it will be taken for granted that methods of dealing with material in the area of knowledge are in reasonably good order. There are, of course, many problems of detail--the warrantability of extrapolation, the application of statistical measures where underlying distributions are unknown, and the like. But these difficulties are small compared with the conceptual vacuum that appears to exist in the area of opinion.

With respect to speculation, it appears very difficult to say anything wise other than to avoid it whenever possible. That isn't very helpful. It appears likely that most major policy decisions involve more than a dash of speculative inputs. Some of the general results described below are applicable to speculation, but how useful it is to the decisionmaker to furnish him with refined speculation is hard to say.

This report sidesteps the even more difficult issue raised by the fact that most practical decision situations involve a mixture of all three types of information. The delicate balancing of the weight to give each kind of material is a second-level sort of "wisdom" that has not yet been investigated.

In discussions of policy analysis it is usual to distinguish two kinds of assertions, factual statements and value judgments. It is an open question whether there is any basic conceptual difference between these two, but there are certainly very large practical differences.

In particular, value judgments tend to be much vaguer and displaced toward the opinion and speculative and of the solidity scale. The experimental results described below are concerned with factual material, but there is a short comment on value judgments in Section 10.

With respect to factual statements, it is worth pointing out that the crude scale of "solidity" is related to the likelihood that assertions are true. In the area of knowledge, by definition the probability of an assertion being true is relatively high; for speculative material the probability is low; and for opinion it is middling (see Fig. 1). This point is rather vital. There is an irrepressible urge on the part of analysts to move the arena of action entirely into the knowledge area. Sometimes this is possible. In general, it is not. When an opinion is expressed, it is an inescapable fact of life that whatever is said, there is a reasonable probability of its being false.

2. TWO HEADS ARE BETTER THAN ONE

There is a kind of technology for dealing with opinion that has been applied throughout historical times and probably in more ancient times as well. The technology is based on the adage "Two heads are better than one," or more generally "n heads are better than one." Committees, councils, panels, commissions, juries, boards, the voting public, legislatures ....the list is long, and illustrates the extent to which the device of pooling many minds has permeated society.*

The basis for the n-heads rule is not difficult to find. It is a tautology that, on any given question, there is at least as much relevant information in n heads as there is in any one of them. On the other hand, it is equally a tautology that there is at least as much misinformation in n heads as there is in one. And it is certainly not a tautology that there exists a technique of extracting the information in n heads and putting it together to form a more reliable opinion. With a given procedure, it may be the misinformation that is being aggregated into a less reliable opinion.

The n-heads rule, then, depends upon the procedures whereby the n heads are used. There is one kind of procedure and one kind of factual judgment where the n-heads rule comes very close to a tautology. Consider the case where the judgment required is a numerical estimate--e.g.,

---

*Most of these groups have more than one function. They can operate to transmit information, to coordinate action, to diffuse responsibility, to formulate policy, etc. All of these functions are important. None of the discussion below should be taken to apply directly to these other functions. In the present context we are concerned with the use of groups to formulate factual judgments. If the results of the present study appear suggestive with regard to the other functions of groups, I can only hope that this tends to generate additional experimentation.

the date at which a certain technological development will occur, or the size of world population in 1990—and assume you have a group of indistinguishable experts with respect to this estimate; that is, you have no way of asserting that one expert is more knowledgeable than another. Is it better to select the opinion of one expert at random or to take some statistical aggregate of the opinions of the group? It is a near-tautology that you are at least as well off to take the mean or the median as to select an expert at random.* This is, of course, a very weak statement. It can be most simply illustrated by using the median as the statistical representative of the group answer. Referring to Fig. 2, it is clear than, independent of the distribution of answers, and independent of the location of the true answer T, the median of the individual answers M is at least as close to the true answer as <u>one-half of the group</u>. If the range of group answers includes the true, then, in general, the median is closer to the true answer than more than half of the group, as in Fig. 3.

In practical situations, the range of answers is very likely to include the true answer, in which case the stronger assertion is valid. Fig. 4 shows the dependence on group size of the mean accuracy of a group response for a large set of experimentally derived answers to factual questions. The curve was derived by computing the average error of groups of various sizes where the individual answers were drawn from the experimental distribution. The error is

---

*The precise statement is: for the median, the probability that the median is at least as close to the true answer as any individual response is at least one half; for the mean, the error of the mean, (measured by the distance to the true answer) is less than or equal to the average error of the individual answers. These two criteria are not equivalent, and for different decision situations one or the other could be more appropriate.

WORST CASE:
MEDIAN BETTER THAN HALF

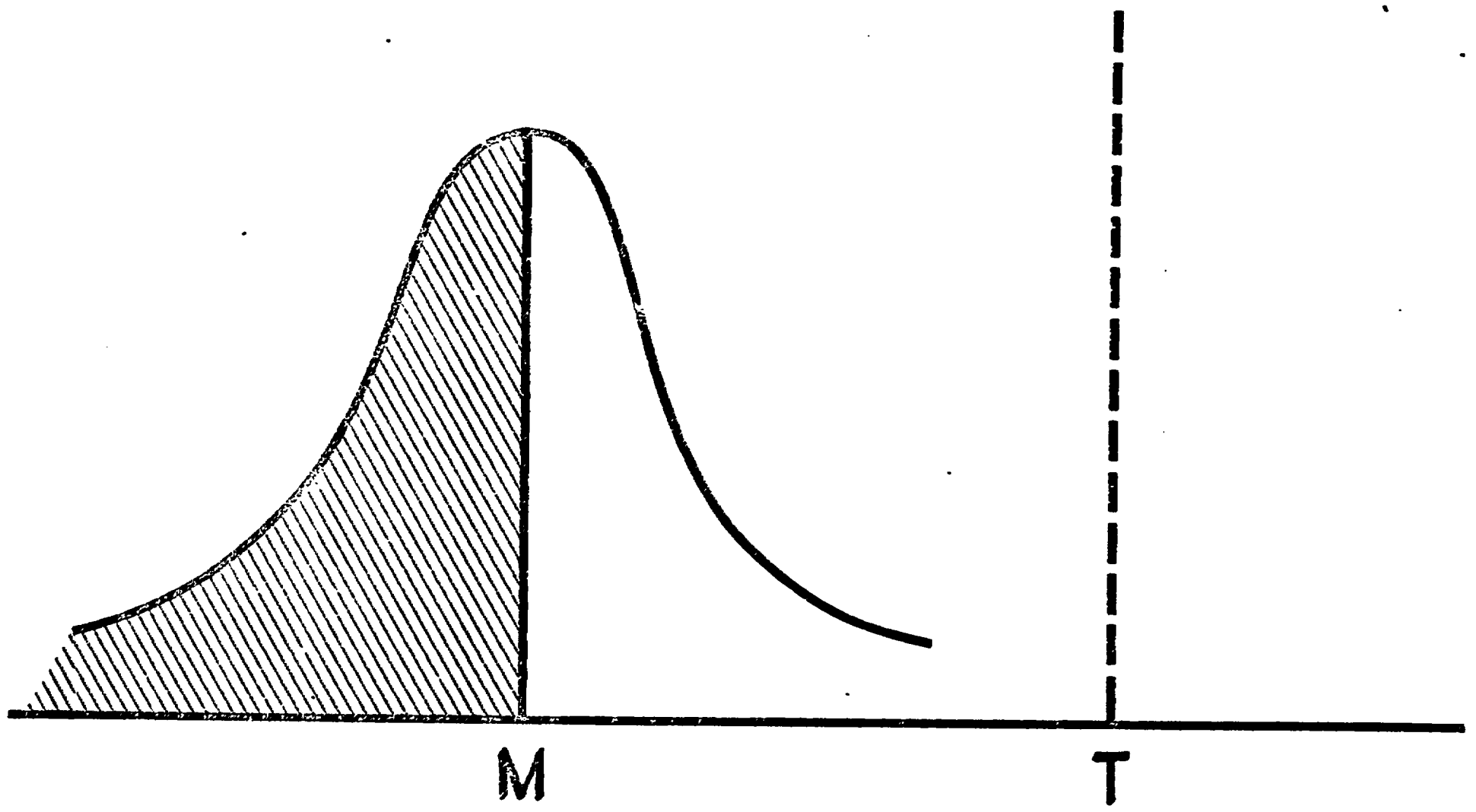

Fig. 2

NORMAL CASE:
MEDIAN BETTER THAN MORE THAN HALF

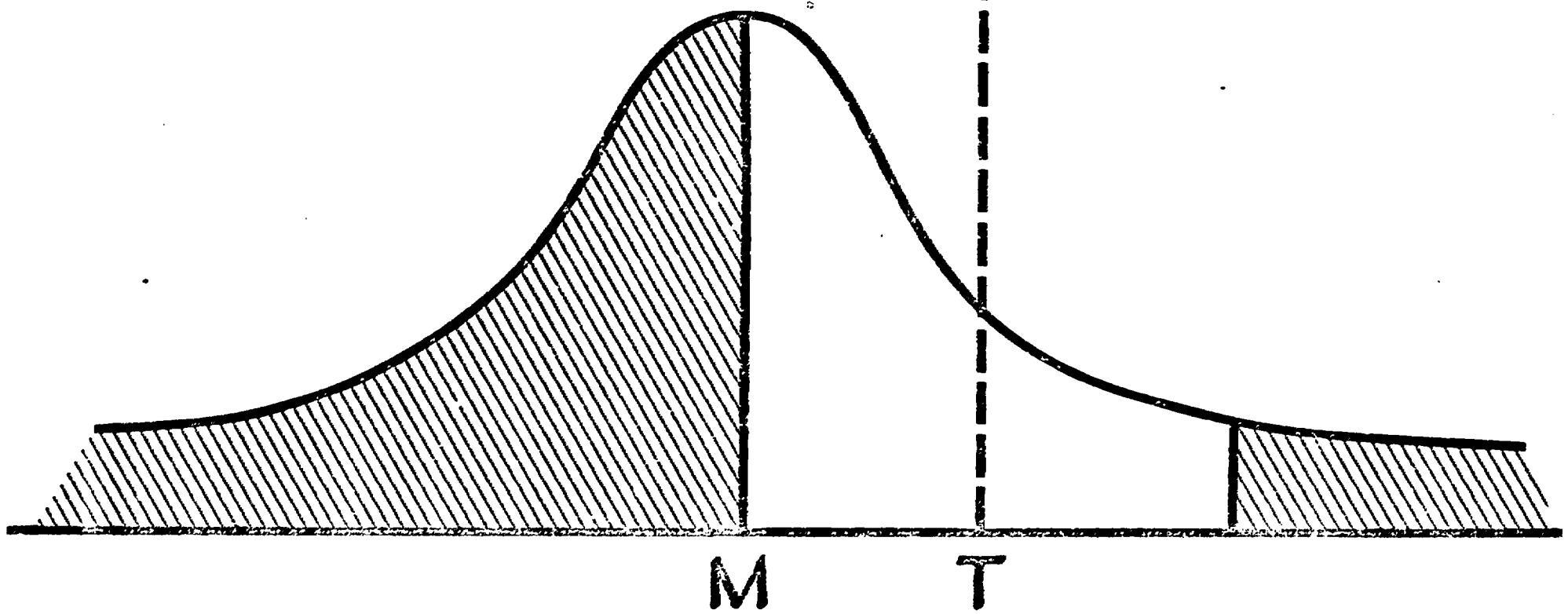

Fig. 3

measured on a logarithmic scale.* It is clear from Fig. 4 that with this population of answers, the gains in increasing group size are quite large. It is interesting that the curve appears to be decreasing in a definite fashion, even with groups as large as 29. This was the largest group size we used in our experiments.

Another important consideration with respect to the n-heads rule has to do with reliability. The most uncomfortable aspect of opinion from the standpoint of the decisionmaker is that experts with apparently equivalent credentials (equal degrees of expertness) are likely to give quite different answers to the same question. One of the major advantages of using a group response is that this diversity is replaced by a single representative opinion.** However, this feature is not particularly interesting if different groups of experts, each made up of equally competent members, come up with highly different answers to the same question.

In general, one would expect that in the area of opinion group responses would be more reliable than individual opinions, in the simple sense that two groups (of equally competent experts) would be more likely to evidence similar answers to a set of related questions than would two

---

*These were questions where the experimenters knew the answer but the subjects did not. The group error is the absolute value of the natural logarithm of the group median divided by the true answer. The groups used to construct Fig. 4 were "synthetic"; i.e., they were randomly selected sets of answers of the appropriate number drawn from the experimental distributions of answers.

**Whether this is the best use of group opinion, or whether the decisionmaker should take into account the full distribution of answers, and also make use of ranges of uncertainty on the part of individual respondents is an important topic in its own right, that will be partially explored in later sections.

EFFECT OF GROUP SIZE

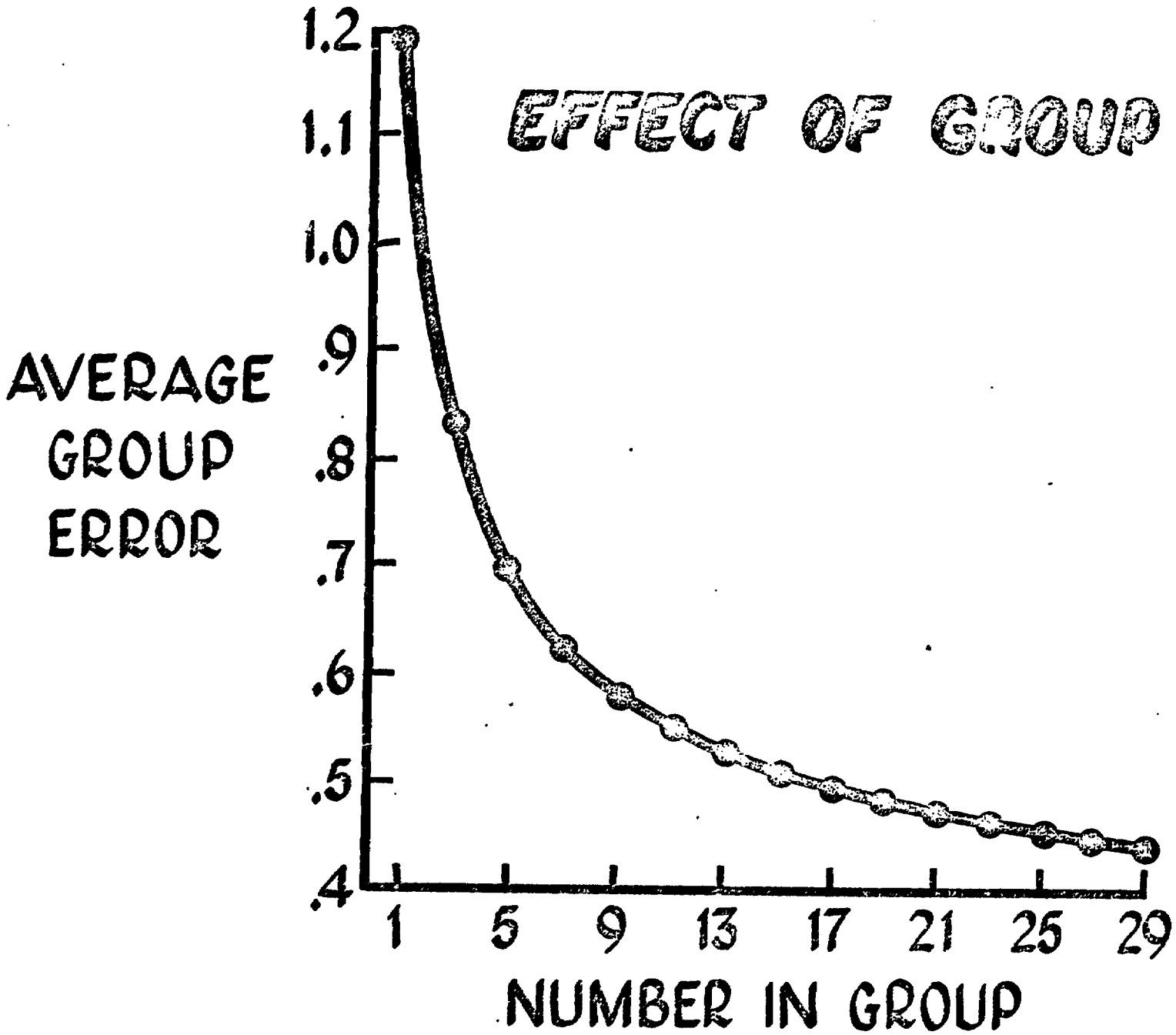

Fig. 4

individuals. This "similarity" can be measured by the correlation between the answers of the two groups over a set of questions. But the assertion that groups will be more reliable than individuals is not a tautology. It depends on the distributions of answers that would be obtained from the total population of potential respondents, and it depends upon the method of selecting the subgroups out of this population. The result can be expected to hold if the distributions of answers for the potential population are not highly distorted, and if the subgroups are selected at random. There are clearly implications of this remark for the rules for selecting members of advisory bodies—in practice small advisory groups are probably never selected at random out of the total potential pool of experts.

For the analyst using expert opinion within a study, reliability can be considered to play somewhat the same role as reproducibility in experimental investigations. It is clearly desirable for a study that another analyst using the same approach (and different experts) arrive at similar results.

Fig. 5 shows the relationship between reliability and group size for the experimental population of answers to questions already mentioned. It was constructed by selecting at random pairs of groups of respondents of various sizes and correlating the median responses of the pairs on twenty questions. The ordinate is the average of these correlations.

It is clear that there is a definite and monotonic increase in the reliability of the group responses with increasing group size. It is not clear why the relationship would be approximately linear between n = 3 and n = 11.

In the area of opinion, then, the n-heads rule appears to be justified by considerations of both improved average accuracy, and reliability. The question remains whether

RELIABILITY VS GROUP SIZE

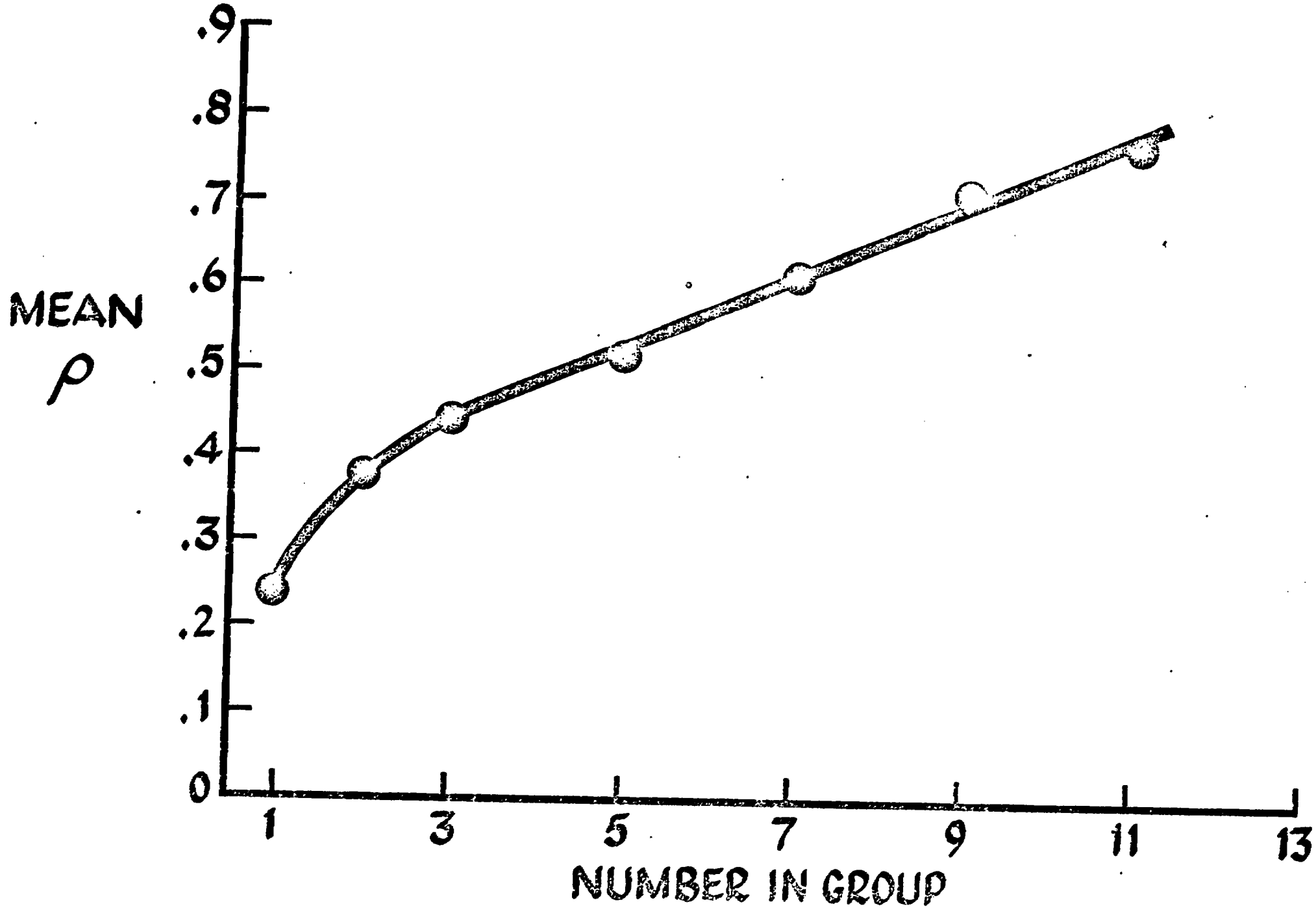

Fig. 5

these quasi-statistical properties of group opinion can be improved upon by allowing more direct pooling of information on the part of the group.

The traditional way of pooling individual opinions is by face-to-face discussion. Numerous studies by psychologists in the past two decades have demonstrated some serious difficulties with face-to-face interaction [2]. Among the most serious are: (1) Influence of dominant individuals. The group opinion is highly influenced, for example, by the person who talks the most. There is very little correlation between pressure of speech and knowledge. (2) Noise. By noise is not meant auditory level (although in some face-to-face situations this may be serious enough!) but semantic noise. Much of the "communication" is a discussion group has to do with individual and group interests, not with problem solving. This kind of communication, although it may appear problem oriented, is often irrelevant or biasing. (3) Group pressure for conformity. The experiments of Asch [3] demonstrate in dramatic fashion the distortions of individual judgment that can occur from group pressure.

In experiments at RAND and elsewhere, it has turned out that, after face-to-face discussion, more often than not the group response is less accurate than a simple median of individual estimates without discussion.

3. DELPHI

There has been a somewhat intermittent series of studies at The RAND Corporation since its early days concerned with the problem of using group information more effectively. The early studies were concerned mainly with improving the statistical treatment of individual opinions [4]. They indicated that some formal properties of individual estimates (precision, definiteness) could be used to rate the success of short-term predictions, and that background information (as measured by a standard achievement test) had a small but significant influence on the success of predictions. Both of these effects were fairly well washed out by combining estimates into group predictions.

In 1953, Dalkey and Helmer [5] introduced an additional feature, namely iteration with controlled feedback. The set of procedures that have evolved from this work has received the name "Delphi"—a somewhat misleading appelation, since there is little that is oracular about the methods.

The Delphi procedures received a very large boost in general interest with the publication of Gordon and Helmer's study of forecasting technological events [6]. In the area of long-range forecasting, it is difficult to dodge the fact that a large part of the activity is at least within the area of opinion, and possibly worse. That particular study happened to coincide with a surge of interest in long-range forecasting itself, with an attendant interest in the systematic use of expert opinion.

In the last three years there has been a very large increase in applications of the procedures, primarily by industry for the forecasting of technological developments[7], but also by a variety of organizations for exploring policy decisions in areas such as education, public transportation, public health, etc. At present, it is difficult to obtain a clear picture of how widespread the applications are; but a crude guess would put the number of studies recently completed, under way, or in the planning stages at well over a hundred.

In light of this widespread exploitation, the question of just how effective the procedures are has considerable practical import.

In general, the Delphi procedures have three features: (1) anonymity, (2) controlled feedback, and (3) statistical group response. Anonymity, effected by the use of questionnaires or other formal communication channels, such as on-line computer communication, is a way of reducing the effect of dominant individuals. Controlled feedback—conducting the exercise in a sequence of rounds between which a summary of the results of the previous round are communicated to the participants—is a device for reducing noise. Use of a statistical definition of the group response is a way of reducing group pressure for conformity; at the end of the exercise there may still be a significant spread in individual opinions. Probably more important, the statistical group response is a device to assure that the opinion of every member of the group is represented in the final response. Within these three basic features, it is, of course, possible to have many variations.

There are several properties of a Delphi exercise that should be pointed out. The procedure is, above all, a rapid and relatively efficient way to "cream the tops of the heads" of a group of knowledgeable people. In general,

it involves much less effort for a participant to respond to a well-designed questionnaire than, for example, to participate in a conference or to write a paper. A Delphi exercise, properly managed, can be a highly motivating environment for respondents. The feedback, if the group of experts involved is mutually self-respecting, can be novel and interesting to all. The use of systematic procedures lends an air of objectivity to the outcomes that may or may not be spurious, but which is at least reassuring. And finally, anonymity and group response allow a sharing of responsibility that is refreshing and that releases from the respondents inhibitions. I can state from my own experience, and also from the experience of many other practitioners, that the results of a Delphi exercise are subject to greater acceptance on the part of the group than are the consensuses arrived at by more direct forms of interaction.

I believe all of these features of a Delphi exercise are desirable, especially if the exercise is conducted in the context of policy formulation where group acceptance is an important consideration. Like any technique for group interaction, the Delphi procedures are open to various misuses; much depends on the standards of the individual or group conducting the exercises.

4. EXPERIMENTS

In addition to questioning the effects on free expression of opinion and group acceptance, it still must be asked whether the use of iteration and controlled feedback have anything to offer over the "mere" statistical aggregation of opinions. I put "mere" in quotation marks; in the area of opinion much can be gained by the simple arithmetical pooling of individual opinions as shown above. To get some measure of the value of the procedures, and also to obtain, as a basis for improving the procedures, some insight into the information processes that occur in a Delphi exericse, we undertook a rather extensive series of experiments at RAND starting in the spring of 1968.* We used upper-class and graduate students, primarily from UCLA, as subjects. They were paid for their participation. For subject matter we chose questions of general information, of the sort contained in an almanac or statistical abstract. Typical questions were: "How many telephones were in use in Africa in 1965?" "How many suicides were reported in the U.S. in 1967?" "How many women marines were there at the end of World War II?" This type of material was selected for a variety of reasons: (1) We wanted questions where the subjects did not know the answer but had sufficient background information so they could make an informed estimate. (2) We wanted questions where there was a verifiable answer to check the performance of individuals and groups. (3) We wanted questions with numerical answers to a reasonably wide range of performance could be scaled. As far as we

---

*The team involved in these experiments consisted, in addition to myself, of Bernice Brown, Tom Brown, Samuel Cochran, Olaf Helmer and Richard Rochberg. The fruitfulness of the experimental program is directly ascribable to the high level of competence of these co-workers.

can tell, the almanac type of question fits these criteria quite well. There is the question whether results obtained with this very restricted type of subject matter apply to other kinds of material. We can say that the general-information type of question used had many of the features ascribable to opinion: namely, the subjects did not know the answer, they did have other relevant information that enabled them to make estimates, and the route from "other relevant information" to an estimate was neither immediate nor direct.*

For about half of the experiments, the design called for a control group and an experimental group, each of about 15 subjects. For the others, the iterative structure allowed the group to be its own control. The experiments were conducted as closed information sessions; no inputs beyond the background information of the subjects were introduced. The standard task was answering 20 questions of an almanac sort. The questions were different from experiment to experiment (to preclude inadvertent transfer of information outside the experiments). The basic feedback between rounds was the median and the upper and lower quartiles of the previous-round answers. Additional feedback, summarized from subject responses, was introduced in some cases for experimental evaluation. Altogether, there were 11 experiments, involving close to 5000 answers to some 300 questions on each of several rounds. I will not describe

---

*The results from other experiments using as subject matter short-range prediction of economic, technological, and social events [8,4] appear to substantiate the assumption that there is very little difference between the general properties of answers to our estimation-type questions and the short-range predictions; e.g., with respect to distribution of answers, convergence on feedback, relative accuracy of individual and group responses, etc. However, this observation should be confirmed with more controlled exercises.

all the details of each experiment but will present a resume of the major results.*

The general outcome of the experiments can be summarized roughly as follows: (1) On the initial round, a wide spread of individual answers typically ensued. (2) With iteration and feedback, the distribution of individual responses progressively narrowed (convergence). (3) More often than not, the group response (defined as the median of the final individual responses) became more accurate. This last result, of course, is the most significant. Convergence would be less than desirable if it involved movement away from the correct answer.

---

*Details of procedure, the list of questions employed, and specific outcomes of the experiments are contained in [9].