EXPERIMENTS IN GROUP PREDICTION

N. C. Dalkey

March 1968

P-3820

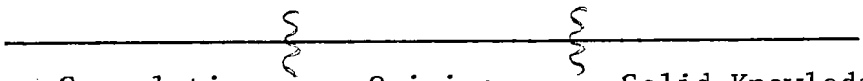
EXPERIMENTS IN GROUP PREDICTION

N. C. Dalkey*

The RAND Corporation, Santa Monica, California

INTRODUCTION

With the advent of long-range planning in military, social, and industrial enterprises, greater attention is being paid to the less rich grades of mental ore. We can grade predictions according to a simple scale:



Statements about the future can (and do) run the gamut from highly confirmed assertions (such as the earth will still exist in the year 2000) all the way to propositions for which there is no foundation whatsoever (examples of which will be left as an exercise for the reader).

As you are all aware, this scale is not precisely related to the truth or falsity of the propositions in question. A highly confirmed assertion can be false; and

---

*Any views expressed in this paper are those of the author. They should not be interpreted as reflecting the views of The RAND Corporation or the official opinion or policy of any of its governmental or private research sponsors. Papers are reproduced by The RAND Corporation as a courtesy to members of its staff.

This paper was prepared for presentation at the National Meeting of the American Chemical Society held in San Francisco, California, April 1 - 4 1968.

a sheer speculation can be true. However, there is an important relationship. A highly confirmed assertion is much more likely to be true than its antipodal cousin.

There does not seem to be a well-established term for the middle region. I call it opinion. Some writers appear to believe that there is a special kind of opinion--often called "wisdom," or "expert opinion" or "informed conjecture." As far as I can tell, these special terms are concerned more with the abstruseness of the subject matter than with the degree of confirmation of the statements so labeled.

Opinions about opinion also run a gamut. At one extreme is the view that anything left of the fuzzy line at the right is worthless for designing policy. At the other extreme is the point of view that decisions cannot wait on highly confirmed knowledge. Faced with a pressing problem, the decision maker must take the "best information available" and act accordingly.

I am inclined to believe that wisdom lies in the middle of these two views. At all events, policy decisions will be made on the basis of opinion; and there remains the separable question whether techniques exist for refining it.

There are a number of traditional approaches in what might be called "opinion technology." (1) Select the best qualified expert. (2) Embed opinions in a more comprehensive conceptual structure, along with whatever knowledge is available. (3) "n heads are better than one."

The three are not incompatible, and, in fact are often combined, as in the PERT methodology for scheduling technological developments. Today I will be concerned primarily with the third approach.

The precept "several heads are better than one" is based on the near tautology that the amount of information available to several heads is greater than the amount available to one. This is not quite a tautology because of the term "available." Although the total knowledge contained in the "backgrounds" of several individuals is undoubtedly greater than that for one, whether this information can be elicited and merged to form a more solid group opinion is not obvious. In addition, there is the counter-vailing tautology that the amount of misinformation available to a group is greater than that available to a single individual, and there is no question that such misinformation can throw sand in the deliberative gears.

The usual way of using a group for the formulation of opinion is by way of face-to-face discussion in a committee, commission, or panel. A large number of recent investigations by social psychologists (1) have demonstrated that face-to-face discussion has serious drawbacks. Chief among these are: influence of dominant individuals, noise, and group pressure toward conformity.

DELPHI

To ameliorate some of these difficulties, a procedure called Delphi has been developed at RAND. This procedure has three basic features: (1) Anonymity. The opinions of the group are recorded separately--usually by questionnaire--and when communicated to other members of the group are not attributed to specific individuals. (2) Controlled feedback. An exercise is conducted in several rounds in which the opinions generated during one round are fed back to the group on the next round, usually in the form of statistical summaries. (3) Statistical group response. The "group opinion" is expressed in terms of a statistical score--the median of final responses has proved to be most suitable for numerical estimates. There is no pressure to arrive at a "consensus."

This procedure has been employed in about thirty exercises at RAND and elsewhere, about a third of these being experiments to investigate the information processes involved, and the remainder have been applications to substantive questions. Most of the applications have been in the area of forecasting long-range social and technological events, presumably because the ratio of opinion to knowledge is particularly high in this area.

The results of these exercises can be briefly summarized as follows: (1) In almost all cases, there is a pronounced convergence of opinion with iteration. On the

initial round, opinions tend to have a wide spread. This spread decreases monotonically on succeeding rounds. (2) the principle decrease is between the first and second rounds. (3) Most significant, for those cases where the accuracy of responses can be checked, the accuracy of the group response increases with iteration.

LOGICAL BOOTSTRAPS

To obtain a clearer understanding of the Delphi process, and to lay a more solid foundation for improvements, we have been conducting a series of experiments at RAND.*

We have concentrated on the closed information case; i.e., during the exercise, no new information concerning the subject matter is introduced into the group. Even in this case, the accuracy of the group response increases with iteration--rather like lifting itself by its logical bootstraps on the part of the group.

The subjects for the experiments have been college upper-class and graduate students. They are paid for participating. No formal screening is performed, but the hiring procedure apparently acts to some extent as a screening device. Very few of the students are from the physical sciences or engineering; most are from the social sciences and the humanities.

The subject matter we have been working with is general information of the sort to be found in an almanac or

*The team involved in these experiments consists of B. Brown, T. Brown, S. Cochran, O. Helmer, and myself.

statistical abstract. Typical questions are: "What was the number of popular votes Kennedy received in Texas in the 1960 presidential election?" "How many billion dollars did consumers in the U.S. spend on recreation in 1965?" This type of material is employed because the accuracy can be checked, and because, although the students would not be expected to know the precise answers, they would be expected to have a large amount of miscellaneous information which is relevant to the answers. As you can see, this is roughly the kind of intellectual situation that exists for long-range forecasts, where we generally have little in the way of precise theory, but a great deal of pertinent knowledge.

We plan in later experiments to extend the subject matter to short-term forecasts of economic, political, and technological events to see if the change to what could be called "objective uncertainty" makes a difference, but have not done so as yet.

A standard task was devised for the experiments, namely, the answering of twenty questions of the almanac type. A new set of twenty questions was devised for each experimental session. In addition to answering the twenty questions, the subjects were asked to rate themselves according to their competence for each question. In most of the sessions, a standardized intelligence test (the Terman Concept Mastery Test) was also administered.

For all but one of the experimental sessions, a uniform experimental design has been employed, consisting of an experimental group and a comparison group. Students are allocated to the two groups at random. In general, the control group follows a "simple" Delphi procedure: answering the twenty questions in the first round, receiving feedback of the medians and 25% and 75% quartiles of the first-round answers, and revising their estimates. The experimental group engages in some variation of this procedure. The experimental design may appear inefficient, but we have discovered that there are large differences in performance due to differences in questions, and the design is intended mainly to control this variation.

RESULTS

It would be much too time consuming to describe all of the experiments and their outcomes in detail. In addition we are right in the middle of the planned set of exercises, and much of what I say in the way of interpretation of results is only tentative. I will try to give some of the highlights, as well as some of the lowlights, of our analyses to date.

An initial experiment conducted to compare the efficacy of the Delphi procedure with face-to-face discussion, indicated that the first-round "off-the-cuff" estimates of the subjects are at least as accurate as the "consensus" reached by the face-to-face group after a

half-hour discussion of each question (2). There was a much richer exchange of information during the face-to-face discussion than occurred in the questionnaire sessions, but degradations due to the three factors mentioned above--dominant individuals, noise, and group pressure--were quite evident.

A basic question is whether the estimation process is a definable task, or whether performance is highly erratic, and crucially dependent upon the particular question being asked. Split-half reliabilities for the questionnaires range from about .4 to about .6. Reliabilities were obtained by computing the correlation between subject scores on odd and even questions. These reliabilities are not as high as would be required for a measuring instrument; but they indicate a reasonable amount of consistency in the subjects' relative ability to estimate answers to questions of a general information type.

Presumably, this consistency would allow the selection of a superior subgroup which would outperform the total group. One of the lowlights of our investigation so far is that we have not been able to find a criterion which enables the selection of such a subgroup. In particular, the use of a self-rating scale, either in terms of the relative confidence the subject has in his answers, or his relative performance vis-a-vis the group, has not offered a reliable way of singling out a superior subgroup.

In this respect, another lowlight has been the lack of any discernable relationship between the Concept Mastery Test scores and any of the other significant variables--especially accuracy of answers. The estimation task is clearly an intellectual one, but we may be tapping a skill that is not closely related to the ability measured by the Concept Mastery Test.

One of the more promising lines of analysis has come out of examining the way in which the initial (before feedback) answers are distributed. If the estimates are transformed to normal scores with a mean of 1 we obtain a distribution that is reasonably log-normal (Fig. 1.)

This leads to a preliminary picture of a very noisy process, in which there are many different bits and pieces of information influencing the subject, some of which tend to move the estimate upward, some downward. The outcome of the process is a random variate, distributed in most cases around the true answer, but in some cases exhibiting significant bias.

The spread in estimates, measured by the standard deviation of the log scores, for a given question, has a correlation of .83 with the accuracy of the group estimate on that question. This is a surprisingly high correlation, and suggests that the spread is a relatively good measure of the "solidity" of the initial estimate. This confirms an intuitive feeling which practically everyone appears

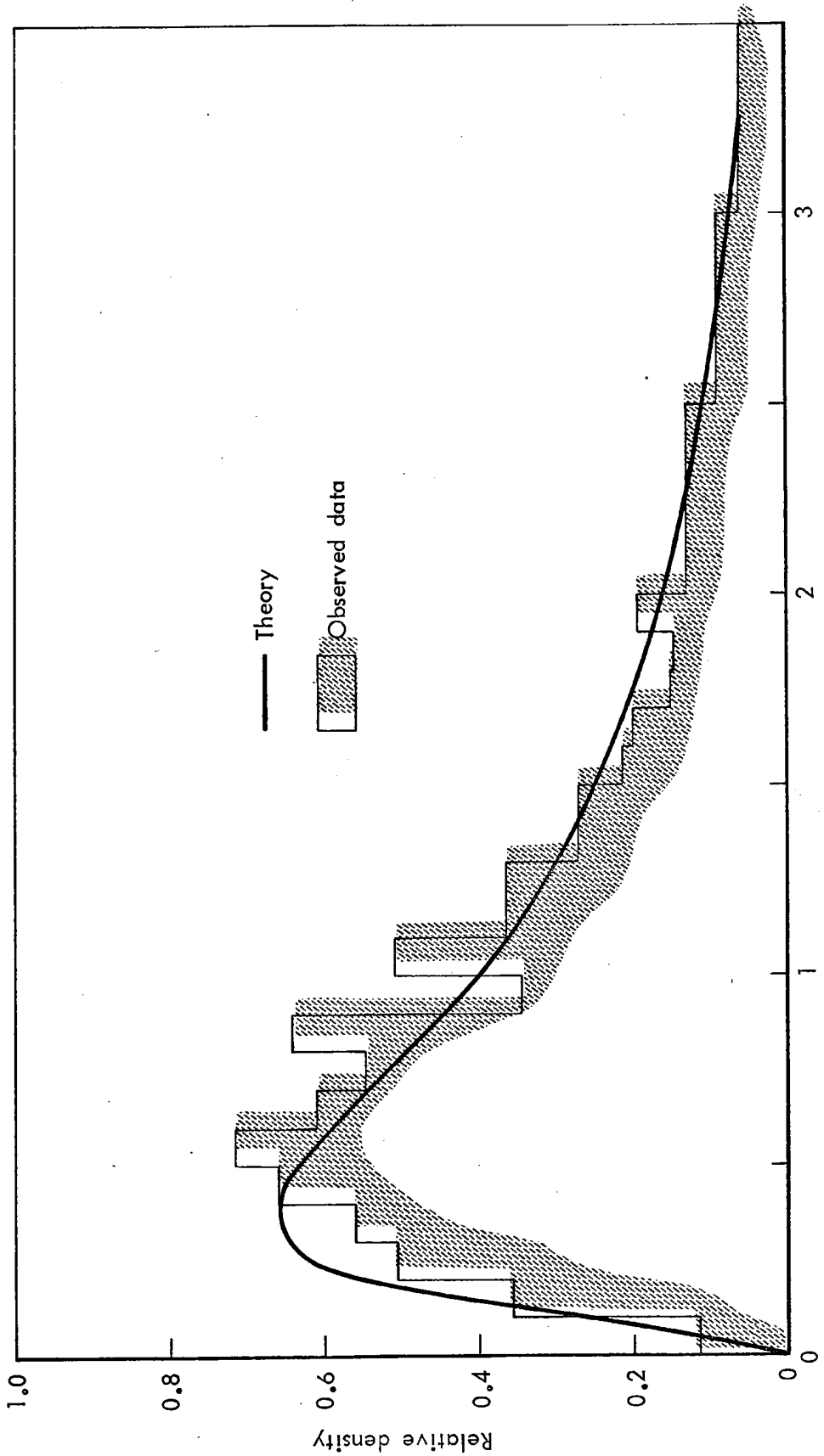


Fig. 1—Distribution of initial estimates

to have in this field. It is one of the bases for trying to reach a consensus.

The picture of estimation as a random process, possibly with bias, is given additional weight by the results obtained when the subjects are requested to give probability distributions rather than point estimates. In two of the sessions the experimental group made their estimates in the form of quartiles--i.e., they listed three numbers: the first, in their judgment, having a 25% chance of being too high, the second a 50% chance and the third a 75% chance. Formulating their estimates in this fashion improved performance in all respects--the initial responses of the experimental group were more accurate than those of the control group who made point estimates, the responses of the control group converged to a greater extent, and the improvement in accuracy between the initial and feedback rounds was greater. To date, aside from the iteration structure, the use of probability estimates is the most powerful device we have found for improving the accuracy and convergence of estimates.

CODA

On the books are experiments to explore the effects of learning, of differential reward structures, and of richer communication as well as the extension to short-range predictions. We have, of course, raised many more questions than we are likely to be able to answer by the

present series of experiments.

One of the most intriguing questions is the role of individual changes of opinion in the improvement of the group response. In some of the sessions, the "holdouts," those who change their opinions little or not at all, appear to be the more accurate estimators and improvement occurs by movement of the more volatile members toward the "center." In other cases, the holdouts do not appear to be the most accurate subgroup; but rather, improvement occurs by very large changes of opinion by the initially deviant members. We have a long way to go in accounting for the "fine structure" of the process.

In any case, the Delphi procedure has turned out to be a highly effective experimental structure with which to investigate group estimation processes. It appears likely that the results of these studies--and others like them--will afford valuable leads for improvements in opinion technology.

REFERENCES

(1) Maier, Norman R. F., "Assets and Liabilities in Group Problem Solving: The Need for an Integrative Function," Psychological Review, Vol. 74, No. 4, July 1967, pp. 239-249.

(2) Dalkey, Norman C., Delphi, The RAND Corporation, P-3704, October 1967.